DE GRUYTER
MOUTON

Francesco Cangemi, Meghan Clayards, Oliver Niebuhr, Barbara Schuppler, Margaret Zellers (Eds.)

RETHINKING REDUCTION

INTERDISCIPLINARY PERSPECTIVES ON CONDITIONS, MECHANISMS, AND DOMAINS FOR PHONETIC VARIATION

PHONOLOGY AND PHONETICS

Francesco Cangemi, Meghan Clayards, Oliver Niebuhr,
Barbara Schuppler and Margaret Zellers (Eds.)
Rethinking Reduction

Phonology and Phonetics

Editor
Aditi Lahiri

Volume 25

Rethinking Reduction

Interdisciplinary Perspectives on Conditions, Mechanisms, and Domains for Phonetic Variation

Edited by
Francesco Cangemi, Meghan Clayards, Oliver Niebuhr, Barbara Schuppler and Margaret Zellers

DE GRUYTER
MOUTON

ISBN 978-3-11-052163-4
e-ISBN (PDF) 978-3-11-052417-8
e-ISBN (EPUB) 978-3-11-052171-9
ISSN 1861-4191

Library of Congress Control Number: 2018934238

Bibliografische Information der Deutschen Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutschen Nationalbibliografie; detailed bibliographic data are available on the Internet http://dnb.dnb.de.

Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Preface

This book began as a result of a reunion workshop for members of Sound to Sense (S2S), an EU-funded Marie Curie Research Training Network (MC-RTN) that ran from 2007 to 2011. Engineers, computer scientists, psychologists, and linguistic phoneticians worked together to explore new ideas about how humans understand speech, and how scientists can use this knowledge to improve the way we communicate with each other and with machines. We used a variety of approaches to investigate what types of information are available in the speech signal, and how listeners use that information when they are listening in their native language, or in a foreign language, or in a noisy place like a railway station when it is hard to hear the speech. These three types of listening situation allow us to see how listeners actively use their pre-existing knowledge, together with the speech they hear, to understand a message.

The workshop leading to this book was held in Kiel in 2012, after S2S's funding period was over. Most PhD students on the grant had graduated and, like the postdoctoral students, moved on to employment around Europe. Twenty original S2S members attended the workshop, along with 10 people who had not been members. I unfortunately could not attend the workshop, but was asked to write this preface in my capacity as Coordinator of S2S. In particular, I was tasked with describing S2S and assessing how well the current book reflects the values and aims of S2S.

The book's title indeed reflects both the focus of the workshop and a major focus of S2S: how to observe, analyze, and describe the changes that words undergo when they are produced in meaningful connected speech, especially spontaneous speech, compared with in isolation; to what extent those changes help or hinder listeners to understand the meaning of utterances – that is, whether, and if so how, listeners exploit those changes to aid understanding; and to what extent the conclusions of such research point to a need to change standard cognitive-linguistic theories and approaches to applications such as users of foreign languages and automatic speech recognition by machines.

Within these general research areas, the chapters vary considerably, for example, in the type of speech studied (e.g. read, spontaneous conversational, careful, and fast), the range of languages and regional varieties (it would be good to extend further to non-European languages), the units of analysis (acoustic features, articulatory trajectories, phones, syllables, words, phrases), the types of measurement and methods (see Chapter 1), and the range of theoretical orientations espoused (see Chapters 1 and 9). There is a wealth of information in here, with some detailed examination of patterns within a language and speech

https://doi.org/10.1515/9783110524178-201

style, and several comparisons across languages. The typical approach is broadly enquiring, open-minded, yet critical, be it about general cognitive processes, or units of analysis such as syllables and phones. This breadth of approach yet attention to phonetic detail likewise reflects the spirit of S2S during its lifetime.

Another positive aspect of S2S was its inclusivity. The consortium began with quite a large membership, and grew bigger as it went on. Moreover, especially in the second half of its lifetime, when research questions and working patterns were fairly well established, it was our practice to broaden our horizons by inviting non-members to speak at our regular workshops. The contributors to this book very much reflect this spirit of outward-looking collaboration. A total of 21 authors have contributed to the 9 chapters. Eleven individuals who were either members of S2S or contributed invited talks and/or technical skills have contributed to six of the chapters (1, 3, 4, 5, 7 and 9). Of these, Chapters 4 and 7 each also includes an author not connected with S2S. The remaining three chapters (2, 6 and 8) are authored by 10 individuals who had no connection with S2S in its lifetime. The majority of the latter group work in North America, which tended to be excluded de facto from major EU funding. It is good to see North American as well as new European work represented in this book.

The five editors of the book, all former S2S fellows, have written helpful and thoughtful "bookends" to the other chapters which go beyond general orientation to the main contents of the book. Chapter 1 introduces the concept of reduction in both its simple and problematic aspects, and provides a comprehensive overview of the contents of the chapters and the overarching themes they represent. Chapter 9 first provides a historical perspective about this complex subject and then draws together conclusions from the various contributions as part of looking to the future. These chapters together remove any need for me to say more about the book's contents.

So I could stop right here. But instead, since some self-indulgence is acceptable in a preface, I would like to add a personal viewpoint. As this book makes clear, the term "reduction" in linguistics is historically predicated on the idea that there is something not quite "right" or "normal" about it – that citation-form pronunciations are to be preferred because they transmit more "information." Yet, as far as we know, the phenomena of reduction are integral to effective communication in every language and subject to interacting processes of every type that affects effective speech communication in its broadest sense. So rather than emphasizing the meaning of reduction as of taking something away from a preferred version, we might get further by defining it in terms of another of its many meanings, that of *essence*. The strong hypothesis would be that reduced forms provide the phonetic essence that allows effective communication in the particular circumstances under consideration. This hypothesis is probably too strong, but

may prove accurate if we develop models sophisticated and broad enough to take account of all the particular circumstances, which must include the utterance's function, and each interlocutor's beliefs about what the others involved in the interaction understand – and this should certainly be our aim. More achievable at the present time might be to adopt a simple concept of "essence." A good analogy may be that of reduction in cooking. When one reduces a sauce, one decreases its volume, but that is not the point. The real aim is to *add* properties that cannot be achieved without the process of reduction: at its best, an exquisite blending of the critical attributes of the ingredients that is intense in itself, and enhances the experience of the food it accompanies, yet bears such indirect resemblance to the original ingredients that only experts (be it chefs or native speakers) can recover what those unreduced ingredients must have been. Yet everyone, expert and novice alike, can eat and appreciate the reduced sauce without necessarily knowing how it relates to other manifestations of its ingredients.

In conclusion, I believe I speak for all the contributors in saying that my hope for this book is that it will help to refocus thinking about speech production and perception. Our field could benefit from greater sensitivity in acknowledging complexity, multiple influences, and context-dependency rather than contrasting single properties or pitting theories against each other in a narrow "either-or" approach. For example, debates about contrast versus gradience may become more nuanced when gradience is conceptualized as the consequence of multiple influences interacting on a sound pattern to different degrees, and perhaps for different reasons. Recognition of complexity may allow us to develop more realistic models. But we are not there yet. Most work on reduction processes is still at the descriptive stage, and it is proving hard to move beyond that. Gathering rich descriptions together is valuable, however, for if we can describe the complexity well, then underlying explanatory principles should be more easily discerned; indeed, as in many areas of biological sciences, explanation may turn out to be relatively simple once the right level of analysis is identified. But we are currently very far from reaching that depth of understanding. Rather, we are still at the stage of widening and redefining our fields of enquiry to include recognition that there is much more to the transmission of spoken information than that conveyed by syllables divorced from meaning, or by meaningful words spoken in isolation, and that only by adopting such broader perspectives will we be able to account for what in the past has been dismissed as inexplicable and probably irrelevant variation. This book is a welcome step toward that goal.

Prof. Sarah Hawkins
University of Cambridge

Acknowledgements

The idea for this book arose out of a post-project meeting of the Marie Curie Research Training Network "Sound to Sense" in December 2012 in Kiel, Germany. All of the editors passed through "Sound to Sense" as part of our training, and we also drew on that network in inviting authors and reviewers to participate in this project, along with other experts in our field. We are very grateful to the attendees of the Kiel meeting for the conversations inspiring this book, as well as to the network as a whole, which opened up doors of interdisciplinary collaboration and influenced our scientific thinking.

We are grateful to the authors for their high-quality contributions and their patience with the process of bringing this book together.

This book would not have been possible without the aid of a large corps of reviewers who contributed their expertise to improving this book's contents. These reviewers include (in alphabetical order): Federico Albano Leoni, Molly Babel, Louis ten Bosch, Dan Brenner, Donna Erickson, Evelin Graupe, Nikolaus Himmelmann, Benjamin Munson, Bob Port, Cristel Portes, Rebecca Scarborough, Odette Scharenborg, Torbjørn Svendsen, György Szaszák, and Francisco Torreira, as well as others who have chosen to remain anonymous.

The work of Barbara Schuppler was supported by a Hertha-Firnberg grant (T572N23) from the Austrian Science Fund (FWF). Work by Francesco Cangemi was supported by the German Research Foundation's Collaborative Research Center 1252 ("Prominence in language").

https://doi.org/10.1515/9783110524178-202

Contents

List of Contributors

Martine Adda-Decker
LPP (Laboratoire de Phonétique et de Phonologie) UMR7018, CNRS, France
madda@univ-paris3.fr

Francesco Cangemi
Institute for Linguistics - Phonetics, University of Cologne, Cologne, Germany
fcangemi@uni-koeln.de

Meghan Clayards
Department of Linguistics, School of Communication Sciences and Disorders, McGill University, Montreal, Canada
meghan.clayards@mcgill.ca

Cynthia G. Clopper
Department of Linguistics, Ohio State University, Columbus, OH, USA
clopper.1@osu.edu

Jennifer Cole
Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana, IL, USA
Department of Linguistics, Northwestern University, Evanston, IL, USA
jennifer.cole1@northwestern.edu

Francesco Cutugno
Dept. of Electrical Engineering and Information Technology, University of Naples "Federico II", Italy
cutugno@unina.it

Wim A. van Dommelen
Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway
wim.van.dommelen@ntnu.no

Mirjam Ernestus
Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
mirjam.ernestus@mpi.nl

Carol Espy-Wilson
Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA
espy@umd.edu

Louis Goldstein
Department of Linguistics, University of Southern California, Los Angeles, CA, USA
louisgol@usc.edu

Lori Lamel
LIMSI (Laboratoire d' Informatique et de Mécanique pour les Sciences de l'Ingénieur) UPR3251, CNRS, France
lamel@limsi.fr

Vikramjit Mitra
Speech technology and Research Lab, SRI International, Menlo Park, CA, USA

Oliver Niebuhr
SDU Electrical Engineering, Mads Clausen Institute, University of Southern Denmark, Sonderborg, Denmark
olni@sku.dk

Antonio Origlia
Dept. of Electrical Engineering and Information Technology, University of Naples "Federico II", Italy
antonio.origlia@unina.it

https://doi.org/10.1515/9783110524178-203

Elliot Saltzman
Department of Physical Therapy and Athletic Training, Boston University, Boston, MA, USA

Valentina Schettino
Department of Literary, Linguistic and Comparative Studies, University of Naples, Italy
vschettino@unior.it

Barbara Schuppler
Signal Processing and Speech Communication Laboratory, Graz, University of Technology, Graz Austria
b.schuppler@tugraz.at

Stefanie Shattuck-Hufnagel
Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA
sshuf@mit.edu

Ganesh Sivaraman
Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA

Rachel Smith
Glasgow University Laboratory of Phonetics, University of Glasgow, Glasgow, UK
rachel.smith@glasgow.ac.uk

Mark Tiede
Haskins Laboratories, New Haven, CT, USA

Rory Turnbull
Department of Linguistics, University of Hawai'i at Mānoa, Honolulu, HI, USA
rory.turnbull@hawaii.edu

Margaret Zellers
Institute for Scandinavian Studies, Frisian Studies and General Linguistics, Kiel University, Kiel, Germany
mzellers@isfas.uni-kiel.de

Margaret Zellers, Barbara Schuppler, and Meghan Clayards

1 Introduction, or: why rethink reduction?

Abstract: In phonetic reduction, segments may be shorter, less clearly articulated, or absent, compared to "canonical" or dictionary forms. While traditionally considered "slurred" or deficient, recent work has shown that reduction phenomena are highly complex. This chapter takes a historical and multidisciplinary approach, describing how views of reduced speech have evolved over time in the domains of phonetics, speech perception, and automatic speech recognition. We bring these perspectives together to raise several questions: Are reduced forms really inferior to canonical forms? Is the scope of reduction best described at the level of the feature, syllable, or larger unit? Does reduction generally occur in places where the content is more predictable? The chapters of this book are then contextualized with regard to these questions.

1.1 Introduction

This book is focused around the phenomenon of phonetic reduction in speech. In phonetic reduction, segments may be shorter, less clearly articulated, or absent, compared to "canonical" or dictionary forms. A classical view on reduction is given by Jakobson and Halle (1956):

> Since in various circumstances the distinctive load of the phonemes is actually reduced for the listener, the speaker, in his turn, is relieved of executing all the sound distinctions in his message: the number of effaced features, omitted phonemes and simplified sequences may be considerable in a blurred and rapid style of speaking. The sound shape of speech may be no less elliptic than its syntactic composition.... But, once the necessity arises, speech that is elliptic on the semantic or feature level, is readily translated by the utterer into an explicit form which, if needed, is apprehended by the listener in all its explicitness.
>
> The slurred fashion of pronunciation is but an abbreviated derivative from the explicit clear-speech form which carries the highest amount of information.... When analyzing the patterns of phonemes and distinctive features composing them, one must resort to the fullest, optimal code at the command of the given speakers.
>
> Jakobson and Halle (1956: 6)

Margaret Zellers, Kiel University
Barbara Schuppler, Graz University of Technology
Meghan Clayards, McGill University

https://doi.org/10.1515/9783110524178-001

Many early studies of reduction take a similar attitude to Jakobson and Halle's characterization of reduced forms as "slurred," "slovenly," or otherwise deficient. However, more recent work has shown clearly that reduction is much more complex. For example, while reduction is primarily thought of as a casual speech phenomenon, it also occurs in read speech, and indeed can make read speech easier to listen to and process; speech synthesis for the blind adopts reduction phenomena to make texts easier to listen to, especially over longer stretches of time (Jande 2003). Furthermore, an increasing amount of evidence suggests that "canonical" and "reduced" forms are not simply categorical oppositions, but may rather fall along a spectrum of pronunciation variants that are more or less clearly articulated (Nolan 1992). An acoustically "absent" segment may still leave prosodic and/or articulatory traces (Kohler and Niebuhr 2011; Niebuhr and Kohler 2011; Torreira and Ernestus 2011); and conversely, segments may be hyperarticulated to a point that, while their pronunciation may be "super-canonical," it does not reflect the typical production of that segment (Clayards and Knowles 2015). Schuppler et al. (2012), for instance, found that in conversational Dutch only 11.7% of the tokens show canonical realizations of word-final /t/ (i.e., a voiceless closure followed by one strong burst, produced at an alveolar place of articulation).

While a great deal of current research addresses the kinds of difficulties posed by dealing with reduced forms and spontaneous speech, cf. a special issue of the *Journal of Phonetics* (39[3], 2011) addressing this very topic, this book will ask the question of whether the ways we think about "reduction" are helpful, and how as researchers we could potentially shift our paradigms and methodologies, leading to greater understanding of this kind of variation in phonetic forms. Thus, this book brings together work from a variety of research and language backgrounds aimed at widening our understanding of what reduction is and how we as language users make use of it.

1.2 Examples of reduction

Reduction phenomena can be highly variable, particularly across languages. A few examples are provided here to illustrate some of the possibilities.

Figure 1.1 shows the acoustic realization of the utterance "we were supposed to see yesterday, but he felt really bad." For native speakers, it is easily understandable, even though it is realized with fewer segments, with only two instead of three syllables, and with different segments than the canonical form. The remaining segments are the initial consonant, the stressed vowels, and the fricative. As frequently shown in the literature, unstressed vowels and plosive

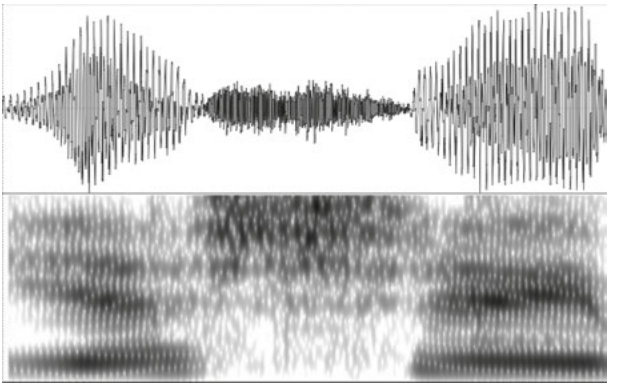

Figure 1.1: yesterday, realiszed as ['jeʃa͡i]; Buckeye corpus

closures are deleted completely. With a segment-based approach, this word would be hard to annotate and segment, as boundaries between the segments are overlapping. (However cf. Xu and Liu 2013, who propose that segments, like prosody, are produced as a series of approximations to dynamic targets, and that segmentation is thus preferable on the basis of gesture onsets in the context of syllable structure. Thus, in the current example, the onset of the fricative occurs before the first vowel gesture is complete, as can be seen by the continuation of voicing as well as by the clear formant structure visible within the frication.)

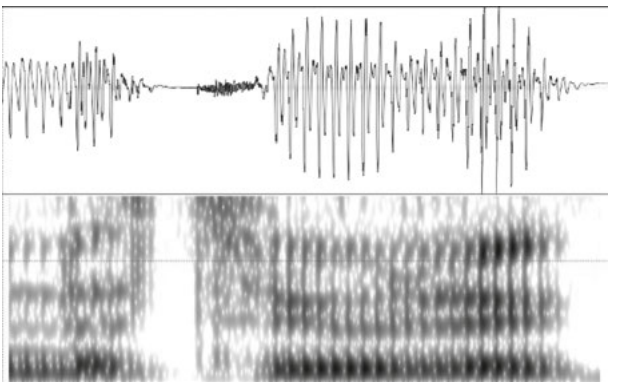

Figure 1.2: *natuurlijk*, canonical, Ernestus Corpus of Spontaneous Dutch

These examples of *natuurlijk* ("naturally") are taken from a Dutch corpus of spontaneous dialogues (Ernestus 2000). In Dutch, *natuurlijk* can be used with different functions; it can be an adjective, as in "natural languages," or it can have the

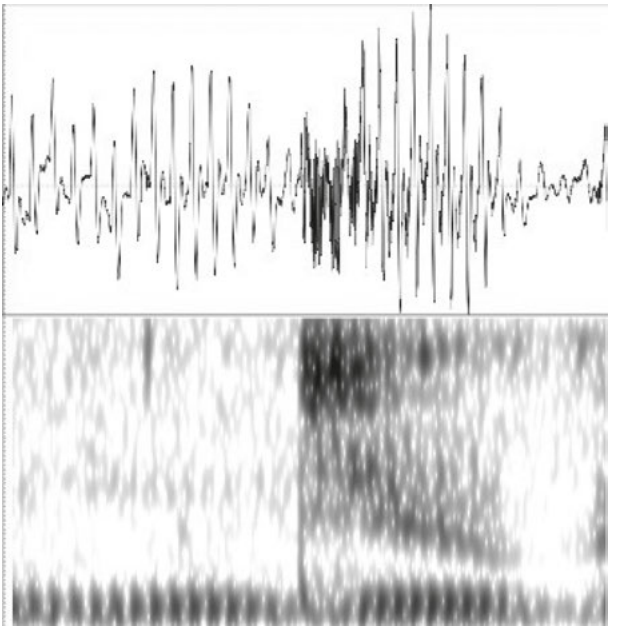

Figure 1.3: *natuurlijk* [nt'yk], Ernestus Corpus of Spontaneous Dutch

discourse-oriented meaning "of course." If presented in isolation, listeners are not able to understand the reduced form (Van de Ven, Ernestus, and Schreuder 2012), even though the canonical form used as "of course" never occurs in the corpus (Schuppler et al. 2011). When used as an adjective, however, the canonical form is observed frequently. Thus, the same sequence of phones is reduced differently for two different functions.

1.3 A historical and multidisciplinary perspective on reduction

1.3.1 Phonetics

Like Jakobson and Halle (1956), most early views on reduction coincided in considering reduced forms as lacking in some way. Jakobson and Halle thus argue that reduced forms are not worth studying compared to canonical forms, which contain the "fullest, optimal" information. However, not all contemporary researchers held the same view, and a number of phonetic studies touched on

topics relating to reduction, particularly in the context of research on lexical stress. Fry (1955, 1958), for example, reports that lexical stress in American English is marked by duration and intensity ratios between syllables in disyllabic words; that is, reduced duration and intensity are associated with the unstressed syllable. Similarly, Lieberman (1960) reports that stressed syllables in American English have higher fundamental frequency (F0), higher amplitude, and longer duration than unstressed syllables. Research on Swedish by Fant (1962) demonstrates that the formants in unstressed vowels move closer to those typical of schwa (i.e., 500, 1,500, and 2,500 Hz for the first three formants). Fónagy (1966) reports for Hungarian that increased articulatory effort (as in stressed syllables), measured by EEG, is associated with higher formant amplitude and broader bandwidth.

Lindblom (1963) addresses reduction phenomena directly by asserting the existence of "physiologically invariant" vowel targets, which are more or less closely approached based on the articulatory context. He treats reduction as a phenomenon resulting from limits on articulatory speed and claims that timing is more important in influencing reduction than lexical stress; that is, targets are undershot on the basis of decreased duration, not on the basis of lexical stress alone.

As phonetic research on reduction began to move beyond analysis of lexical stress alone, a large body of work also arose in the phonological community, providing contextual analyses that provided rules for certain types of reduction (e.g., schwa allophones of vowels or elision of consonants). Kohler (1974), for example, reports rules for the common elision of schwa from the German suffix *-en*. Gimson (1977), Elgin (1979), and Lass (1984) provide detailed grammars including elision rules. For Lass, deletion is phonologically defined as a segment merging with a null segment, although he indirectly addresses the phenomenon of reduction by stating that deletion is often the last stage of a lenition process (1984: 187). Since he deals with a phonological process rather than a phonetic one, however, the kinds of deletions he reports are not synchronically variable (in contrast with, for example, Dutch *natuurlijk*, as discussed in Section 1.2).

Reduction was also addressed from a sociolinguistic perspective in this time period. Labov (1972) reports on, among other phenomena, elision of word-final alveolar plosives in New York Black American English, and social distribution of syllable-final /r/ in department store workers. He does not study reduction as a set of individual phonetic cases, but rather as a way in which people modify their speech to demonstrate group membership. Ohala (1981) studies reduction in the context of explaining sound change over time. His work is related to Lindblom's on timing, arguing that articulatory timing is constrained by physical characteristics of articulators.

A turning point for the analysis of reduction came with Lisker's (1984) argument that invariance is not actually something to be expected of the speech signal. He points out that while invariance was considered as important as an explanation for why listeners are able to constantly identify sequences of a limited set of sounds from the continuous speech signal, in fact, "phonemes ... are not perceptual constants" (1984: 1201). This argument grew in part out of contemporaneous phonological analyses, in which articulatorily and acoustically different sounds appeared as allophones for the same phoneme. In fact, the argument that contrastiveness, rather than invariance, is essential for speech processing is fundamental for modern work on reduction.

Following this, more phonetic studies of reduction phenomena began to appear. Dalby (1986) characterizes a number of reduction phenomena occurring in fast American English speech. Kohler (1990), moving beyond his earlier phonological analysis, argues that phonology-based characterizations of reduction are too restrictive, and that reduction rules should instead be based on phonetic features, generating a larger variety of alternatives. Kohler specifically argues that phonology, being an abstraction, cannot account for the physical and/or phonetic processes that lead to reduction, and that therefore it is an insufficient basis for analyzing reduction. Instead, he refers back to "motor economy," as proposed by Lindblom (1963; see discussion above) and others, as a fundamental consideration.

Lindblom (1990) himself also argues against the "problem" of invariance. His theory of hyper- and hypo-articulation (H&H theory) states that articulation is influenced by both constraints on the production system (i.e., a preference for minimal expenditure of effort) and constraints on the output (i.e., the need to make oneself understood). Since segmental production must take both of these constraints into account, reduction is not a problem but simply a response to a particular set of system settings. Again, the goal is to create sufficient contrast (and thus understandability) rather than to connect to something invariant.

Articulatory Phonology (Browman and Goldstein 1990) provides a somewhat different analysis of reduction, while staying within a similar tradition. In this theory, variation in forms occurs on the basis of varying articulation rate and thus varying degrees of gestural overlap. At extreme degrees of overlap gestures can be "hidden" so that they are not acoustically/perceptually available. For instance, Browman and Goldstein's measurements showed that word-final /t/ in word combinations like *perfect memory* could be absent in the acoustic signal, even though the tongue clearly moved to the alveolar ridge. Since all form variations can be accounted for in this model by increased gestural overlap and decreased gestural magnitude (i.e., gestures are never added, changed, or deleted), this model necessarily asserts that all reduction is gradient. Johnson, Flemming, and Wright

(1993) follow up on the idea of a single-gradient system for speech sound forms: "[F]ortitions are more accurately seen as descriptions of the pronunciation of phonetic targets in the absence of lenitions, and hence it is apparently the case that for every lenition there is an equal and opposite fortition" (524).

More recently, new insights about reduced words and the conditions under which they occur have been achieved by two parallel growing interests: (1) in conversational speech and the collection of large conversational speech corpora, for example, for English, Pitt, Johnson, Hume, Kiesling, and Raymond (2005); Godfrey, Holliman, and McDaniel (1992); for Dutch, Ernestus (2000); for German, Peters (2005); and for French, Torreira, Adda-Decker, and Ernestus (2010); and (2) in creating automatic tools for phonetic annotation, for example, forced alignment, Adda-Decker and Lamel (2000), and Adda-Decker and Snoeren (2011). Studies based on such large, automatically annotated corpora obviously do not focus on the detailed pronunciation of words. However, this quantitative approach can identify trends for certain speaking styles and allows for the calculation of sophisticated statistical models in order to learn about the conditions under which certain reductions are likely to occur. For instance, Schuppler et al. (2012) found that the realization of word-final /t/ in Dutch spontaneous dialogues is conditions by word frequency, bigram frequency, segmental context, morphological structure, and phrase position. For English, Yuan and Liberman (2009) investigated conditions for /l/ variation in English. Chapters 4 (Adda-Decker and Lamel) and 7 (Cutugno et al.) present studies carried out with such quantitative, corpus-based approaches.

1.3.2 Speech perception

The "lack of invariance" problem (Lisker 1984) was a central issue for speech perception studies as much as for acoustic-phonetic studies. That is, the acoustic cues that signal a phone are very context dependent, even in carefully produced citation forms such as one encounters in lab speech. Since the early studies the sources and varieties of variation that have been considered have been greatly multiplied and include casual speech phenomena like assimilations, flapping, or deletions as well as changes in speaking style and speaker. The traditional approach has been to assume that these variations pose a problem for the (canonically based) speech recognition system and the search has been for processes or representations that can accommodate this variation such as talker normalization (see Johnson 2005 for summary); general auditory processes that normalize coarticulation and assimilation (e.g., Diehl, Lotto, and Holt 2004; Fowler 1986; Gow 2003; Mitterer, Csépe, and Blomert 2006); and statistical processes that infer canonical forms from variable data (e.g., Gaskell and Marslen-Wilson 1996;

McMurray and Jongman 2011; Snoeren 2011; Sonderegger and Yu 2010). Other approaches explicitly move away from the idea of a phone-based canonical form as the mental representation (e.g., Clayards 2010; Goldinger 1998; Hawkins and Smith 2001; Johnson 1997, 2006; Pisoni 1997; Port 2007).

Thus a central set of questions in the perception literature has been what level of processing deals with pronunciation variation (pre-lexical, lexical, context) and what kind of units or representations best accommodate or even incorporate variation. Another important question has been the role of the canonical form versus other forms in representation. These questions are mirrored in the phonetic and automatic speech recognition (ASR) literature as discussed in Sections 1.4.1 and 1.4.2. Despite the central role of representation and processing of phonetic variation of all kinds to the speech perception literature, perception studies have overwhelmingly focused on non-spontaneous and non-conversational speech. Some notable exceptions focusing on reduction include Pickett and Pollack (1963); Ernestus and Baayen (2007); Janse and Ernestus (2011); Kohler and Niebuhr (2011); Van de Ven, Schreuder, and Ernestus (2012); and Brouwer, Mitterer, and Huettig (2013). Examining perception of spontaneous, and especially conversationally produced, speech is clearly an important direction for future research in order to enrich our understanding of speech perception more generally. Chapter 6 (Cole and Shattuck-Hufnagel) takes a novel approach to examining perception of spontaneous speech through the use of imitation.

1.3.3 Automatic Speech Recognition

Since the first ASR experiments in the 1970s mainly focused on the recognition of isolated words, there was no need to investigate methods to deal with reduced pronunciations. At that time, researchers thought that speech recognition was soon to be a solved problem. However, as new applications continued to be proposed for ASR systems, ASR research needed to move beyond recognizing only isolated words. When researchers began using connected words and read speech, the need arose to find ways of incorporating coarticulation and pronunciation variation. In the ensuing decades, interest progressed more and more toward spontaneous and conversational speech, and recently to conversations between more than two people. Since the frequency of reduction phenomena is highest in conversations and lowest in read words, the importance of dealing with reduction has steadily increased alongside the increasing focus on conversational speech. For instance, a recent study on Austrian German database GRASS (Schuppler et al. 2014a) shows that while in read texts only 33.1% of the words are produced with a pronunciation different from the canonical form, in conversational speech

this number rises to 63.2% (Schuppler, Adda-Decker, and Morales-Cordovilla 2014b). The performance of ASR systems drops in proportion to the degree of spontaneity of the speech in question, as shown by Adda-Decker et al. (2013): an ASR system which was trained on read speech reaches nearly 96% word accuracy on read speech, whereas only 27% on spontaneous conversations involving the same speakers (without further adaptations and/or additional pronunciation modeling). It is thus clear that in order to make ASR systems work, pronunciation modeling has to be taken into account.

In general, reduction has been dealt with in the speech recognition community using the same methods as other sources of pronunciation variation (e.g., regional and social variation, variation due to anatomical differences of speakers, and emotional status; for an overview of different methods used, see Strik and Cucchiarini 1999). In a typical ASR system, the basic unit chosen is the phone; alternative approaches that are assumed to deal better with reduced words are the use of syllable (see Chapter 7, Cutugno et al.) or phonetic features. Apart from the choice of the basic unit, an additional question arising is which component of the ASR system needs to be adapted in order to make it robust to reduced speech; this is discussed further in Section 1.4.3.

1.4 Some open questions about reduction

1.4.1 What do "canonical" and "reduced" actually mean?

When we talk about "reduced" forms, we are usually implying that they are reduced compared to something else. This something else is generally a "canonical" form, which might also be called a full or citation form, or a dictionary form. It can be defined as the form of a lexical item that is used in clear speech, with all underlying phonemes having a phonetic realization; indeed, this is the traditional definition that researchers like Jakobson and Halle (1956) assume. However, recent research shows that such a simple, binary contrast is not sufficient to characterize the many different facets of reduction that can be observed in speech, and that some parts of a phonemic (or other underlying) structure may be more necessary for conveying information than others.

In the first place, the language surrounding canonical and reduced forms may be problematic in that it carries with it an implicit assumption that reduced items are deficient compared to canonical forms – indeed, the implication is that phonetic information is somehow missing from the reduced form, taken away from the "full" canonical form. Even in the Phonetics of Talk-in-Interaction,

where the potential meaningfulness of all phonetic forms is a basic tenet, the terms phonetic "upgrade" and "downgrade" may imply a hierarchical relationship between these kinds of forms, with downgraded forms being somehow "less" than upgraded forms (Traci Walker, pers. comm.). Lindblom's influential H&H theory (cf. Section 1.3.1) relies on the assumption that reduced forms are lacking by suggesting that talkers reduce effort and therefore reduce phonetic information when the needs of the listener are minimal. Chapter 2 (Clopper & Turnbull) questions the premise that reduction and listener needs are as tightly linked as proposed in H&H theory. However, it still shares the assumption that reduced forms are less useful for extracting lexical information from the signal.

One problem inherent in this loaded value comparison between canonical and reduced forms is the assumption about what linguistic forms are primary. Despite criticisms of "sloppy" conversational speech, researchers like Abercrombie (1965) have long been pointing out that "spoken prose," the focus of most traditional linguistic analyses, is derived from conversation, rather than vice versa. If we consider, following Abercrombie, that conversation is primary, it becomes increasingly difficult to make the leap to saying that reduced forms are less privileged compared to citation/canonical forms. Why then have "canonical forms" become so prominent? And how do we reconcile this with psycholinguistic findings that lend at least some support to the idea of canonical forms being privileged in perception, even outside of contexts where they would be produced naturally.

In basic word recognition tasks, canonical forms do seem to have an advantage. They are recognized faster (Ernestus and Baayen 2007; Janse 2004; Janse, Nooteboom, and Quené 2007; Ranbom and Connine 2007; Tucker 2011) and more easily (Pitt, Dilley, and Tat 2011); they prime more effectively (Andruski, Blumstein, and Burton 1994; Ranbom, Connine, and Yudman 2009), and exhibit stronger lexical biases on perception (e.g., Pitt 2009), than corresponding reduced forms. In cases where a pronunciation variant is extremely frequent, such as the flap variant of word internal intervocalic d/t in American English, it may behave similar to the canonical form (Connine 2004; Pitt, Dilley, and Tat 2011).

One limitation of findings related to the so-called canonical advantage is that they are often tested in the context of single spoken words, where canonical forms are very much more expected than reduced forms (e.g., Pitt 2009; Tucker 2011). In fact research has shown that the perception of reduced words is very much dependent on aspects of the context, including the speaking rate (Dilley and Pitt 2010). Other studies such as those by Ranbom, Connine, and Yudman (2009) and Viebahn, Ernestus, and McQueen (2015) include the canonical and reduced forms in full sentences that should favor reduced forms to a greater extent. Ranbom, Connine, and Yudman (2009) further vary whether the prosodic

environment favors the reduced or canonical variant (presence vs. absence of a prosodic break) and found a canonical advantage even in the environment favoring flapping. Viebahn, Ernestus, and McQueen (2015) varied the predictability of the words, with more predictable environments favoring the reduced variant, and again found a consistent advantage for canonical forms. However, even for these studies, because they involved deliberately and not spontaneously produced reduction, it is unclear whether the speaking style truly favored the canonical or reduced variant. Sumner (2013) found that when reduced forms are produced in a casual speaking style, they are recognized just as well as canonical forms produced in a careful speaking style. Tucker (2011) also found that the canonical advantage depended on word frequency, such that very high frequency words were processed more quickly with reduced variants. Thus, some of the processing advantage observed for canonical productions may be due to how well they match contextual expectations. Furthermore, the phonetic details of deliberately and spontaneously produced variation may not be the same. Gow (2002, 2003) has shown that spontaneous nasal place assimilations are phonetically distinct from deliberately produced ones, and that these phonetic differences may facilitate speech perception.

On the other hand, research on the perception of spontaneously produced reduced speech continues to find that recognition of reduced words is difficult. When the perception of severely reduced words from spontaneous productions is tested, recognizing reduced forms in isolation is very difficult (Ernestus, Baayen, and Schreuder 2002; Janse and Ernestus 2011), though recognition goes up when the context is provided. Furthermore, Brouwer et al. (2013) present both reduced and clearer pronunciations of words in their original contexts and find that canonical forms still have an advantage.

Several of the psycholinguistic studies discussed earlier raise the question of whether reduction can be treated as a simple application of rules (as in the Viebahn et al. study, in which "reduced" variants were carefully produced for an experiment) or whether the influence of a larger context is strictly necessary to bring about a correct form. In ASR systems, the most typical component where reductions are incorporated is the Pronunciation Dictionary. When the basic unit chosen is the phone, pronunciation variants are typically incorporated in the form of deletions, substitutions, and insertions of segments to the canonical form. Starting in the early 1970s, pronunciation variants were created automatically by formulating phonological rules and applying them to the canonical forms in terms of "re-write rules." Barnett (1974), for instance, developed a phonological rule compiler for American English, which took the phonetic features of the sounds into account (manner and place of articulation, voiced vs. voiceless), as well as the stress pattern of the word and the position of the

segment within the word. These phonological rules mostly deal with coarticulation (degemination, flap generation, homorganic stop insertion, etc.) and include few reductions for conversational speech, since they were not necessary for the data in question.

Since then, new approaches continue to be developed, all of them with the aim of creating a lexicon, but not all of them based on the development of overt, human-defined rules. Besides the rule-based approach (mentioned above; cf. also Van Bael et al. 2007), the increasing number of transcribed speech databases has allowed for the development of data-driven approaches (e.g., Hämäläinen, Ten Bosch, and Boves 2008; Kessens, Cucchiarini, and Strik 2003). A set of variants derived with a data-driven approach is specific to the database from which the variants were extracted and tends to contain fewer pronunciation variants for most words than a lexicon created with the knowledge-based approach. Not all plausible variants will be present for all word types, especially for words with a low frequency of occurrence. Since low-frequency words occur seldom by definition, correspondingly few variants of those words will appear in the speech material. For highly frequent words, however, the data-driven approach yields a good set of pronunciation variants. In order to compensate for the disadvantages, knowledge- and data-driven approaches have been combined (Wester 2002, Schuppler et al. 2011; Schuppler et al. 2014b; for a broader overview, see Barry and van Dommelen 2005; Hain 2005).

The question of the relationship between canonical and reduced forms is further complicated by research results from within Conversation Analysis/Phonetics of Talk-in-Interaction, demonstrating that the context in which canonical versus reduced (or upgraded vs. downgraded) forms may occur is not necessarily simply determined. Curl (2002, 2005) and Curl, Local, and Walker (2006) pursued a number of lines of study shedding light on the reduced (or not) properties of repeated elements in conversation, responding to a body of literature arguing that first mentions of lexical items tend to be more "canonical" in form, while mentions of something that is already present in the conversation tend to be more reduced (cf. Bard, Lowe, and Altmann 1989; Bell et al. 1999; Fowler 1988; Fowler and Housum 1987; Jurafsky et al. 1998, *inter alia*; also discussion in Section 1.4.3 about informativeness). While Ohala (1994) reports that creating a context of mishearing leads to speakers repeating their productions with more "canonical" features, Curl (2002) finds that not all repetitions as part of repair sequences, that is, those in which speakers resolve some kind of misunderstanding or incorrect information, are produced the same. Instead, these repetitions are sensitive to how the repair process fits in with the rest of the conversation. If it is well fitted to the current location, the repaired item is repeated with a phonetically "upgraded" form: louder,

expanded pitch range, increased duration, as well as differences in articulatory form. If the repair turn is not well fitted in its current location, however, the phonetic form employed in the repetition tends to closely resemble the phonetic form in the earlier production. Both of these findings contrast with the finding in experimental contexts and corpus studies that second mentions tend to be produced in a reduced form compared to the first mention (e.g., Baker and Bradlow 2009). They also highlight that a range of discourse factors play a role in determining the form a word takes. Curl also points out that since these turns are all produced clearly enough to obtain a display of understanding from the interlocutor, appropriateness in context is not necessarily the same thing as optimally clear speech.

Several chapters address the question of the identity of "canonical" and "reduced" forms, and how they relate to one another. For example, Ernestus and Smith (Chapter 5) propose that the essence of a word (similar to Niebuhr and Kohler's term "phonetic essence," Kohler and Niebuhr 2011; Niebuhr and Kohler 2011) is something other than a canonical form, and that we should perhaps be more interested in what may not be reduced away than in what is present when the "full form" is produced. Cole and Shattuck-Hufnagel (Chapter 6) similarly propose that the essence of a word may be captured in the form of "landmarks," which reduced forms still aim at. Espy-Wilson et al.'s research in articulation during spontaneous speech (Chapter 8) provides evidence for an underlying remnant of articulatory form, as would be predicted in the articulatory phonology framework, even when acoustic evidence of a segment is absent. van Dommelen (Chapter 3) addresses the issue of exposure to canonical and reduced forms in second-language learning, and the extent to which this influences L2 speakers' production behavior.

1.4.2 Units and scope of reduction phenomena

As discussed in Section 1.3.1, the bulk of the body of research into phonetic reduction has been focused at the level of the segment. However, the scope of reduction phenomena can also be considered to be much larger. Work in ASR fields, in particular, has investigated some larger domains for reduction, such as syllables or longer range acoustic features (AFs).

1.4.2.1 Syllable

Greenberg (1999) presents a quantitative analysis of pronunciation variation in conversational speech taken from the Switchboard corpus (Godfrey, Holliman,

and McDaniel 1992), and shows that variation is systematic if analyzed at the level of the syllable. This study concludes that syllabic onsets are realized in their canonical form much more frequently than nuclei or codas, and that word stress has systematic effects on the pronunciation of syllables. Greenberg (1999) concludes that the syllable is a more suitable unit than the phone for describing the variation occurring in spontaneous speech.

Since Greenberg's quantitative phonetic analyses, many studies in the field of speech technology have investigated whether a syllable-based speech recognition system may have advantages, especially in the case of spontaneous speech. The benefit of using the syllable, however, is not always straightforward. Hämäläinen et al. (2008), for instance, compare context-independent single-path and multi-path syllable models with context-dependent phone models. In contrast to their original hypothesis, single-path syllable models and context-dependent phone models outperform multi-path syllable models. Their analysis shows that word recognition is mostly conditioned by syllabic context and lexical confusability. Their results suggest that multi-path syllable models are only beneficial to an ASR system if the pronunciation variation described at the syllable level of pronunciation can be linked with the word level in the language model (LM).

The aforementioned study by Hämäläinen et al. can be seen as involving a full syllabic system, since the basic units of the acoustic models (AMs) are syllables. The disadvantage of this approach is that there are more different syllables and syllabic contexts than phones and phonetic contexts (e.g. for triphones), and thus more data are necessary in order to have enough material to train the syllable models. Promising methods, however, have been found by combining the training of acoustic phone models with information about their positions within the syllabic structure. For example, Shafran and Ostendorf (2003) incorporate syllabic structures into AM clustering. They found that phone model clustering on the basis of syllabic structures outperforms traditionally trained pentaphones in a recognition task on the spontaneous speech material from switchboard.

Whereas the aforementioned methods deal with reductions implicitly (i.e., the AMs are trained on both reduced and fully realized segments in the training material), there are also explicit ways to incorporate reductions specific to certain syllabic structures and/or properties into the pronunciation modeling component of the ASR system. Schuppler et al. (2011), for instance, created pronunciation variants by applying reduction rules, which were dependent on syllabic structure and stress patterns, to the canonical pronunciations of words. Thus, different reduction rules apply to nuclei of stressed versus unstressed syllables, to onset versus coda consonants and consonant-clusters.

1.4.2.2 Acoustic phonetic features

Even though the syllable is a larger unit than the word, it is still treated as linear in most analyses. One problem with segment-based approaches in ASR is that deletions are seen as "complete" deletions, and that there are no ways to capture "traces of segments left" in surrounding segments (i.e. overlapping features), as in the "yesterday" example in Section 1.2. Two chapters in this volume further explore the relationship between features and reduction in terms of acoustic phonetics. Both Ernestus and Smith (Chapter 5) and Cole and Shattuck-Hufnagel (Chapter 6) show that certain AFs are more or less likely to be reduced than others and that features that are left behind are often blended together. Researchers following a Firthian tradition (cf., e.g., Firth 1948; Ogden and Walker 2001) have begun to investigate segments in terms of their parallels to prosody; cf. also Xu and Liu's (2013) extension of the target approximation model to account for segment dynamics. The approach of articulatory phonology (Browman and Goldstein 1990, 1992), discussed in Section 1.3.1, seeks to model speech as a set of overlapping and dynamic gestures that are represented directly by both the speaker and the listener and is able to capture many reduction phenomena.

A study by Ostendorf (1999) proposes moving beyond the traditional "beads-on-a-string" model of speech by using representations of speech based on AFs, which strongly resemble the articulatory gestures that articulatory phonology considers as primitives). Such a representation could consist of several layers: for instance, one for manner of articulation, one for place of articulation, one for voicing, and one for nasality. Boundaries on different layers are placed independently of each other, and are thus capable of capturing the asynchronous gestures of the articulators, that is, the acoustic correlates of the articulatory gestures. Thus, AFs seem to offer a natural way for representing (semi-) continuous articulatory gestures and the ensuing acoustic characteristics of speech signals (e.g., Frankel, Wester, and King 2007).

Promising results using AFs as the basic unit instead of phones have been forthcoming since the early 1990s. Deng and Erler (1992) compare multidimensional (or multivalued) feature representation of speech with a phone-based representation in a Hidden Markov Model (HMM) recognition framework. They show that because of the high degree of data sharing, training data can be used well, and the resulting models are very capable of capturing coarticulatory effects such as feature spreading. For the task of stop consonant discrimination, they show performance gains for the AF-based system in comparison with the phone-based system.

Since the development of these early AF classifiers, automatic AF classification has been continuously further developed and used for speech recognition in adverse conditions (e.g., Kirchhoff 1999; Kirchhoff, Fink, and Sagerer 2002;

Schutte and Glass 2005), to build language-independent phone recognizers (Siniscalchi, Svendsen, and Lee 2007; Siniscalchi and Lee 2014), and in computational models of human spoken-word recognition (Scharenborg 2010).

1.4.2.3 Prosody

Since many studies have shown that acoustic reduction is conditioned by the prosodic properties of a word, the potential benefit of incorporating prosodic information into ASR systems has also been investigated. For these purposes, a large number of prosodic features are usually automatically extracted from the speech signal, including fundamental frequency (F0), energy and rhythm features such as timing, durations, and silent pauses. According to Ostendorf et al. (2003), "A major problem in computational modeling of prosody is that these acoustic correlates provide cues to many different types of information associated with different time scales, from segmental to phrasal to speaker characteristics" (148).

Prosody is incorporated into speech technology systems for various purposes: for example, the detection of utterance endings or possible places for the realization of backchannels in dialogue systems and the detection of paralinguistic characteristics such as emotions and speaker uncertainty. It has also been demonstrated that the incorporation of prosodic information into acoustic modeling and pronunciation modeling shows benefits, as, for instance, Ostendorf et al. (2003) do for a recognition task on the conversational speech material of the American English Switchboard corpus (Greenberg 1997). Ostendorf et al.'s model makes use of both intermediate symbolic representations and acoustic correlates of prosody. For the incorporation of prosody into large-vocabulary speech recognition, however, some obstacles are still to be overcome. Whereas acoustic properties are relatively easily extracted automatically, symbolic representations of prosody in conversational, spontaneous speech can still only be created by hand, a very time-consuming project, despite its great promise. One attempt at addressing this problem is the development of "silver standard" corpora, which involve high-quality automatic segmentation and labeling (cf. Mahlow et al. 2014).

1.4.3 Role of predictability

It has long been observed that there is a relationship between reduced speech and predictable speech, and one common assumption has been that words that are harder to perceive or produce (because they are less predictable) are produced more clearly and this is sometimes explained with appeals to audience design

(e.g., Lindblom's H&H theory). Clopper and Turnbull (Chapter 2) address this issue by reviewing the evidence in favor of the conclusion that more predictable and higher frequency words and segments are on average shorter, and vowels in them are more centralized. However, when one looks at very reduced tokens, they are often not especially predictable in their immediate context (Brouwer, Mitterer, and Huettig 2013). This means that in online production of speech, there is a lot of variability in the degree of reduction that is not accounted for by predictability. Clopper and Turnbull (Chapter 2) also note a number of complications to the relationship. It is also clear from much of the perception literature that more reduced words are harder to recognize, even (or especially) when they are spontaneous productions heard in their original context (Brouwer, Mitterer, and Huettig 2013). This again calls into question the idea that talkers reduce on-line where they can "get away with it" because the context makes up for their lack of clarity. Seyfarth (2014) argues that word durations are explained in part by mean predictability and not just contextual predictability. That is, words that are on average more predictable will tend to be reduced even when they occur in an unpredictable context. This result is more in line with a representational account of reduction rather than an online or listener-oriented account (cf. Clopper and Turnbull, Chapter 2).

The phonetic literature on Talk-in-Interaction has also noted the discrepancy between accounts of reduction related to predictability. For example, Local, Kelly, and Wells (1986; see also Niebuhr, Görs, and Graupe 2013), in their discussion of turn-taking in Tyneside English, include more centralized vowel quality, that is, a reduced production, as one feature of turn-constructional unit ends which are followed by speaker transition. The centralized vowels in these turn-final locations need not be part of repetitions of previous lexical material, although they may be. Instead, they are produced in a reduced form (in concert with other cues) in order to indicate a speaker's intention to stop speaking. Local et al. (1986) do not find evidence that lexical material produced in these locations is less informative than in other locations. Thus, the use of "reduced" forms must be context-appropriate on some basis other than it simply being easier to access. Curl and colleagues' observations about repetitions in the context of repair (discussed in Section 1.4.1) also show that some repairs are not clearer than the original productions, again challenging the notion that reduction occurs in places where the listener requires less information to recognize the word.

In light of this discussion, it is also of interest to consider how the structure of a typical ASR system can be compared with the H&H model: a top-down LM makes hypotheses about how likely different words are given a specific N-gram context, while a bottom-up AM plus a lexicon makes the same hypotheses based on the acoustic signal. If the H&H model were strictly true, that is, loss of information in the bottom-up signal occurs in locations where the top-down model should

make strong predictions, these two should balance each other out. However, the fact that ASR models struggle with reduction seems to challenge this assumption. One factor may be that in a traditional ASR system, the LM, AM, and the pronunciation model (i.e., the lexicon) are trained independently. In order to make ASR systems more robust for conversational speech (which has a high number of reduced words), methods have been developed to combine the modeling of the components. For instance, this can be achieved by using the variants themselves (instead of the underlying words) to calculate the N-grams of the LM. First, this would easily allow incorporation of cross-word processes and reductions at the word boundaries. Second, variants that tend to occur together (for instance, in multi-word expressions, highly frequent bigrams and trigrams such as "I have done" and "I don't know") are also modeled together. This approach seems simple, but runs into problems of data-coverage. Whereas for regular LMs, all text sources available can be used to train the N-grams, for such a combined LM phonetically transcribed speech material is required. Kessens et al. (1999) presents a way to get around this "lack of data" issue; he trains an LM to which pronunciation variants that are created from phonological rules are added for each word, including both within-word and cross-word processes. With this approach, the system's word error rate (WER) improved by 8.8%.

1.5 Chapter summaries

Each chapter of this book addresses some aspect of the issues we have raised surrounding our understanding of reduction phenomena. The chapter by *Clopper and Turnbull* addresses sources of reduction, including listener-oriented, talker-oriented, and structural/evolutionary accounts of reduction. They find evidence for a complex relationship between these different sources.

The next two chapters examine reduction from a cross-linguistic perspective. The chapter by *van Dommelen* investigates influences on reduction for second-language speakers of a language, who often have more access to canonical productions in their target language. He finds evidence that second-language learners may reduce in similar ways to native speakers, though to differing degrees. The chapter by *Adda-Decker and Lamel* uses ASR tools to conduct automatic analysis of reduction phenomena across different languages and speaking styles, taking advantage of canonical word dictionaries to identify areas with reduced phonetic production.

The last four chapters question the common assumption of phone-based reduction. The chapter by *Ernestus and Smith* reports factors conditioning the reduction of *eigenlijk* in Dutch, and raises the issue of the automaticity of reduction

phenomena, as well as that of invariant landmarks occurring even in the most reduced productions of a word. The chapter by *Cole and Shattuck-Hufnagel* further investigates phonetic landmarks in imitations of reduced forms, reporting on which kinds of landmarks are most stable across speakers and contexts. The chapter by *Cutugno, Origlia, and Schettino* addresses mismatches between expected and observed production in speech on the syllable level. They use automatic methods to identify syllables with reduced forms, and propose a rule-based relationship between reduced and non-reduced forms. The chapter by *Espy-Wilson, Tiede, Mitra, Sivaraman, Saltzman, and Goldstein* uses a speech inversion system to identify areas of overlapping gesture, even where acoustic evidence for reduced segments may be lacking, with the aim of improving ASR systems.

Finally, the *conclusion* considers reduction from the perspective of its role in the history of the field(s) and draws together some of the implications of the studies reported in this book for future conceptions of and research on phonetic reduction.

References

Abercrombie, D. 1965. Studies in phonetics and linguistics. London: Oxford University Press.

Adda-Decker, M. & L. Lamel 2000. Modeling reduced pronunciations in German. Phonus 5, Institute of Phonetics, University of the Saarland, 129–143.

Adda-Decker, M., Schuppler, B., Lamel, L., Morales-Cordovilla, J-A., Adda, G. 2013. What we can learn from ASR errors about low-resourced languages: A case-study of Luxembourgish and Austrian. ERRARE Workshop 2013, Ermenonville.

Adda-Decker, M. & N. D. Snoeren 2011. Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics* 39 (3). 261–270.

Andruski, J. E., S. Blumstein & M. Burton. 1994. The effect of subphonetic differences on lexical access. *Cognition* 52. 163–187.

Baker, R. E. & A. R. Bradlow. 2009. Variability in word duration as a function of probability, speech style, and prosody. *Language & Speech* 52 (4). 391–413.

Barnett, J. 1974. A phonological rule compiler. In: L. Erman., L. (ed.), *Proceedings of the IEEE Symposium on Speech Recognition*, Carnegie-Mellon University, Pittsburgh, PA, 188–192.

Bard, E. G., A. J. Lowe & G. T. M. Altmann. 1989. The effect of repetition on words in recorded dictations. In *Eurospeech '89: Proceedings of the European Conference on Speech Communication and Technology*, vol. 2, 573–576.

Barry, W. J. & W. A van Dommelen. 2005. *The integration of phonetic knowledge in speech technology*. Dordrecht, the Netherlands: Springer. ISBN 1-4020-2635-8.

Bell, A., D. Jurafsky, E. Fosler-Lussier, C. Girand & D. Gildea. 1999. Forms of English function words – effects of disfluencies, turn position, age and sex, and predictability. *Proceedings of ICPhS*, San Francisco, USA.

Brouwer, Susanne, Holger Mitterer & Falk Huettig. 2013. Discourse context and the recognition of reduced and canonical spoken words. *Applied Psycholinguistics* 34. 519–539. doi:10.1017/s0142716411000853.

Browman, C. P. & L. Goldstein. 1990. Tiers in articulatory phonology, with some implications for casual speech. Papers in laboratory phonology I: Between the grammar and physics of speech, 341–376.

Browman, C. P. & L. Goldstein. 1992. Articulatory phonology: An overview. *Phonetica* 49 (3–4). 155–180.

Clayards, M. 2010. Using probability distributions to account for recognition of canonical and reduced word forms. In LSA Annual Meeting Extended Abstracts (Vol. 1, 4 pages). http://dx.doi.org/10.3765/exabs.v0i0.529

Clayards, M. & T. Knowles. 2015. Prominence enhances voiceless-ness and not place distinctions in English voiceless sibilants. *Proceedings of ICPhS 2015*, Glasgow, Scotland.

Connine, C. 2004. It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. *Psychonomic Bulletin & Review* 11 (6). 1084–1089.

Curl, T. S. 2002. *The phonetics of sequence organization: An investigation of lexical repetition in other-initiated repair sequences in American English*. Ph.D. dissertation, University of Colorado.

Curl, T. S. 2005. Practices in other-initiated repair resolution: The phonetic differentiation of 'repetitions'. *Discourse Processes* 39 (1). 1–43.

Curl, T. S., J. Local & G. Walker. 2006. Repetition and the prosody-pragmatics interface. *Journal of Pragmatics* 38 (10). 1721–1751.

Dalby, J. 1986. Phonetic structure of fast speech in American English. Bloomington, IN: Indiana University Linguistics Club.

Deng L. & K. Erler. 1992. Structural design of HMM speech recognizer using multi-valued phonetic features: Comparison with segmental speech units. *Journal of the Acoustical Society of America* 92. 3058–3067.

Diehl, R. L., A. J. Lotto & L. L. Holt. 2004. Speech perception. *Annual Review of Psychology* 55. 149–179. doi:10.1146/annurev.psych.55. 090902.142028

Dilley, L. C. & M. A. Pitt 2010. Altering context speech rate can cause words to appear or disappear. *Psychological Science* 21 (11). 1664–1670. 10.1177/0956797610384743

Elgin, S. H. 1979. *What is linguistics*. Second Edition. Englewood Cliffs: Prentice-Hall.

Ernestus, M. 2000. *Voice assimilation and segment reduction in casual Dutch. A corpus-based study of the phonology-phonetics interface*. Ph.D. dissertation, LOT, Vrije Universiteit Amsterdam, The Netherlands.

Ernestus, M. & R. H. Baayen 2007. The comprehension of acoustically reduced morphologically complex words: The roles of deletion, duration and frequency of occurrence. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 773–776.

Ernestus, M., H. Baayen & R. Schreuder. 2002. The recognition of reduced word forms. *Brain and Language* 81. 162–173.

Fant, G. 1962. *Den akustika fonetikens grunder*. Kungliga Tekniska Högskolan, Taltransmissionslab. 2nd printing. Stockholm: Royal Institute of Technology, no. 7.

Firth, J. R. 1948. Sounds and Prosodies. Reprinted in Palmer (ed.) 1970 *Prosodic analysis*. London: Oxford University Press, 1–26.

Fónagy, I. 1966. Electrophysical and acoustic correlates of stress and stress perception. *Journal of Speech and Hearing Research* 9.231–244.

Fowler, C. A. 1986. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* 1. 3–28.

Fowler, C. A. 1988. Differential shortening of repeated content words produced in various communicative contexts. *Language & Speech* 31 (4). 307–20.
Fowler, C. A. & J. Housum. 1987. Talkers signaling of "new and 'old' words in speech and listeners" perception and use of the distinction. *Journal of Memory and Language* 26. 489–504.
Frankel, J., M. Wester & S. King. 2007. Articulatory feature recognition using dynamic Bayesian networks. *Computer Speech and Language* 21 (4). 620–640.
Fry, D. B. 1955. Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America* 27 (4). 765–768.
Fry, D. B. 1958. Experiments in the perception of stress. *Language & Speech* 1 (2). 126–152.
Gaskell, M. G. & W. D. Marslen-Wilson. 1996. Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance* 22. 144–158. doi:10.1037/0096-1523.22.1.144
Gimson, A. C. 1977. *An introduction to English pronunciation*. London: Edward Arnold.
Godfrey, J. J., E. C. Holliman & J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of ICASSP*, vol. 1, 517–520.
Goldinger, S. D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105 (2). 251.
Gow, D. W., Jr. 2002. Does English coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception and Performance* 28. 163–179.
Gow, D. W., Jr. 2003. Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics* 65. 575–590. doi: 10.3758/BF03194584
Greenberg, S. 1997. The Switchboard Transcription Project in Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD (56 pp.).
Greenberg, S. 1999. Speaking in shorthand. A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29. 159–176.
Hain, T. 2005. Implicit modelling of pronunciation variation in automatic speech recognition. *Speech Communication* 46. 171–188.
Hawkins, S. & R. Smith. 2001. Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistics* 13. 99–188.
Hämäläinen, A., L. Ten Bosch & L. Boves. 2008. Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider. *Speech Communication* 51 (2). 130–150.
Jakobson, R. & M. Halle. 1956. *Fundamentals of language*. The Hague: Mouton & Co.
Jande, P.-A. 2003. Evaluating rules for phonological reduction in Swedish. *PHONUM* 9. 149–152.
Janse, E. 2004. Word perception in fast speech: Artificially time-compressed vs. naturally produced fast speech. *Speech Communication* 42. 155–173.
Janse, E., S. Nooteboom & H. Quené. 2007. Coping with gradient forms of /t/- deletion and lexical ambiguity in spoken word recognition. *Language and Cognitive Processes* 22. 161–200.
Janse, E. & M. Ernestus. 2011. The roles of bottom-up and top-down information in the recognition of reduced speech: Evidence from listeners with normal and impaired hearing. *Journal of Phonetics* 39 (3). 330–343.
Johnson, K. 1997. Speech perception without speaker normalization. In K. Johnson & J. W. Mullenix (eds.), *Talker variability in speech processing*, 145–165. San Diego: Academic Press.

Johnson, K. 2005. Speaker normalization. In R. Remez & D. B. Pisoni (eds.), *The handbook of speech perception*, 363–389. Oxford: Blackwell.
Johnson, K. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34 (4). 485–499.
Johnson, K., E. Flemming & R. Wright. 1993. The hyperspace effect: Phonetic targets are hyperarticulated. *Language* 69 (3). 505–528.
Jurafsky, D., A. Bell, E. Fosler-Lussier, C. Girand & W. Raymond. 1998. Reduction of English function words in switchboard. *Proceedings of ICSLP*, vol. 7, Sydney, Australia, 3111–3114.
Kessens, J. M., C. Cucchiarini & H. Strik 2003. A data-driven method for modeling pronunciation variation. *Speech Communication* (40). 517–534.
Kessens, J. M., M. Wester & H. Strik 1999. Improving the performance of a Dutch CSR by modelling within-word and cross-word pronunciation variation. *Speech Communication* 29 (2–4). 193–207.
Kirchhoff, K. 1999. *Robust speech recognition using articulatory information*. Ph.D. dissertation, University of Bielefield.
Kirchhoff, K., G. A. Fink & G. Sagerer 2002. Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication* 37. 303–319.
Kohler, K. J. 1974. Koartikulation und Steuerung im Deutschen. In *Sprachsystem und Sprachgebrauch: Festschrift für Hugo Moser*, 172–192, Teil 1. Düsseldorf: Schwann.
Kohler, K. J. 1990. Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In W. J. Hardcastle & A. Marchal (eds.), *Speech production and speech modelling*, 69–92. Dordrecht: Kluwer Academic Publishers.
Kohler, K. J. & O. Niebuhr 2011. On the role of articulatory prosodies in German message decoding. *Phonetica* 68: 1–31.
Labov, W. 1972. Sociolinguistic patterns. Philadelphia: University of Pennsylvania Press.
Lass, R. 1984. *Phonology: An introduction to basic concepts*. Cambridge: Cambridge University Press.
Lieberman, P. 1960. Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America* 32 (4). 451–454.
Lindblom, B. 1963. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35 (11). 1773–1781.
Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In Speech production and speech modelling, 403–439. Springer, Dordrecht.
Lisker, L. 1984. The pursuit of invariance in speech signals. *Journal of the Acoustical Society of America* 77 (3).1199–1202.
Local, J., J. Kelly & W. Wells 1986. Towards a phonology for conversation: Turn-taking in Tyneside English. *Journal of Linguistics* 22. 411–437.
Mahlow, C., K. Eckart, J. Stegmann, A. Blessing, G. Thiele, M. Gärtner & J. Kuhn 2014. Resources, tools and applications at the CLARIN center Stuttgart. *Proceedings Konvens 2014*, Hildesheim, Germany.
McMurray, B. & A. Jongman 2011. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review* 118 (2). 219.
Mitterer, H., V. Csépe & L. Blomert 2006. The role of perceptual integration in the recognition of assimilated word forms. *Quarterly Journal of Experimental Psychology* 59. 1395–1424. doi:10.1080/ 17470210500198726
Niebuhr, O., K. Görs & Graupe, E. 2013. Speech reduction, intensity, and F0 shape are cues to turn-taking. Proceedings of the 14th Annual SigDial Meeting on Discourse and Dialogue, Metz, France, 261–269.

Niebuhr, O. & K. J. Kohler 2011. Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics* 39. 319–329.
Nolan, F. 1992. The descriptive role of segments: Evidence from assimilation. In G. Docherty & D. R. Ladd (eds.), *Laboratory Phonology 2*, 261–280. Cambridge: Cambridge University Press.
Ogden, R. & G. Walker 2001. We speak prosodies and we listen to them. *Symposium on Prosody and Interaction*, Uppsala, Sweden.
Ohala, J.J. 1981. The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, & M. F. Miller (eds.), Papers from the Parasession on Language and Behavior. Chicago: Chicago Linguistic Society, 178 - 203.
Ohala, J. J. 1994. Acoustic study of clear speech: A test of the contrastive hypothesis. *International Symposium on Prosody*, Vol. 18, 75–89.
Ostendorf, M., 1999. Moving beyond the "Beads-on-a-String" model of speech. *Proceedings of 1999 IEEE Workshop on Automatic Speech Recognition and Understanding*, Vol. 1. 79–83.
Ostendorf, M., I. Shaffran & R. Bates 2003. Prosody models for conversational speech recognition. *Proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, 147–154.
Peters, B. 2005. The database – The Kiel Corpus of Spontaneous Speech. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), 35a, 1–6.
Pickett, J. M. & I. Pollack 1963. Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language & Speech* 3. 151–164.
Pisoni, D. B. 1997. Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullenix (eds.), *Talker variability in speech processing*, 9–32. San Diego: Academic Press.
Pitt, M. A. 2009. The strength and time course of lexical activation of pronunciation variants. *Journal of Experimental Psychology: Human Perception and Performance* 35. 896–910.
Pitt, M. A., L. Dilley & M. Tat 2011. Exploring the role of exposure frequency in recognizing pronunciation variants. *Journal of Phonetics* 39 (3). 304–311.
Pitt, M. A., K. Johnson, E. Hume, S. Kiesling & W. D. Raymond 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication* 45. 89–95.
Port, R. 2007. How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology* 25 (2). 143–170.
Ranbom, L. & C. M. Connine 2007. Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language* 57 (2). 273–298.
Ranbom, L. J., C. M. Connine & E. M. Yudman 2009. Is phonological context always used to recognize variant forms in spoken word recognition? The role of variant frequency and context distribution. *Journal of Experimental Psychology: Human Perception and Performance* 35 (4). 1205–1220.
Scharenborg, O. 2010. Modeling the use of durational information in human spoken-word recognition. *Journal of the Acoustical Society of America* 127 (6). 3758–3770.
Schuppler, B., M. Adda-Decker & J. A. Morales-Cordovilla 2014b. Pronunciation variation in read and conversational Austrian German. *Proceedings of Interspeech 2014*, 1453–1457.
Schuppler, B., M. Ernestus, O. Scharenborg & L. Boves 2011. Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics* 39. 96–109.
Schuppler, B., M. Hagmüller, J. A. Morales-Cordovilla & H. Pessentheiner 2014a. GRASS: The Graz Corpus of Read and Spontaneous Speech. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1465–1470.

Schuppler, B., W. van Dommelen, J. Koreman & M. Ernestus 2012. How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. *Journal of Phonetics* 40. 595–607.

Schutte, K. & J. Glass 2005. Robust detection of sonorant landmarks. *Proceedings of 6th Interspeech*, Lisbon, Portugal, 1005–1008.

Seyfarth, S. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133 (1). 140–155.

Shafran, I. & M. Ostendorf 2003. Acoustic model clustering based on syllable structure. *Computer Speech & Language* 17 (4). 311–328.

Siniscalchi, S. M. & C.-H. Lee 2014. An attribute detection based approach to automatic speech recognition. *Loquens* 1 (1). e005. doi:http://dx.doi.org/10.3989/loquens.2014.005.

Siniscalchi, S. M., T. Svendsen & C.-H. Lee 2007. Towards bottom-up continuous phone recognition. *Proceedings of IEEE ASRU Workshop*, 566–569.

Snoeren, N. D. 2011. *Psycholinguistique cognitive de la parole assimilée*. Saarbrücken, Germany: Editions Universitaires Européennes.

Sonderegger, M. & A. C. L. Yu 2010. A rational account of perceptual compensation for coarticulation. In S. Ohlsson & R. Catrambone (eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin: Cognitive Science Society, 375–380.

Strik, H. & C. Cucchiarini 1999. Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication* 29 (2–4). 225–246.

Sumner, M. 2013. A phonetic explanation of pronunciation variant effects. *Journal of the Acoustical Society of America* 134 (1). EL26–EL32.

Torreira, F., M. Adda-Decker & M. Ernestus 2010. The Nijmegen Corpus of Casual French. *Speech Communication* 52 (3). 201–212.

Torreira, F. & M. Ernestus 2011. Vowel elision in casual French: The case of vowel /e/ in the word c'était. *Journal of Phonetics* 39 (1). 50–58.

Tucker, B. V. 2011. The effect of reduction on the processing of flaps and/g/in isolated words. *Journal of Phonetics* 39 (3). 312–318.

Van Bael, C., L. Boves, H. van den Heuvel & H. Strik 2007. Automatic phonetic transcription of large speech corpora. *Computer Speech and Language* 21. 652–668.

Van de Ven, M., M. Ernestus & R. Schreuder 2012. Predicting acoustically reduced words in spontaneous speech: The role of semantic/syntactic and acoustic cues in context. *Laboratory Phonology* 3. 455–481.

Viebahn, M. C., M. Ernestus & J. M. McQueen 2015. Syntactic predictability in the recognition of carefully and casually produced speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41 (6). 1684–1702.

Wester, M. 2002. *Pronunciation Variation Modeling for Dutch Automatic Speech Recognition*. PhD dissertation, Radboud University Nijmegen, The Netherlands.

Xu, Y. & F. Liu 2013. Intrinsic coherence of prosodic and segmental aspects of speech. In O. Niebuhr & H. R. Pfitzinger (eds.) *Prosodies: Context, function, and communication,* 1–26. Berlin/New York: de Gruyter.

Yuan, J. & M. Liberman 2009. Investigating /l/ variation in English through forced alignment. *Proceedings of Interspeech 2009*, 2215–2218.

Cynthia G. Clopper and Rory Turnbull

2 Exploring variation in phonetic reduction: Linguistic, social, and cognitive factors

Abstract: Substantial empirical research has revealed that temporal and spectral phonetic vowel reduction occurs in "easy" processing contexts relative to "hard" processing contexts, including effects of lexical frequency, lexical neighborhood density, semantic predictability, discourse mention, and speaking style. Theoretical accounts of this phonetic reduction process include listener-oriented approaches, in which the reduction reflects the talker's balancing the comprehension needs of the listener with production effort constraints, talker-oriented approaches, in which reduction is argued to result entirely from constraints on speech production processes, and evolutionary approaches, in which reduction results directly from long-term interactive communication within a community. Recent research in our laboratory has revealed complex interactions among the linguistic, social, and cognitive factors involved in phonetic vowel reduction processes. These interactions reveal variation in the robustness of phonetic reduction effects across linguistic factors, as well as different patterns of interactions among linguistic, social, and cognitive factors across acoustic domains. These interactions challenge aspects of each of the three existing models of phonetic reduction. We therefore propose that a more complex view of the relationship between processing demands and phonetic vowel reduction processes is necessary to account for these observed patterns of variation.

Keywords: lexical frequency, neighborhood density, contextual predictability, speaking style, regional dialect

2.1 Introduction

Phonetic reduction is one of many processes contributing to variation in the acoustic-phonetic realization of speech. We define phonetic reduction as the phenomenon in which linguistic units (e.g., segments, syllables, or words) are realized with relatively less acoustic-phonetic substance (e.g., shorter duration and/or less extreme articulation) in a given context relative to other contexts. We assume that phonetic reduction involves acoustic-phonetic variation in realized

Cynthia G. Clopper, Ohio State University
Rory Turnbull, University of Hawai'i at Mānoa

https://doi.org/10.1515/9783110524178-002

segments along a continuum from hypoarticulated or reduced to hyperarticulated or enhanced. We therefore consider phonetic reduction to reflect a reduced degree of acoustic-phonetic substance in comparison to more hyperarticulated or enhanced forms (Johnson, Flemming, and Wright 1993). This variation in the degree of acoustic-phonetic substance is assessed using measures of segment and word duration, vowel space expansion, and f0, among others.

We limit our discussion in this chapter primarily to phonetic variation along measurable acoustic dimensions and therefore do not include categorical reduction processes, such as segmental alternations (e.g., full vowels alternating with schwa or stop consonants alternating with flap) or the deletion of segments, syllables, or words (cf. Ernestus 2014; Johnson 2004; Schuppler et al. 2011). We similarly focus on lexical and contextual factors that have been described in the literature as contributing to phonetic reduction as we have defined it here and therefore do not include phonetic reflexes of phonological properties such as segmental context (cf. Klatt 1976; Luce and Charles-Luce 1985; Peterson and Lehiste 1960), lexical stress (cf. de Jong 1995, 2004; Fourakis 1991; van Bergem 1993), or prosodic structure (cf. Lehiste 1971; Wightman et al. 1992). This division between continuous, phonetic reduction and categorical, phonological processes provides us with a more clearly circumscribed focus of discussion in this chapter, but it most likely does not reflect a true, natural division in language processing.[1] We therefore expect our conclusions to extend to categorical reduction processes (see also Cohen Priva 2015) and encourage more work that examines the intersection of prosodic structure and the lexical and contextual factors we discuss in this chapter (see, e.g., Baker and Bradlow 2009; Burdin and Clopper 2015; Turnbull et al. 2015; Watson, Arnold, and Tanenhaus 2008).

We further focus in this chapter primarily on phonetic vowel reduction, which involves both temporal (i.e., duration) and spectral (i.e., vowel space peripherality) dimensions, although we also discuss some preliminary findings in the domain of prosodic (i.e., f0 and timing) reduction. The phenomena we discuss are not unique to vowels, however, and we expect our conclusions to be applicable to phonetic reduction in other domains, including consonantal phenomena (see, e.g., Baese-Berk and Goldrick 2009; Bouavichith and Davidson 2013; Goldrick, Vaughn, and Murphy 2013; Warner and Tucker 2011), coarticulatory phenomena (see, e.g., Lin, Beddor, and Coetzee 2014; Scarborough 2013), and other dimensions of prosodic structure in which duration, vowel quality, and f0 play a critical role (see, e.g., Arnold, Kahn, and Pancani 2012; Calhoun 2010a, 2010b; Watson, Arnold, and Tanenhaus 2008).

1 For discussion of the essentially arbitrary division between categorical and continuous aspects of phonetic and phonological structure, see Ladd (2011) and Munson et al. (2010).

2.2 Phonetic reduction in "easy" contexts

The unifying observation in previous work on phonetic reduction processes is that linguistic forms are reduced in "easy" contexts relative to "hard" contexts, where easy and hard are defined with respect to the assumed processing demands imposed by the context on the talker and/or the listener. The linguistic factors that have been shown to contribute to phonetic reduction include lexical properties (e.g., lexical frequency and neighborhood density), contextual properties (e.g., semantic predictability and discourse mention), and speaking style. The definitions of easy and hard contexts for each of these factors are summarized in Table 2.1.

For each of these factors, phonetic reduction is observed in the "easy" contexts relative to the "hard" contexts, although the identification of easy versus hard contexts differs across factors. For the lexical factors, easy and hard contexts are typically defined with respect to the processing demands of the listener and, although the lexical factors are themselves continuous, "easy" and "hard" contexts are typically treated categorically (e.g., Luce and Pisoni 1998; Munson and Solomon 2004; Wright 2004; cf. Baese-Berk and Goldrick 2009; Gahl, Yao, and Johnson 2012). Speaking style is an explicitly listener-oriented manipulation and is also typically defined categorically (Picheny, Durlach, and Braida 1985, 1986). In contrast, for the contextual factors, easy and hard contexts are more often defined with respect to the processing demands of the talker, which are often treated continuously as a reflection of continuous measures of predictability or accessibility (e.g., Bard et al. 2000; Bell et al. 2009; Kahn and Arnold 2012, 2015; cf. Aylett and Turk 2004). Thus, the observed relationship between processing ease and phonetic reduction has been defined in different ways and has been argued to result from a number of different processing mechanisms.

Lexical frequency is typically defined as the number of occurrences of a target word per million words in a corpus of written or spoken language. Early research on speech intelligibility revealed that high-frequency words are easier for listeners to identify than low-frequency words (Broadbent 1967; Howes 1957).

Table 2.1: Linguistic factors contributing to phonetic reduction.

Factor	"Easy"/Reduced	"Hard"/Unreduced
Lexical frequency	High frequency	Low frequency
Neighborhood density	Low density	High density
Semantic predictability	More predictable	Less predictable
Discourse mention	Second mention/given	First mention/new
Speaking style	Plain	Clear

High-frequency words are also produced more quickly than low-frequency words, suggesting an effect of lexical frequency on lexical access and/or motor planning in production (Balota and Chumbley 1985). Thus, high-frequency words exhibit fewer processing demands than low-frequency words for both talkers and listeners. Phonetic reduction is also observed for high-frequency words relative to low-frequency words. This effect of lexical frequency on phonetic reduction has been observed in both the temporal domain for words and vowels (Arnon and Cohen Priva 2013; Aylett and Turk 2004; Bell et al. 2009; Gahl, Yao, and Johnson 2012; Munson and Solomon 2004; Myers and Li 2009; Pate and Goldwater 2011; Pluymaekers, Ernestus, and Baayen 2005b) and the spectral domain for vowels (Munson 2007; Munson and Solomon 2004), and in both isolated word production (Munson 2007; Munson and Solomon 2004; Myers and Li 2009) and continuous speech production (Arnon and Cohen Priva 2013; Aylett and Turk 2004; Bell et al. 2009; Gahl, Yao, and Johnson 2012; Pate and Goldwater 2011; Pluymaekers, Ernestus, and Baayen 2005b).

Lexical neighborhood density is a measure of phonological similarity across words in the lexicon and is typically defined as the number of words that differ from a target word by one phoneme insertion, deletion, or substitution (Luce and Pisoni 1998). Competition during lexical access among phonologically similar words leads to more difficult perceptual identification of words with many phonological neighbors (i.e., high neighborhood density) than for words with few phonological neighbors (i.e., low neighborhood density; Luce and Pisoni 1998; Vitevitch and Luce 1998, 1999). However, the activation of multiple similar word forms leads to faster and less error-prone production for high-density words than low-density words (Vitevitch 2002).[2] Thus, high-density words are difficult to perceive and easy to produce, whereas low-density words are easy to perceive and hard to produce. Consistent with the processing demands exhibited for neighborhood density in perception, phonetic vowel reduction is typically observed in low-density words relative to high-density words in read speech. This effect of neighborhood density on phonetic reduction has been observed for read speech in both the temporal domain for stop consonants (Fox, Reilly, and Blumstein 2015; Peramunage et al. 2011) and the spectral domain for vowels (Clopper and Tamati 2014; Munson 2013; Munson and Solomon 2004), but not in the temporal domain for vowels (Munson and Solomon 2004). Further, in conversational speech, the effect of neighborhood density may be more consistent with the processing demands exhibited for neighborhood density in production: temporal

2 High neighborhood density also facilitates perceptual processing in tasks involving nonwords (Vitevitch and Luce 1998, 1999).

and spectral vowel reduction is observed for high-density words relative to low-density words (Gahl, Yao, and Johnson 2012).

Several composite measures of neighborhood density and lexical frequency have also been developed to provide a single metric to account for the combined effects of these two lexical factors on phonetic reduction. Similar to the results with the simple measures, these composite measures reveal phonetic vowel reduction in the temporal and spectral domains for high-frequency words with few, low-frequency neighbors (i.e., "easy words") relative to low-frequency words with many, high-frequency neighbors (i.e., "hard words"; Munson and Solomon 2004; Scarborough 2010, 2013; Wright 2004). Thus, both individually and in combination, the two lexical factors consistently predict greater phonetic reduction for easy words relative to hard words.

Turning to the contextual factors, semantic predictability captures a range of phenomena related to the syntactic, semantic, and nonlinguistic context that a target word is produced in. Words that are predictable given the preceding sentence context are more intelligible when presented in context than less predictable words (Kalikow, Stevens, and Elliott 1977; Miller and Isard 1963). Similarly, in production, predictable words are less likely to be preceded by a hesitation indicating disfluency than less predictable words (Beattie and Butterworth 1979). Thus, predictable words exhibit fewer processing demands than less predictable words for both talkers and listeners. Words that are predictable in their context also exhibit phonetic reduction relative to words that are less predictable in their context. This effect of semantic predictability on phonetic reduction has been observed in the temporal domain for words and vowels (Aylett and Turk 2006; Bell et al. 2009; Clopper and Pierrehumbert 2008; Engelhardt and Ferreira 2014; Gahl and Garnsey 2004; Hunnicutt 1987; Jurafsky et al. 2001; Lieberman 1963; Moore-Cantwell 2013; Pate and Goldwater 2011; Pluymaekers, Ernestus, and Baayen 2005a; Tily and Kuperman 2012), the spectral domain for vowels (Aylett and Turk 2006; Clopper and Pierrehumbert 2008; Jurafsky et al. 2001), and the prosodic domain for words (Kaland, Swerts, and Krahmer 2013; Wagner and Klassen 2015; Watson, Arnold, and Tanenhaus 2008). These effects of semantic predictability are consistent across a range of measures of predictability, including syllable *n*-gram conditional probabilities (Aylett and Turk 2006), lexical *n*-gram conditional probabilities (Bell et al. 2009; Jurafsky et al. 2001; Pate and Goldwater 2011; Pluymaekers, Ernestus, and Baayen 2005a; Tily and Kuperman 2012), syntactic structure probabilities (Gahl and Garnsey 2004; Moore-Cantwell 2013), cloze probabilities (Clopper and Pierrehumbert 2008; Hunnicutt 1987; Lieberman 1963), information structure (Wagner and Klassen 2015), and nonlinguistic contextual information (Engelhardt and Ferreira 2014; Kaland, Swerts, and Krahmer 2013; Watson, Arnold, and Tanenhaus 2008).

Discourse mention is a contextual factor that captures whether the target word is new or old in the context. Repeated words have real-world referents that are already in the common ground of the conversation and are therefore expected to be easier to access for the talker and the listener than new words that introduce new real-world referents (Chafe 1974; Fowler and Housum 1987). Consistent with this hypothesis, phonetic reduction is observed for easier, second mentions of target words than for harder, first mentions of the same word within a given discourse context.[3] This second mention reduction has been observed primarily in the temporal domain for words and vowels in both read speech (Baker and Bradlow 2009; Fowler 1988) and spontaneous speech (Bard et al. 2000; Fowler and Housum 1987; Galati and Brennan 2010; Kahn and Arnold 2012, 2015; Kaiser, Li, and Holsinger 2011; Lam and Watson 2010, 2014; Pate and Goldwater 2011; Sasisekaran and Munson 2012; Shields and Balota 1991). Thus, for both contextual factors, phonetic reduction is observed for easy (i.e., predictable or given) words relative to hard (i.e., less predictable or new) words.

The final linguistic factor, speaking style, refers to the adoption of a particular mode of speaking that is appropriate for the discourse context and the interlocutors. Style can be explicitly manipulated by the talker and is therefore a potentially different type of linguistic factor contributing to phonetic reduction than the lexical and contextual factors discussed above, which are assumed to be largely implicit. In the context of phonetic reduction research, the primary speaking styles that have been investigated are plain lab speech, which is directed toward an imagined friend, and clear lab speech, which is directed toward an imagined hearing-impaired or nonnative listener.[4] Clear lab speech is more intelligible than plain lab speech (Picheny, Durlach, and Braida 1985), reflecting the talker's explicit adoption of a style that is appropriate for a listener who is assumed to exhibit speech processing difficulties. That is, although clear lab speech is easier to perceive than plain lab speech, it is produced in a context in

3 Second mention reduction may also be linked to other aspects of the communicative domain: Hoetjes et al. (2015) observed that co-speech gesturing that accompanies second mentions tends to be reduced in magnitude relative to gesturing which accompanies first mentions. Similarly, Hoetjes et al. (2012) documented second mention reduction effects in Dutch Sign Language.

4 Speaking style is also a focus of a substantial body of work in variationist sociolinguistics (e.g., Eckert and Rickford 2001) and is therefore related to our discussion below of the effect of social factors, including dialect variation, on phonetic reduction. However, we limit our discussion here to clear and plain lab speech styles because this stylistic variation involves a similar continuum of reduced and enhanced speech as the other linguistic factors described in this section.

which perceptual processing is assumed to be difficult, given the characteristics of the listener. Thus, plain lab speech exhibits phonetic reduction relative to clear lab speech, consistent with the unifying claim across domains that phonetic reduction is observed in easy contexts relative to hard contexts. This speaking style effect on phonetic reduction has been observed in read speech in both the temporal domain for words and vowels (Ferguson and Kewley-Port 2007; Picheny, Durlach, and Braida 1986; Scarborough and Zellou 2013; Smiljanic and Bradlow 2005) and the spectral domain for vowels (Ferguson and Kewley-Port 2007; Moon and Lindblom 1994).

In addition to their individual effects on phonetic reduction, the linguistic factors listed in Table 2.1 have also been shown to have independent effects on phonetic reduction when presented in combination. For example, lexical frequency and neighborhood density exhibit independent effects on spectral vowel reduction (Munson and Solomon 2004), neighborhood density and semantic predictability exhibit independent effects on both temporal and spectral vowel reduction (Scarborough 2010), and neighborhood density and speaking style exhibit independent effects on both temporal and spectral vowel reduction (Scarborough and Zellou 2013). The observed independent phonetic reduction effects across linguistic factors suggest a simple additive system related to processing demands. As processing demands are decreased, phonetic reduction is increased, and vice versa. Thus, high-frequency words that are highly predictable are very easy to process and are therefore more reduced than high-frequency words that are less predictable, which in turn are easier (and more reduced) than low-frequency words that are less predictable.

However, interactions between the various linguistic factors have also been observed, suggesting that phonetic reduction may not simply reflect an additive function of the processing demands imposed by the linguistic context. In particular, Baker and Bradlow (2009) observed a three-way interaction among lexical frequency, discourse mention, and speaking style on temporal reduction, in which high-frequency words exhibited more second mention reduction than low-frequency words in plain lab speech, but not in clear lab speech. Baker and Bradlow (2009) attributed this interaction to a maximal reduction in the easiest context (high-frequency, second mention, plain speech), but a lower bound on the permissible degree of reduction in clear speech that reduces the effects of lexical frequency and discourse mention in that style relative to the effects observed in plain speech. Bell et al. (2009) observed a similar interaction between lexical frequency and semantic predictability, in which high-frequency words exhibited a greater effect of semantic predictability on temporal reduction than low-frequency words. As in the Baker and Bradlow (2009) data, this interaction suggests maximal reduction in the easiest context (high-frequency, high-predictability),

but a lower bound on the permissible degree of reduction for low-frequency and/or low-predictability targets.[5]

Taken together, these interactions suggest that a more complex analysis of the phonetic reduction process may be warranted to account for the potential limits on phonetic reduction in various contexts. The observed interactions suggest lower bounds on reduction in some hard contexts, but lower bounds on reduction in extremely easy contexts are also expected. For example, in the temporal domain, the absolute lower bound on phonetic reduction is deletion. That is, the duration of a linguistic unit (i.e., segment, syllable, or word) cannot be reduced to a value less than 0 ms, which may mean that the combined effects of the various linguistic factors contributing to phonetic reduction cannot be additive because the minimum allowable duration is 0 ms (i.e., deletion). Similarly, in the spectral domain, the lower bound on phonetic vowel reduction is potentially a categorical change to schwa. As noted in the Introduction, we consider segmental alternations and deletions to be categorical phenomena that potentially differ from the continuous, phonetic reduction processes we are focused on in this chapter. However, the possibility of segmental alternations and deletions, as well as their effects on how the various linguistic factors in Table 2.1 must interact in promoting phonetic reduction, must be borne in mind as we consider theoretical models of and further empirical evidence for phonetic reduction processes.

2.3 Theoretical approaches to phonetic reduction

A number of theories have been proposed to capture the insight that phonetic reduction emerges in contexts with limited processing demands. As previewed in the previous section, one of the primary dimensions that differentiates these various theories is whether it is the processing demands for the listener (listener-oriented) or the processing demands for the talker (talker-oriented) that are driving the phonetic reduction process.

5 These findings are only partially consistent with Wright's (2004) predictions about the potential interactions among these factors. In particular, although Wright (2004) predicted maximal reduction of "easy" words (i.e., high-frequency words with few neighbors) in easy contexts, as observed by Baker and Bradlow (2009) and Bell et al. (2009), Wright (2004) also predicted maximal enhancement of "hard" words (i.e., low-frequency words with many neighbors) in hard contexts, which was not observed by either Baker and Bradlow (2009) or Bell et al. (2009).

2.3.1 Listener-oriented approaches

From the listener-oriented perspective, phonetic reduction serves the functional purpose of enhancing communicative success while minimizing talker effort. The underlying assumption of this approach is that some segments or words are more likely to be misperceived by the listener than other segments or words, due to acoustic-perceptual factors (such as masking of acoustic cues in certain phonological contexts) and/or linguistic predictability factors (such as the likelihood of an adjective following a noun). According to the listener-oriented perspective, talkers have a tacit awareness of these potential comprehension difficulties and attempt to enhance the acoustic-phonetic prominence of words that are likely to be difficult for the listener to process. Conversely, the talker is free to phonetically reduce easy words, which are likely to be perceived correctly by the listener. These models assume that it is easier for the talker to produce reduced variants than enhanced variants, leading to reduced variants when the listener's successful perception is likely. The listener-oriented perspective, then, claims that hyperarticulation exists to facilitate successful perception by the listener, and reduction exists to ease the articulatory burden on the talker (see also Brouwer, Mitterer, and Huettig 2013; Mitterer and Russell 2013, for evidence that reduction in appropriate contexts can facilitate perception). Successful communication is therefore central to the listener-oriented account (Jaeger 2013; Ramscar and Baayen 2013).

One of the earliest listener-oriented models was Lindblom's (1990) Hyper- & Hypospeech (H&H) theory, which he argued could account for the observation that segmental realization is affected by a range of contextual factors, including those discussed in the previous section. According to H&H theory, speech is produced along a continuum from hyper- to hypoarticulated as a function of the competing goals of the talker to conserve energy (hypoarticulate) and to be understood (hyperarticulate). Contexts in which lexical access is expected to be easier for the listener lead to phonetic reduction relative to contexts in which lexical access is expected to be more difficult.

Aylett and Turk's (2004; see also Aylett 2000; Aylett and Turk 2006; Turk 2010) smooth signal redundancy hypothesis is very similar in spirit to Lindblom's (1990) H&H theory, in that two competing constraints are argued to be at play: reliable communication and conservation of effort. For communication to be reliable, the signal needs to be clear enough for the message to be transmitted successfully. On the other hand, conservation of effort demands that the talker exert minimal effort to convey the message. Too much focus on reliability and the talker's speech provides unnecessarily redundant information; too much focus on brevity and the talker is not understood. According to Aylett and Turk (2004), redundancy in speech communication is of two kinds. One kind is language

redundancy, which is broadly equivalent to the concept of semantic predictability discussed above. More predictable parts of a message are more redundant than less predictable parts. The other kind of redundancy is acoustic redundancy, which is conceptualized as the likelihood that the signal will be perceived correctly based on the acoustic properties alone. The sum of these two redundancies is the total signal redundancy. Aylett and Turk (2004) proposed that language users strive to ensure that the total signal redundancy is smooth (i.e., constant) throughout an utterance. Thus, the balance between language and signal redundancies accounts for the observed relationships between linguistic factors, such as lexical frequency and semantic predictability, and phonetic reduction. When language redundancy is high, because the target word is highly predictable or frequent, signal redundancy can be low, leading to phonetic reduction.

A number of similar listener-oriented accounts of phonetic reduction have been proposed, which invoke concepts qualitatively similar to smooth signal redundancy, including uniform information density (Jaeger 2010; Levy and Jaeger 2007; Qian and Jaeger 2012), communicative efficiency (van Son and Pols 2003; van Son and van Santen 2005), and Bell's (1984) audience design (Galati and Brennan 2010; Schober 1993). Consistent with the functional underpinnings of the listener-oriented approach, these theories are typically linked to broader claims about the critical role of communication in the functional structure of language (e.g., Hawkins 2014).

2.3.2 Talker-oriented approaches

From the talker-oriented perspective, phonetic reduction arises from interactions in the cognitive architecture of the speech production system. The precise formulation and reasoning behind the process is generally theory specific, but the shared theme of these models is that easy words are accessed or processed more quickly and more easily than hard words, which leads to a faster and less precise (i.e., reduced) production for easy words relative to hard words. By contrast, hard words are accessed or processed less quickly and less easily, resulting in a more effortful and precise (i.e., unreduced) production. From this perspective, the ability of the listener to understand the message plays no direct role in shaping the phonetic realization of the speech signal, and successful communication between interlocutors is therefore not central to the process.

One line of evidence for the talker-oriented approach is research demonstrating that talkers do not always take into account the perspective of their interlocutor, and instead appear to rely on their own perspective, in the implementation of phonetic reduction processes. For example, Bard et al. (2000) conducted an

investigation of second mention reduction in the HCRC map task corpus (Anderson et al. 1991). In the materials of interest in Bard et al.'s (2000) study, after the instruction-giver had finished guiding their partner through a map, their partner changed and they had to guide the new partner through the same map. All of the mentions of the landmarks were, in this context, discourse-given from the perspective of the instruction-giver, but discourse-new from the perspective of the partner being led. Bard et al. (2000) found that the productions in this second trial with the new partner were both shorter in duration and less intelligible in isolation than the productions from the first trial, suggesting that second mention reduction had taken place. That is, despite the instruction-giver's awareness that their interlocutor had changed and was therefore not familiar with the discourse context, the instruction-giver still reduced tokens that were discourse-given from the instruction-giver's perspective. Bard et al. (2000) interpreted this finding as evidence of an egocentric pattern of phonetic reduction, in which the talker's situational knowledge assumes primacy over their modeling of the listener's knowledge (see also Keysar 2008; Keysar and Barr 2005; Keysar et al. 2000).

More recently, Baese-Berk and Goldrick (2009) carried out a series of experiments in which a participant instructed their partner to click on an item on a computer display. Both the instructor and the partner saw the same display of items. In a condition where two of the displayed items were referents of a voice-onset-time (VOT) minimal pair (e.g., *cod* and *god*), more extreme VOT values were observed on the target word relative to a condition in which the target word did not have a real-word minimal pair competitor (e.g., *cog*, where *gog* is not a real word in English). This phonetic enhancement of the aspiration contrast in a potentially ambiguous context is consistent with a listener-oriented perspective. However, when the same target item *cod* was displayed without its minimal pair competitor, VOT enhancement was still observed, albeit to a smaller degree. Baese-Berk and Goldrick (2009) argued that this VOT enhancement in an unambiguous context cannot be accounted for by a listener-oriented perspective and suggested instead that lexical competition in production drives the enhancement effect. Various replications of Baese-Berk and Goldrick's (2009) findings in situations that do not involve a communicative partner (Bullock-Rest et al. 2013; Fox, Reilly, and Blumstein 2015; Kirov and Wilson 2012; Peramunage et al. 2011) provide further evidence against a purely communicative account of the phenomenon. In particular, in the absence of a live interlocutor, the communicative imperative to speak clearly to distinguish minimal pair targets is arguably absent.

However, this line of argumentation suggests that most of the data presented in the previous section should be taken as evidence for a talker-oriented approach to phonetic reduction. In particular, the effects of lexical frequency, neighborhood density, semantic predictability, and discourse mention described

above are all observed in the absence of a live interlocutor. If real-time communication is required for listener-oriented adjustments, such adjustments should not be observed in laboratory settings without an immediate communicative task. That is, if phonetic reduction reflects an adjustment for the listener, phonetic reduction should not be observed when a listener is not physically present. However, numerous studies have shown that talkers can make explicit speaking style adjustments for imagined interlocutors in this kind of noncommunicative laboratory setting (e.g., Ferguson and Kewley-Port 2007; Picheny, Durlach, and Braida 1986; Smiljanic and Bradlow 2005), suggesting that real-time communication is not necessary for listener-oriented adjustments to take place. Furthermore, adjustments in duration and vowel space size can be comparable to those that are produced when a live interlocutor is present, although other processes such as coarticulation and speaking rate show significant effects of real versus imagined interlocutors (Scarborough et al. 2007; Scarborough and Zellou 2013). Although participants in many laboratory studies are not talking to another person, they are producing speech in a laboratory setting and are aware that their speech is being recorded, implying that someone (e.g., the researcher or participants in a future study) will eventually listen to their speech (see also Wagner, Trouvain, and Zimmerer 2015). Thus, recorded speech in a laboratory is not comparable to true self-directed speech with no communicative intent, and the lack of an explicit communicative context may not be sufficient to negate a listener-oriented interpretation of phonetic reduction processes.

2.3.3 Evolutionary approaches

In addition to the listener-oriented and talker-oriented perspectives, a third explanation for the relationship between processing demands and phonetic reduction has been proposed. This set of theories differs from the previous two perspectives in holding that no active force is responsible for phonetic reduction. Specifically, rather than communicative pressure or cognitive architecture producing these effects, phonetic reduction simply exists as a natural consequence of patterns of language acquisition and change over generations. Segments or words that are easy to perceive are generally perceived correctly, whereas segments or words that are difficult to perceive are only perceived correctly if they are sufficiently acoustically prominent. Over time, segments and words that are perceived correctly (i.e., easy words and acoustically prominent hard words) become the principal component of language; all other modes of production fall into disuse (Garrett and Johnson 2012; Pierrehumbert 2001a, 2002; Silverman 2012). Silverman (2012, p. 147) expressed this position as follows (emphasis in original):

> Successful speech propagates; unsuccessful speech does not. Confusing speech tokens may be misunderstood, and thus not pooled with the exemplars of the intended word, and so the system maintains its state of semantic clarity. Anti-homophony is thus not an *active* pressure for which there is an abundance of overt evidence. Rather, it is a *passive* result of the pressures that inherently act upon the interlocutionary process.

One of the few explicit formulations of this perspective comes from Pierrehumbert's (2001a, 2001b, 2002, 2003a, 2003b) work on exemplar-based phonology. Her description involves an exemplar model (Goldinger 1998; Johnson 1997; Tenpenny, 1995) in which each perceived word token has its own representation in a perceptual cloud (see also Blevins and Wedel 2009; Tupper 2014; Wedel 2006, for refinements and extensions of these mechanisms). In Pierrehumbert's (2002) model, phonetic reduction effects emerge as a simple consequence of the acquisition process. In particular, when a high-frequency word is uttered, the listener can guess the word's identity with relative ease, even if it is not acoustically prominent, due to its high frequency. When the word is identified, the token is added to the listener's exemplar cloud and becomes part of that word's representation. However, when a low-frequency word is uttered, the listener cannot as easily guess the word's identity (due to its low frequency), and the token therefore needs to be more acoustically prominent than the high-frequency word for its identity to be ascertained correctly. When the word is not correctly identified, the token is not added to the listener's exemplar cloud and does not become part of the target word's representation. Thus, the low-frequency word token will only be added to the exemplar space if it is sufficiently acoustically prominent (Tupper 2014). Over time, then, the exemplar space will contain acoustically prominent low-frequency words, and both prominent and nonprominent high-frequency words. In speech production, the talker selects a token at random from the exemplar space of the target word (see Pierrehumbert 2001a, 2002, for mathematical details of the implementation). High-frequency words will tend to be reduced in production relative to low-frequency words because their exemplar clouds contain both reduced and unreduced variants, whereas the exemplar clouds of the low-frequency words contain primarily unreduced variants, leading to unreduced productions of these targets. Within a speech community, this behavior facilitates a positive feedback loop leading to clear productions of low-frequency words and reduced productions of high-frequency words.

Additional support for this evolutionary perspective comes from animal behavior research, suggesting that nonhuman animal communication systems are structured to allow for maximal information transmission with minimal effort (see, e.g., Bezerra et al. 2010; Semple, Hsu, and Agoramoorthy 2010; Semple et al. 2013, on primates and Luo et al. 2013, on bats). Ferrer-i-Cancho et al. (2013) explicitly argued that all communication systems, including human language, are

governed by basic distributional properties that enhance efficiency of coding. For a communication system to persist, successful communication with the lowest possible energy expenditure is necessary (see Ferrer-i-Cancho and Elvevåg 2010; Ferrer-i-Cancho and Moscoso del Prado 2011, for statistical approaches to this reasoning). Under this view, the observed linguistic effects on phonetic reduction are a necessary consequence of natural selection and no appeal to cognitive or psychological mechanisms is needed.

2.4 Complexifying our understanding of phonetic reduction

The listener-oriented, talker-oriented, and evolutionary approaches differ considerably in their assumptions regarding the root cause of phonetic reduction. However, these models share the assumption that one underlying factor (e.g., cognitive processing demands) drives the phonetic reduction effects that are observed across temporal and spectral acoustic domains and across lexical, contextual, and stylistic contexts. However, recent research in our laboratory has revealed variation in phonetic reduction processes, suggesting that a simple relationship between processing demands and phonetic reduction may not be sufficient to account for these various effects on segmental realization. In particular, we have observed complex interactions among linguistic factors, social factors, and cognitive factors in temporal, spectral, and prosodic phonetic reduction processes. These interactions reveal variation in the robustness of the linguistic effects on phonetic reduction, as well as different patterns of interactions among linguistic, social, and cognitive factors in temporal and spectral reduction, suggesting diverse phonetic reduction processes across acoustic domains. These findings challenge the notion of a simple linear mapping between phonetic reduction and processing difficulty.

2.4.1 Interactions among linguistic factors

One component of our recent research on phonetic reduction has explored interactions among linguistic factors on temporal and spectral vowel reduction. This work builds on previous research demonstrating both independent and interactive effects of these factors on temporal and spectral reduction (e.g., Baker and Bradlow 2009; Bell et al. 2009; Munson and Solomon 2004; Scarborough 2010; Scarborough and Zellou 2013), and extends the analysis to consider the relationships among more factors simultaneously.

We conducted a large experiment to explore phonetic reduction in read passages in which we manipulated lexical frequency, lexical neighborhood density, semantic predictability, discourse mention, and speaking style in a fully crossed design (Burdin, Turnbull, and Clopper 2015; Clopper, Turnbull, and Burdin in press). The materials were a set of short stories read by Midwestern undergraduates and containing target words with the stressed vowels /i, ɛ, æ, ɑ, ɔ, u/. The target words varied in lexical frequency (as presented in the Hoosier Mental Lexicon; Nusbaum, Pisoni, and Davis 1984), lexical neighborhood density (as presented in the Hoosier Mental Lexicon; Nusbaum, Pisoni, and Davis 1984), and semantic predictability (as assessed by an independent cloze task with Midwestern undergraduates). Each word was included twice in the same story to elicit discourse mention effects. The complete set of stories was read twice by each talker – first to an imagined friend and then again to an imagined hearing-impaired or nonnative listener – to elicit plain and clear lab speech, respectively. For each target word in each story, vowel duration and vowel dispersion, defined as the Euclidean distance from the center of the F1 × F2 vowel space in Bark, for the primary stressed vowel were obtained.

Mixed-effects regression models predicting vowel dispersion from the five linguistic factors (lexical frequency, lexical neighborhood density, semantic predictability, discourse mention, and speaking style) and their interactions revealed the expected main effects of lexical frequency, discourse mention, and speaking style, as well as a four-way interaction between lexical frequency, lexical neighborhood density, semantic predictability, and discourse mention. None of the other main effects or interactions were significant for the vowel dispersion measure. As shown in the top left panel of Figure 2.1, vowels in high-frequency words exhibited less dispersion in the vowel space than vowels in low-frequency words, and vowels in plain speech exhibited less dispersion in the vowel space than vowels in clear speech. Vowels in second mention words also exhibited less dispersion in the vowel space than vowels in first mention words (2.08 vs. 2.14 Bark, respectively). The four-way interaction further revealed effects of lexical neighborhood density and semantic predictability in the expected directions: vowels in low-density words exhibited less dispersion than vowels in high-density words and vowels in high-predictability words exhibited less dispersion than vowels in low-predictability words. Unlike the main effects of lexical frequency, discourse mention, and speaking style, however, these effects of lexical neighborhood density and semantic predictability were more variable across contexts, and thus, significant main effects did not emerge.

Although previous research has not examined vowel dispersion as a function of discourse mention, significant effects of lexical neighborhood density and semantic predictability on vowel dispersion have been reported in previous work (e.g., Aylett and Turk 2006; Clopper and Pierrehumbert 2008; Jurafsky et al. 2001; Munson and

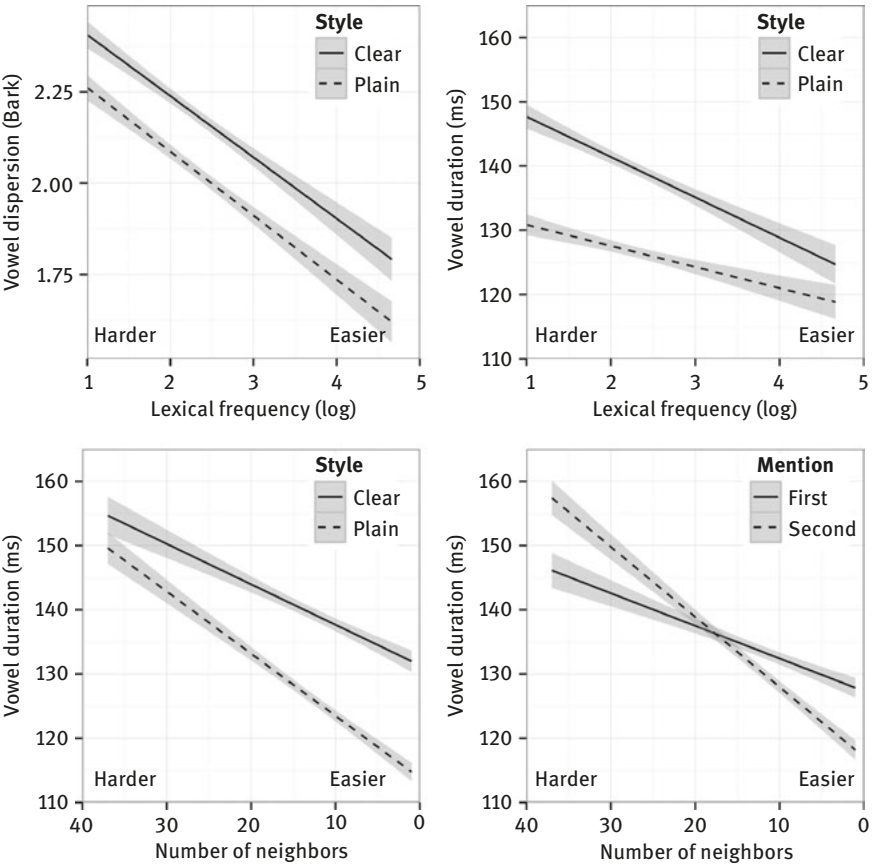

Figure 2.1: Lexical frequency (log occurrences per million words) and speaking style effects on vowel dispersion, defined as the Euclidean distance from the center of the F1 × F2 Bark space (top left), lexical frequency and speaking style effects on vowel duration (top right), lexical neighborhood density and speaking style effects on vowel duration (bottom left), and lexical neighborhood density and discourse mention effects on vowel duration (bottom right) in read short stories. Adapted from Burdin, Turnbull, and Clopper (2015).

Solomon 2004). Thus, the lack of significant main effects of lexical neighborhood density and semantic predictability on vowel dispersion in Burdin, Turnbull, and Clopper's (2015) study is somewhat surprising. This null result may reflect variability in these effects across vowel categories (see, e.g., Clopper and Pierrehumbert 2008; Scarborough 2010; Wright 2004) or the relative sizes of the effects. Munson and Solomon (2004) reported a much larger effect size for lexical frequency than lexical neighborhood density on vowel space dispersion and Clopper et al. (2017) reported more robust effects of speaking style than neighborhood density or discourse mention on vowel space dispersion (see below). Thus, the significant effects

of lexical frequency and speaking style in Burdin, Turnbull, and Clopper's (2015) study may have swamped any smaller effects of the other factors.

Mixed-effects regression models predicting vowel duration from the five linguistic factors and their interactions also revealed the expected main effects of lexical frequency and speaking style. As shown in the top right panel of Figure 2.1, high-frequency words had shorter vowels than low-frequency words and vowels in plain speech were shorter than vowels in clear speech. None of the other main effects were significant, although a number of significant interactions were observed for vowel duration. As shown in the top right panel of Figure 2.1, lexical frequency and speaking style interacted such that the lexical frequency effect was larger in clear speech than in plain speech. This pattern of interaction contrasts with the interaction observed for lexical neighborhood density and speaking style, shown in the bottom left panel of Figure 2.1, in which the lexical neighborhood density effect was larger in plain speech than in clear speech. The interaction between lexical neighborhood density and discourse mention, shown in the bottom right panel of Figure 2.1, parallels the neighborhood density × speaking style interaction and reveals a larger lexical neighborhood density effect for second mentions than for first mentions.[6] Thus, consistent with previous findings in which the effects of lexical frequency and discourse mention were greater in plain speech than in clear speech (Baker and Bradlow 2009), we find evidence for a larger effect of lexical neighborhood density in plain speech than in clear speech and for second mentions than for first mentions. These results are consistent with the maximization of temporal reduction in easier (i.e., low-density, second mention, plain speech) contexts relative to harder (i.e., high-density, first mention, clear speech) contexts. In contrast, the interaction between lexical frequency and speaking style is not consistent with this interpretation and may reflect a lower bound on temporal reduction in easy contexts. That is, high-frequency words in plain speech may not be maximally reduced because further temporal reduction would lead to deletion.

Three findings emerge from these results that suggest that phonetic reduction may reflect a more complex process than simple additive effects of processing difficulty. The first of these findings is that the patterns of phonetic reduction differ

6 This interaction between lexical neighborhood density and discourse mention also exhibits a cross-over effect, suggesting that first mentions were phonetically reduced relative to second mentions for words with many phonological neighbors. This apparent reversal of the discourse mention effect for words with many neighbors may reflect other factors contributing to vowel duration, including prosodic structure (see Burdin and Clopper 2015) or information structure (see, e.g., Wagner and Klassen 2015).

across acoustic domains. Although most previous research has focused on the temporal reduction of words and vowels in easy contexts relative to hard contexts (e.g., Aylett and Turk 2004; Bell et al. 2009; Fowler and Housum 1987; Gahl, Yao, and Johnson 2012), studies that have examined spectral reduction have typically observed spectral reduction in the same easy contexts in which temporal reduction is typically observed (e.g., Aylett and Turk 2006; Clopper and Pierrehumbert 2008; Munson and Solomon 2004; Scarborough 2010, 2013; Wright 2004). However, Burdin, Turnbull, and Clopper's (2015) results reveal comparable main effects of lexical frequency and speaking style on temporal and spectral vowel reduction, but different patterns of interactions among these and other linguistic factors in the two acoustic domains, suggesting that temporal and spectral reduction exhibit different linguistic constraints and may arise from different processes associated with processing difficulty.

Further evidence for differences in phonetic reduction across acoustic domains comes from Turnbull's (2017) analysis of data obtained in an experiment conducted by Ito and Speer (2006). This experiment involved a naïve participant instructing a confederate in the decoration of a Christmas tree. The type of ornament to be hung and its location on the tree were presented to the participant on a computer screen, but no explicit instructions were provided about how to phrase the instructions to the confederate. Thus, the speech elicited in this task was truly spontaneous and interactive. Ornaments varied in both color and shape, necessitating their description as adjective-noun phrases, like *blue drum*. The target words were coded for discourse mention as either the first or a subsequent mention in the decoration of the Christmas tree. The effect of discourse mention on vowel duration and peak f0 of the stressed syllables of the target adjectives and nouns were examined, after controlling for variation in phonological pitch accent type. As shown in Figure 2.2, reduction in duration for subsequent mentions was observed relative to first mentions, consistent with prior research (Baker and Bradlow 2009; Fowler and Housum 1987). However, no effect of discourse mention was observed for peak f0, suggesting that second mention reduction may not extend to the domain of intonation.

The second critical finding from Burdin, Turnbull, and Clopper's (2015) results is that the interactions observed for temporal reduction suggest both maximization of reduction in some easy contexts (e.g., low-density, plain speech), as suggested in previous research (Baker and Bradlow 2009; Bell et al. 2009), and a lower bound on reduction in other easy contexts (e.g., high-frequency, plain speech). Additional evidence from our laboratory for a lower bound on reduction comes from a recent analysis of segmental deletion in interview speech (Turnbull 2015a, in press). Previous studies of segmental deletion in interview speech have revealed widespread deletion in words of all sizes (Johnson 2004), as well as more frequent

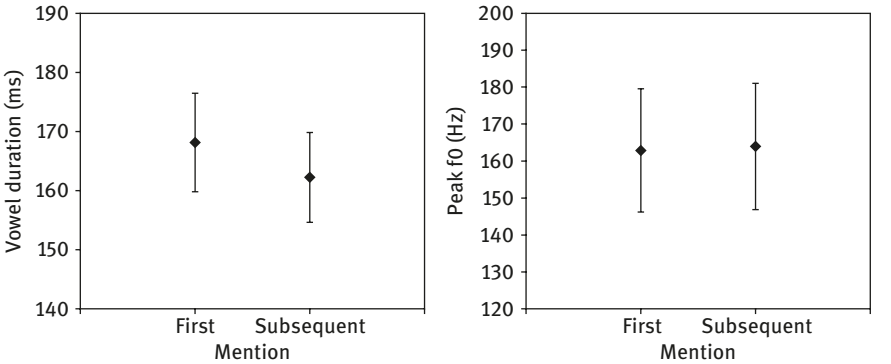

Figure 2.2: Effect of discourse mention on mean vowel duration (left) and peak f0 (right) in Ito and Speer's (2006) Christmas tree decorating task. Error bars are standard error of talker means. Adapted from Turnbull (2017).

/t, d/ deletion in easy words (i.e., high-frequency or high-predictability words) than in hard words (i.e., low-frequency or low-predictability words; Raymond, Dautricourt, and Hume 2006; Jurafsky et al. 2001). Turnbull's (2015a) analysis of the Buckeye Corpus of Conversational Speech (Pitt et al. 2007), which is a phonetically aligned corpus of approximately 40 hours of spontaneous, interview speech, extended this previous work and considered the roles of lexical frequency and lexical neighborhood density on segmental deletion. Each word in the Buckeye Corpus is tagged with both a phonemic (dictionary) transcription and a phonetic (narrow) transcription. By comparing these transcriptions, the number of deleted phones in each of the 282,435 word tokens in the corpus was determined.

A mixed-effects Poisson regression model predicting the number of deleted phones from the target's lexical frequency, lexical neighborhood density, and the number of phonemes in the target's citation form revealed the expected effect of number of phonemes – longer words tended to exhibit more deletions than shorter words, because shorter words can only delete so many phonemes before the word is unintelligible. The expected effects of lexical frequency and lexical neighborhood density were also observed. Harder words in denser neighborhoods tended to have fewer deleted phones than easier words in sparser neighborhoods and harder, less frequent words tended to have fewer deleted phones than easier, more frequent words. However, as shown in Figure 2.3, these two factors interacted such that lexical frequency effects were observed for words in denser neighborhoods (i.e., more than 3 neighbors), but no effect of lexical frequency was observed for words in sparser neighborhoods (i.e., 0–3 neighbors). These low-density words, regardless of lexical frequency, exhibited a high mean phone deletion rate of just over 0.6 phones per word. Thus, the easy, low-density

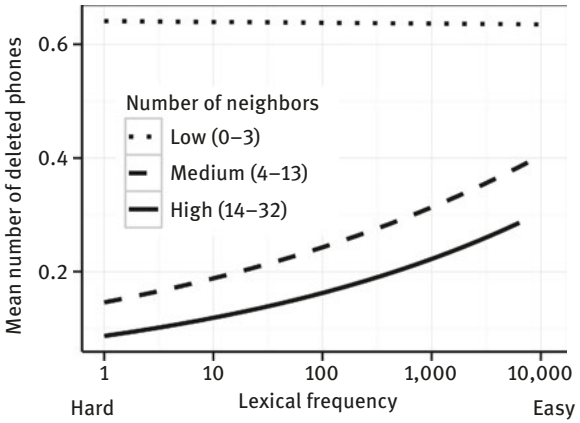

Figure 2.3: Effects of lexical frequency (number of occurrences in the Buckeye Corpus) and lexical neighborhood density on the mean number of deleted phones across words in the Buckeye Corpus. Adapted from Turnbull (2015a).

words exhibit an upper bound on deletion that is comparable to the lower bound on temporal reduction observed for the high-frequency words produced in plain speech in Burdin, Turnbull, and Clopper's (2015) study. That is, low-density words exhibit the maximal number of deleted phonemes and high-frequency words exhibit the minimal vowel duration that the production system allows. This parallel in bounds on reduction suggests a strong connection between the phenomena that we have characterized as categorical versus continuous (see also Cohen Priva 2015), further suggesting that such a distinction is ultimately arbitrary.

The third critical finding from the Burdin, Turnbull, and Clopper (2015) study is that cloze predictability was unexpectedly not a significant independent predictor of either temporal or spectral reduction. As noted above, this null result may reflect variability in the magnitude of the effect across vowel categories or a relatively small effect size. However, in a series of recent studies, we have explored alternative dimensions of semantic predictability and their relative contributions to phonetic reduction in the temporal and prosodic domains. In one of these studies, Turnbull (2017) analyzed a set of data from an experimental investigation of focus marking in English to explore potential effects of semantic predictability on the realization of word duration and peak f0. Crucially, as in the analysis of the Christmas tree data described above, this analysis took phonological pitch accenting into account, so the results cannot be reduced to phonological effects of accent choice, but rather must be attributed to adjustments at the phonetic level. The data set was drawn from an experiment conducted by Turnbull et al. (2015) and Burdin, Phillips-Bourass et al. (2015), which featured a naïve participant instructing a confederate in an object-placing task. The task involved placing

tiles depicting colored objects into numbered boxes on a game board. The objects depicted on the tiles were differentiated in both color and shape, and the participants' instructions were of the form "put the ADJECTIVE NOUN in box NUMBER." The order of the tiles to be placed was manipulated to elicit focus on different linguistic expressions across utterances. Following Rooth (1992), we consider focus to be a semantic property denoting a set of alternatives to the asserted content, not a phonological prosodic property of the utterance. Thus, for example, in the sequence *green lion … blue lion*, the adjective *blue* is focused as a contextually relevant alternative to *green*, whereas in *blue train … blue lion*, the noun *lion* is focused as a contextually relevant alternative to *train*. The set of available tiles was finite and visually salient to both interlocutors, which meant that, as more tiles were played, the individual probability of any one tile being played increased.

Turnbull's (2017) analysis of these data revealed an inverse relationship between peak f0 and utterance probability given the number of remaining available tiles, as shown in Figure 2.4. This result extends previously established effects of probability on duration (Aylett and Turk 2004) to the f0 dimension. Turnbull (2017) also observed an effect of utterance probability on word duration, such that more probable items were produced with a shorter duration, as expected, but this effect held only for nonfocused nouns. An effect of utterance probability on word duration was not observed for focused nouns or for any adjectives. This result suggests that semantic predictability and focus can interact, with focus essentially "blocking" temporal effects of predictability. This interaction is similar to the interaction that Baker and Bradlow (2009) observed in which the effects of lexical frequency and discourse mention were reduced (or "blocked") in clear speech relative to plain speech.

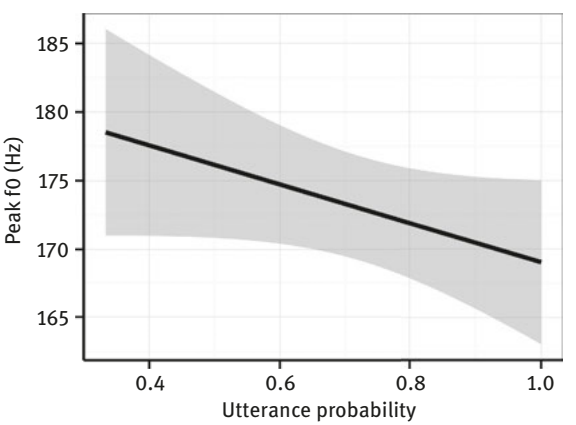

Figure 2.4: Effect of utterance probability on syllable peak f0. Adapted from Turnbull (2017).

A final relevant manipulation in Turnbull et al.'s (2015) study was that the constituent in the instructions that would be focused was either predictable or unpredictable from the global context of each game board. For example, one game board consisted of all red tiles, in which case the noun in each instruction was focused (*red* LION, *red* TRAIN). In other boards, the focused constituent was not predictable from the global context, and which constituent was focused changed from utterance to utterance. The hypothesis under investigation was that phonetic cues to focus, such as word duration and peak f0, would be less prominent in the predictable condition than in the unpredictable condition, due to the contribution of the context to the listener's interpretation of the utterance. The analyses presented by both Turnbull (2017) and Turnbull et al. (2015) found support for this hypothesis. Differences in word duration and peak f0 were larger across focus conditions in the unpredictable condition than in the predictable condition, independent of phonological pitch accent type or phrasing. As shown in the top two panels of Figure 2.5, the effect of context was more robust for peak f0 than for word duration, which was more variable within and across conditions. However, taken together, the results demonstrate that when the context provides information about the relevant semantic contrasts, the talker produces smaller prosodic cues to indicate those semantic contrasts, consistent with Aylett and Turk's (2004, 2006) proposal for a trade-off between language and acoustic redundancies to produce a constant signal redundancy.

The results of Turnbull's (2017) study therefore provide further evidence for variation in phonetic reduction across acoustic domains, as well as evidence for variation in phonetic reduction across different dimensions of semantic predictability. Whereas context condition (predictable vs. unpredictable) exhibited consistent effects across acoustic domains (word duration and peak f0), utterance probability exhibited a robust effect only in the f0 domain. Thus, different dimensions of semantic predictability reveal different phonetic reduction patterns within the same data set. Further, Turnbull's (2017) analysis of utterance probability revealed complex interactions between semantic predictability and other linguistic factors (i.e., focus and word class), which exhibit independent prosodic effects of pitch-accenting and phrasing on phonetic prominence. Given these complex patterns of interaction, the analysis of phonetic reduction must involve careful consideration of all potentially relevant linguistic and contextual factors that contribute to the realization of acoustic-phonetic prominence.

To explore the cross-linguistic generalizability of the American English findings, Turnbull et al. (2015) also examined data from Paraguayan Guaraní in the same tile-placing game that was used with the American English participants. Paraguayan Guaraní has a similar overall prosodic structure to American English, including lexical stress and phrase-level prominences realized through

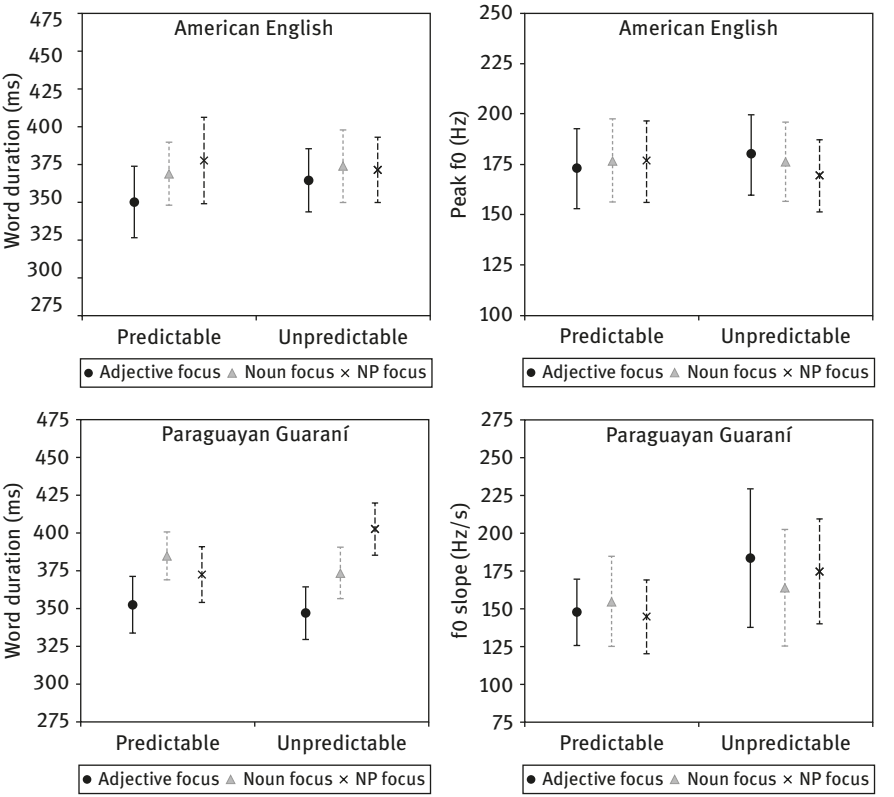

Figure 2.5: Mean word duration (left panels) and f0 prominence (right panels) of adjectives and nouns in American English (top panels) and Paraguayan Guaraní (bottom panels) noun phrases as a function of the focused expression in the noun phrase (adjective, noun, or noun phrase) and experimental context (predictable or unpredictable). Error bars are standard error of talker means. Adapted from Turnbull et al. (2015).

pitch accenting, but differs in the size of its pitch accent inventory (two in Paraguayan Guaraní vs. five in American English) and the number of levels of prosodic phrasing above the word (one in Paraguayan Guaraní vs. two in American English; see also Turnbull et al. 2015; Burdin, Phillips-Bourass et al. 2015). The similarities in the overall prosodic structure allow for a meaningful comparison across languages, whereas the differences allow variation in the realization of contextual effects to emerge. As shown in the bottom two panels of Figure 2.5, word duration in Paraguayan Guaraní was affected by focus condition, with shorter words in adjective focus and longer words in noun and noun phrase focus, independent of pitch accent type and phrasing, but word duration was not significantly affected by context or its interaction with focus condition. However,

context had a significant effect on the f0 slope of the Paraguayan Guaraní pitch accents, independent of phonological pitch accent type. The slope of the pitch accents was steeper when the focused expression was not predictable from the context relative to when the focused expression was predictable from the context. Thus, although both American English and Paraguayan Guaraní exhibit a pattern that can be interpreted as prosodic reduction in an easier (i.e., more predictable) context, the patterns differ considerably across languages. In Guaraní, prosodic prominence was globally reduced through shallower f0 slopes in the easier predictable context relative to the harder unpredictable context and these effects of context did not interact with focus.

Taken together, the recent findings in our laboratory suggest substantial variation in phonetic reduction across acoustic domains and across linguistic factors. Phonetic reduction is most robust in the temporal domain and effects are more variable in the spectral and prosodic domains. These differences across acoustic domains may reflect the relative contributions of these sources of information to phonological contrasts in English. Whereas vowel spectral information and f0 information are critical for distinguishing vowel quality and intonational contrasts, respectively, duration plays a relatively minor role in distinguishing vowel contrasts and may therefore be available for conveying other lexical or contextual information. With respect to linguistic factors, we have observed variation in the strength of phonetic reduction effects across dimensions of semantic predictability, as well as different patterns of interactions among the linguistic factors that contribute to phonetic reduction and between those factors and other factors that contribute to variation in acoustic-phonetic prominence. Segmental duration in particular is shaped by numerous linguistic and contextual factors (Klatt 1976) and these factors must be considered when phonetic reduction is examined. Finally, our results from Paraguayan Guaraní suggest that phonetic reduction processes may also differ in their implementation across languages.

We interpret these results as strong evidence that a simple dichotomy between easy and hard processing contexts, such as that presented in Table 2.1, is insufficient to account for phonetic reduction patterns within or across languages. Minimally, the distinction between easy and hard contexts must be elaborated to account for the observed variability in effect sizes across acoustic domains and linguistic factors. For example, processing demands could be conceptualized as a continuum from easy to hard, with different linguistic factors covering different ranges of the continuum or exhibiting different constraints on their possible realization along the continuum. This idea is consistent with Baker and Bradlow's (2009) proposal for a lower bound on phonetic reduction in clear speech: if clear speech is constrained to a particular range of the "hard" end of the processing demands continuum, the combined effects of other linguistic factors contributing

to phonetic reduction may not lead to as much reduction in clear speech as in plain speech if plain speech has fewer constraints on its possible range. Similarly, the permissible range of variation may vary across acoustic domains, so that larger differences in processing demands are required for phonetic reduction effects to emerge in spectral or prosodic domains than in the temporal domain.

The adoption of a gradient, nonbinary interpretation of processing difficulty is relatively trivial and not at odds with any of the previous work in this area. As noted above, many of the linguistic factors are themselves continuous variables (e.g., lexical frequency, lexical neighborhood density, some measures of semantic predictability) or could straightforwardly be transformed to ordinal (e.g., style) or numerical (e.g., discourse mention) variables. The nature of the nonlinear relationships will be more difficult to determine, but primarily requires substantially more data from production and perception to allow us to characterize not only the nature of the processing difficulties imposed by each of the relevant factors, but also the magnitude of phonetic reduction effects for each of the relevant factors in various combinations across acoustic domains. Thus, the next stage of research in this area will require us to untangle the nonlinear relationships among these numerous continuous variables. Understanding these nonlinear relationships is an essential first step toward determining how much of phonetic reduction can be accounted for by this proposed elaboration of the processing demands explanation.

2.4.2 Interactions between linguistic factors and dialect variation

A second component of our recent research on phonetic reduction has explored the interactions between dialect variation and the linguistic factors contributing to phonetic reduction. A small, but growing, literature suggests that talkers produce more marked social information in easy processing contexts relative to hard processing contexts. For example, Oprah Winfrey, an African-American talk-show host, produces more African-American features for easy, high-frequency words than for harder, low-frequency words (Hay, Jannedy, and Mendoza-Denton 1999). Similarly, gender differences are more pronounced in easy, low-density words than in harder, high-density words (Munson 2007; see also Scarborough 2010, and commentary by Flemming 2010, suggesting more extreme dialect-specific variants are produced in low-density words than high-density words).

In a series of recent studies (Clopper, Mitsch, and Tamati 2017; Clopper and Pierrehumbert 2008; Clopper and Tamati 2014; Turnbull and Clopper 2013),

we have confirmed this general observation that more extreme dialect variants are observed in easier (i.e., low-density, high-predictability, second mention, plain speech) contexts than in harder (i.e., high-density, low-predictability, first mention, clear speech) contexts. However, a closer inspection of the results of these studies reveals variation in the interactions between dialect variation and linguistic factors across vowels and across acoustic domains (see Figures 2.6–2.9).

First, in an investigation of the effect of lexical neighborhood density on vowel reduction and dialect-specific variants in the Midland and Northern dialects of American English (Clopper, Mitsch, and Tamati 2017), more extreme dialect-specific variants were observed consistently for low-density words relative to high-density words in the spectral domain, but dialect differences were enhanced in the temporal domain only for /i/. In particular, although no effect of lexical neighborhood density on vowel duration was observed (see the left panel of Figure 2.6), the Northern vowels were longer than Midland vowels overall and this difference was exaggerated for easy, low-density /i/ words relative to hard, high-density /i/ words. In the spectral domain, we observed more extreme dialect-specific variants, including raising and fronting of /æ/ by the Northern talkers and fronting of /u/ for both talker dialects, in the easy, low-density words than in the hard, high-density words, as shown in the right panel of Figure 2.6. Thus, in the spectral domain, the observation that dialect information is marked more strongly in easy contexts relative to hard contexts was robust across vowel categories, but in the temporal domain, lexical neighborhood density interacted with dialect variation only for one of the four vowels examined.

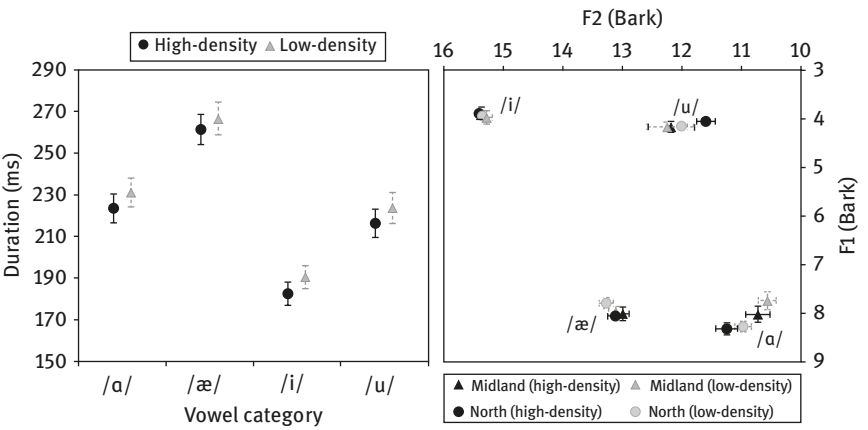

Figure 2.6: Effects of lexical neighborhood density on mean vowel duration (left) and mean vowel formant frequencies (right). Error bars show standard error of talker means. Adapted from Clopper et al. (2017).

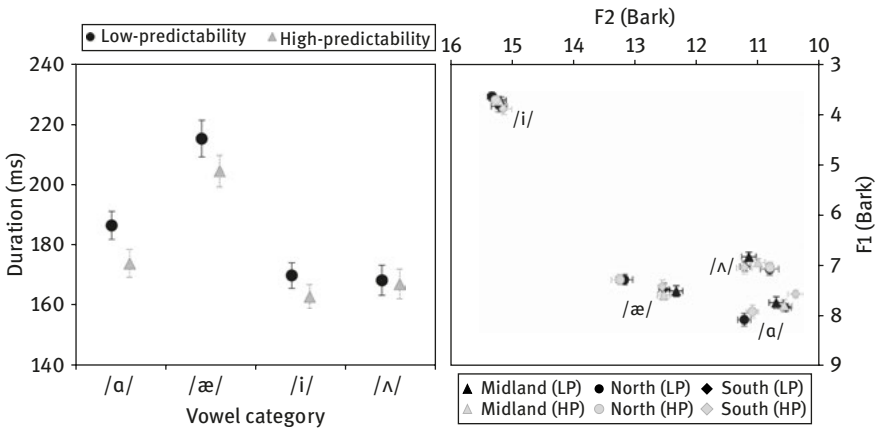

Figure 2.7: Effects of semantic predictability on mean vowel duration (left) and mean vowel formant frequencies (right). Error bars show standard error of talker means. Adapted from Clopper and Pierrehumbert (2008).

Second, in an investigation of the effect of semantic predictability on vowel reduction and dialect-specific variants in the Midland, Northern, and Southern dialects of American English (Clopper and Pierrehumbert 2008), semantic predictability did not interact with dialect variation in the temporal domain and interacted with dialect variation for only one vowel in the spectral domain. In the temporal domain, we observed the expected effect of semantic predictability for three of four vowels (/i, æ, ɑ/, but not /ʌ/), as shown in the left panel of Figure 2.7. Vowels were shorter in easy, high-predictability words than in harder, low-predictability words. In the spectral domain, we observed the expected effect of semantic predictability on vowel dispersion for the Southern talkers for /i, æ, ɑ/ and for the Northern talkers for /i, ɑ/, as shown in the right panel of Figure 2.7. Vowels were less dispersed in the vowel space in easy, high-predictability words than in harder, low-predictability words. In addition, as shown in the right panel of Figure 2.7, we observed greater dialect-specific fronting of /æ/ for the Northern talkers in the easy, high-predictability context relative to the hard, low-predictability context. Thus, for semantic predictability, the interaction between phonetic reduction and dialect variation processes is limited to the spectral domain and to one of the four vowels we examined.

Third, in an investigation of the effect of discourse mention on vowel reduction and dialect-specific variants in the Midland and Northern dialects of American English (Clopper, Mitsch, and Tamati 2017), discourse mention did not interact with dialect variation in the temporal domain and interacted with dialect variation for only one vowel in the spectral domain. In the temporal domain, we observed the expected effect of discourse mention for three of the four vowels (/æ,

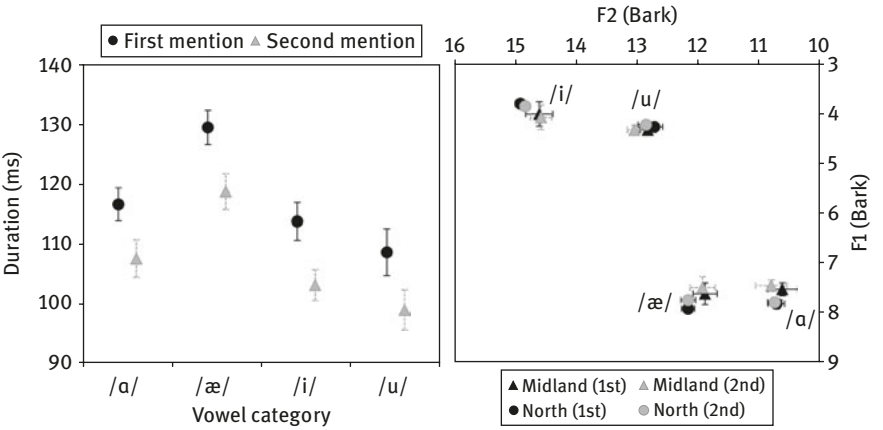

Figure 2.8: Effects of discourse mention on mean vowel duration (left) and mean vowel formant frequencies (right). Error bars show standard error of talker means. Adapted from Clopper et al. (2017).

ɑ, u/, but not /i/), as shown in the left panel of Figure 2.8. Vowels were shorter in easy, second mentions than in harder, first mentions. In the spectral domain, we also observed the expected effect of discourse mention on vowel dispersion for three of the four vowels (/i, æ, u/, but not /ɑ/). Vowels were less dispersed in the vowel space in easy, second mentions relative to harder, first mentions, as shown in the right panel of Figure 2.8. In addition, we observed greater dialect-specific fronting of /u/ for both dialects in second mentions than in first mentions, consistent with the findings for lexical neighborhood density. Unlike the findings for both lexical neighborhood density and semantic predictability, however, no effect of discourse mention was observed for the raising and/or fronting of the Northern /æ/. Thus, for discourse mention, the interaction between phonetic reduction and dialect variation processes is also limited to the spectral domain for a single vowel. This conclusion is qualitatively similar to the conclusions drawn from the analysis of semantic predictability, except that the vowels that exhibit the interaction in the spectral domain differ across linguistic factors (/æ/ for semantic predictability and /u/ for discourse mention).[7]

7 Note, however, that /u/ was not examined in the semantic predictability study, so it is possible that the spectral variation patterns observed for lexical neighborhood density could be replicated with semantic predictability. The studies of lexical neighborhood density and discourse mention examined the same set of vowels, however, so direct comparison between those results is highly interpretable.

Finally, in an investigation of the effect of speaking style on vowel reduction and dialect-specific variants in the Midland and Northern dialects of American English (Clopper, Mitsch, and Tamati 2017), speaking style did not interact with dialect variation in the temporal domain and interacted with dialect variation for two vowels in the spectral domain. In the temporal domain, we observed the expected effect of speaking style for all four vowels, as shown in the left panel of Figure 2.9. Vowels were shorter in plain lab speech than in clear lab speech. In the spectral domain, we observed the expected effect of speaking style on vowel dispersion for three of the four vowels (/i, æ, u/, but not /ɑ/), as shown in the right panel of Figure 2.9. Vowels were less dispersed in the vowel space in plain lab speech than in clear lab speech. In addition, we observed more spectral reduction overall for the Northern talkers than Midland talkers in plain lab speech relative to clear lab speech. As in the previous studies, we also obtained evidence for more fronting of /u/ for both talker dialects and more raising of /æ/ for the Northern talkers in plain lab speech than in clear lab speech. In a separate study investigating the effects of speaking style on /ɑj/ monophthongization in Midland and Southern American English, Turnbull and Clopper (2013) observed the expected effects of talker dialect and speaking style, but the two factors did not interact. Southerners produced more monophthongal /ɑj/ than Midland talkers in both speaking styles and both groups of talkers produced more monophthongal /ɑj/ in plain lab speech than in clear lab speech. Thus, in the spectral domain, the observation that dialect information is marked more strongly in easy contexts relative to hard contexts was robust across vowel

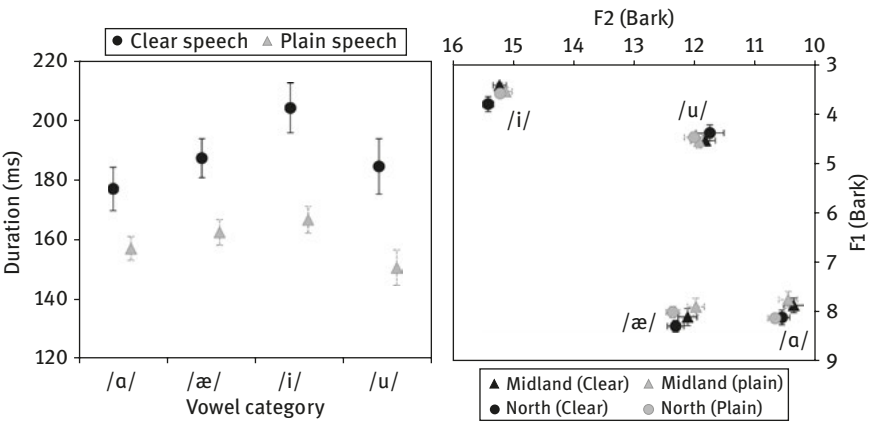

Figure 2.9: Effects of speaking style on mean vowel duration (left) and mean vowel formant frequencies (right). Error bars show standard error of talker means. Adapted from Clopper et al. (2017).

categories, but in the temporal domain, including both vowel duration and vowel trajectory, no interactions between speaking style and talker dialect were observed.

Taken together, the results of these studies provide support for the hypothesis that dialect information is marked more strongly in the same contexts that lead to phonetic reduction. In particular, the results of the previous studies (Hay, Jannedy, and Mendoza-Denton 1999; Munson 2007) and the work in our laboratory described above have demonstrated this relationship between sociolinguistic marking and phonetic reduction for lexical frequency, lexical neighborhood density, semantic predictability, discourse mention, and speaking style. Thus, across linguistic factors, when dialect variation interacts with linguistic context in the temporal and/or spectral realization of vowels, more extreme dialect-specific variants are observed in easier processing contexts relative to harder contexts.

However, the recent research in our laboratory has also shown that this interaction between dialect variation and linguistic context does not emerge robustly across vowel categories or acoustic domains. Although the effects are relatively robust in the spectral domain, they are much weaker in the temporal domain. Whereas interactions between linguistic factors and dialect variation have been observed in the spectral domain for at least some vowel categories in all of the relevant studies, the only interactions between linguistic factors and dialect variation that have been observed in the temporal domain are for dialect differences in duration of /i/ in our work and /ɑj/ monophthongization in Oprah Winfrey's speech in Hay et al.'s (1999) study. This difference between the observed effects in the temporal and spectral domains may reflect the relative importance of temporal and spectral information in conveying dialect information in English. However, the pattern presents an interesting contrast to the results discussed in the previous section in which the temporal domain exhibited more robust phonetic reduction effects than the spectral domain.[8]

The analyses of the interactions between dialect variation and linguistic factors in phonetic reduction processes described in this section were necessarily separated by vowel category because different vowels exhibit different patterns of variation across dialects. These by-vowel analyses revealed variation in phonetic reduction processes across vowels in both acoustic domains, as well as

8 Lexical neighborhood density may present an exception to this general observation. Although we observed significant effects of lexical neighborhood density on vowel duration, but not dispersion, in our study (Burdin, Turnbull, and Clopper 2015), some previous studies on lexical neighborhood effects on phonetic reduction have reported the opposite pattern (e.g., Munson and Solomon 2004).

variation across vowel categories in the interaction between dialect variation and linguistic factors. Specifically, temporal reduction due to semantic predictability and discourse mention was variable across vowels, with only three out of four vowels in each study exhibiting a robust effect. Although the sets of vowels differed in the two studies, some direct comparisons are possible. For example, temporal reduction of /i/ was observed for semantic predictability, but not discourse mention. Similarly, spectral reduction due to semantic predictability, discourse mention, and speaking style was variable across vowels, with only three out of four vowels in each study exhibiting a robust effect. Again, although the sets of vowels differed in the three studies, spectral reduction of /ɑ/ was observed for semantic predictability, but not for discourse mention or speaking style. Wright (2004) also observed variation in spectral reduction across vowel categories in his study of lexical neighborhood density effects and concluded that the point vowels are more likely to exhibit spectral reduction than other vowels because they have more space to centralize. However, our results show a mixed pattern of reduction even for the point vowels, suggesting that additional linguistic and/or nonlinguistic constraints beyond those considered here may be at play in phonetic reduction processes (see also Gahl 2015; Holliday and Turnbull 2015).

Further, although more advanced fronting of /u/ was observed in the easy context for both Midland and Northern talkers in all three studies in which we examined /u/ (i.e., lexical neighborhood density, discourse mention, speaking style), the raising and fronting of /æ/ by Northern talkers was more variable across studies. We observed more advanced raising and/or fronting of /æ/ by the Northern talkers in the low-density, high-predictability, and plain speech contexts, but not in the second mention context. Thus, similar to the overall phonetic reduction effects discussed above, the observed interactions between dialect variation and linguistic factors vary across vowel categories and the linguistic factors have different effects on dialect-specific variants within and across dialects. Although dialect variants have different social meanings and may therefore exhibit different patterns of variation across contexts, the variable interactions across linguistic factors suggest that the linguistic factors themselves may reflect different underlying processes that interact differently with dialect variation. As suggested above, the linguistic factors may represent different locations along an easy/hard processing continuum and dialect variation may interact with that continuum in a nonlinear way. Dialect variation is therefore another dimension that must be considered in further explorations of the hypothesis that all sources of phonetic reduction reflect the same underlying processing demands.

Several of our findings also suggest that there is variation in phonetic reduction processes across regional dialects. For example, we observed more temporal reduction of /i/ due to lexical neighborhood density for the Northern talkers than

for the Midland talkers, as well as more spectral reduction due to speaking style for the Northern talkers than for the Midland talkers. Additional evidence for dialect variation in reduction processes, including segmental alternations such as flapping and vowel reduction to schwa, comes from Byrd's (1994) study of the TIMIT corpus and Clopper and Smiljanic's (2015) study of variation in temporal organization in regional dialects of American English. In particular, American English dialects differ in speaking rate and pausing, but Clopper and Smiljanic (2015) observed additional effects of dialect variation on consonant and vowel timing that cannot be attributed to speaking rate variability. Clopper and Smiljanic (2015) hypothesized that this timing variability may be due to variation in reduction phenomena across dialects and provided some preliminary evidence that consonant cluster reduction and coda /t/ deletion and glottalization differ across dialects. We may therefore also expect phonetic vowel reduction to vary across dialects and other social categories, which may lead to further complex interactions among social and linguistic factors in phonetic reduction processes which are independent of the variability we have observed within and across linguistic factors, vowel categories, and acoustic domains.

2.4.3 Interactions between linguistic and cognitive factors

A third component of our recent research on phonetic reduction has explored the interactions between individual cognitive factors and the linguistic factors contributing to phonetic reduction. Within linguistics, the literature on the effects of individual cognitive differences on speech production is largely limited to developmental and clinical studies. However, a small but growing body of work is critically examining the role of individual differences in explaining variation in linguistic behaviors (see also Doherty et al.'s 2013, analysis of the role (or lack thereof) of variation in psychology research).

One recent study was conducted by Yu (2010), who examined individual differences in perceptual accommodation to coarticulation. Previous research demonstrated that listeners adjust their phoneme category boundaries in coarticulatory contexts (Beddor, Harnsberger, and Lindemann 2002). For example, when [s] is adjacent to [u], it has a lower centroid frequency, making it more [ʃ]-like. Listeners are aware of this coarticulatory pattern and are more likely to classify a sound that is ambiguous between [s] and [ʃ] as /s/ when in the context of [u]; that is, they perceptually accommodate the coarticulation (Mitterer 2006). Yu's (2010) study examined the role of autistic traits in neurotypical adults in this kind of perceptual accommodation to coarticulation. Autistic traits were assessed via the Autism-spectrum Quotient (AQ; Baron-Cohen et al. 2001), a short self-report

questionnaire designed to probe the extent to which someone's cognitive style mirrors that of a person with autism. The AQ was specifically designed to assess the dimensions of social skills, attention switching, communication, imagination, and attention to detail, with the notion that people with autism have deficits in the former four dimensions, and a surplus in the latter dimension. In particular, people with autism tend to exhibit strong attention to physical detail, while missing contextual or global cues (Happé and Frith 2006). The literature suggests that people with autism are less able to recognize global properties of speech, such as emotional content (Kleinman, Marciano, and Ault 2001) and regional dialect (Clopper, Rohrbeck, and Wagner 2012) than neurotypical individuals, and that proportionally more of their attention is devoted to acoustic detail over linguistic detail (Järvinen-Pasley, Pasley, and Heaton 2008). With this background in mind, Yu (2010) obtained the result that neurotypical adults with a greater prevalence of autistic traits in their personality (i.e., higher AQ scores) exhibited larger perceptual accommodation effects, while people with very few autistic traits (i.e., lower AQ scores) only accommodated to the coarticulation to a minor degree. This result is somewhat surprising, because rather than ignoring context and focusing on the acoustic signal alone, the participants with higher AQ scores (i.e., greater autistic traits) were instead paying more attention to the context and adjusting their perceptions accordingly. Nevertheless, this result has been replicated by Yu and Lee (2014) and Turnbull (2015a).

Yu (2010) explained these results in terms of an enhanced capacity to "systemize," that is, to create associations between objects and rules, in the participants with higher AQ scores. This capacity allows these individuals to keep track of contextually conditioned phonetic variation, such as coarticulation, which then allows them to perceptually accommodate the variation to a greater degree than other individuals. Yu's (2010) account also posits that these high-AQ individuals expend less cognitive effort on attention to social context and cues, which explains their relative deficits in attention switching and communication skills, and in turn means that these resources are freed up for attending to patterns of phonetic variation. This explanation is theoretically consistent with the mechanisms of perceptual accommodation to coarticulation outlined by Sonderegger and Yu (2010), although the main empirical claims are as yet untested.

Yu's (2010) finding of a relationship between patterns of perceptual accommodation and individual variation in cognitive style, as well as the findings from similar studies by Stewart and Ota (2008), naturally prompt the question of whether other linguistic phenomena are similarly influenced by such individual differences. To the extent that perception is mirrored in production (Beddor, Harnsberger, and Lindemann 2002; Casserly and Pisoni 2010; cf. Pardo 2012), and to the extent that processes of coarticulation are related to processes of

reduction (Deng, Yu, and Acero 2006; Moon and Lindblom 1994; Mooshammer and Geng 2008; cf. Browman and Goldstein 1992; Scarborough 2013), individual variation in cognitive style in general, and autistic traits in particular, may influence phonetic reduction. To explore this hypothesis, interactions between linguistic factors (lexical frequency, lexical neighborhood density, semantic predictability, and discourse mention) and individual AQ scores in phonetic reduction were examined in a series of studies by Turnbull (2015a, 2015b). The results demonstrate that talkers with higher AQ scores tended to have a larger difference between their word productions in semantically predictable versus unpredictable contexts, relative to talkers with lower AQ scores. This effect is depicted in the left panel of Figure 2.10 and is broadly consistent with Yu's (2010) "systemizing" account: the high-AQ talkers are able to determine the subtle systems and patterns within speech, such as noting the statistical trend for phonetic reduction in highly predictable contexts. This pattern is then reflected in their productions. The low-AQ talkers, on the other hand, do not notice the trend or only learn it inconsistently, leading to nonexistent or small reductions in highly predictable contexts. The modeling also revealed no significant interaction between AQ and discourse mention, as shown in the right panel of Figure 2.10: all participants, regardless of AQ, produced shorter words for second mentions than first mentions to the same degree (approximately a 25 ms reduction). This distinction between the effects of semantic predictability and discourse mention highlights their potentially different cognitive sources.

For lexical frequency and lexical neighborhood density, the statistical models revealed a third pattern. For these factors, participants with higher AQ scores were less affected by lexical frequency and lexical neighborhood density than

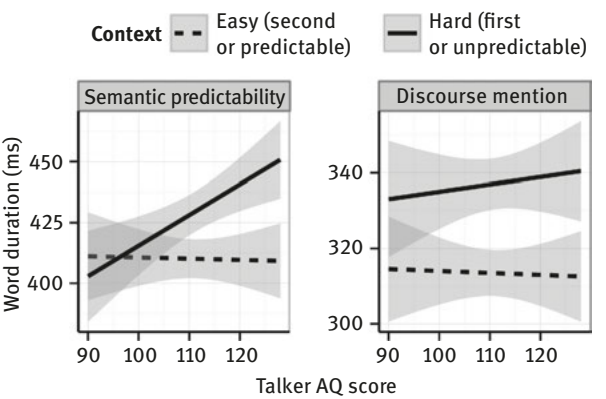

Figure 2.10: Effects of talker AQ score and semantic predictability (left) and discourse mention (right) on word duration. Adapted from Turnbull (2015a).

the lower AQ participants. That is, the acoustic differences – the magnitude of the phonetic reduction – between high- and low-frequency and -density words were smaller for the high-AQ participants than for the low-AQ participants. This result does not immediately appear to be consistent with Yu's (2010) account. However, these results are interpretable in light of the broader research on the autism phenotype. In particular, Stewart and Ota (2008) demonstrated that neurotypical individuals with higher AQ scores exhibit a weaker Ganong effect (Ganong 1980) than individuals with lower AQ scores, suggesting a weaker link between the perceptual system and the lexicon for higher AQ individuals. A weaker link to lexical knowledge could explain the smaller effect sizes for the lexical factors for the higher AQ participants in Turnbull's (2015a) study. Another possible explanation for these results involves an appeal to theory of mind, the ability to impute mental states to others. One of the components of the autism phenotype is proposed to be a weak theory of mind (Baron-Cohen, Leslie, and Frith 1985), and it is therefore possible that higher AQ individuals possess a less well-developed theory of mind than lower AQ individuals. Given a listener-oriented model of phonetic reduction, talkers must have a well-developed theory of mind to model their interlocutor's knowledge, because it is crucial for knowing when to reduce and when to speak clearly. Thus, weaker or more inconsistent phonetic reduction is an expected behavior of individuals with poorer theory of mind and, by extension, a high AQ score. However, this explanation fails to account for the observed interaction with semantic predictability or the lack of an interaction with discourse mention.

Thus, as in our exploration of dialect variation and phonetic reduction in the previous section, we observe considerable variability across linguistic factors in the relationship between cognitive factors and phonetic reduction processes, suggesting that a more nuanced understanding of the relationship between processing demands and phonetic reduction processes is warranted. In particular, the differences we observed across linguistic factors suggest that these factors may reflect different underlying cognitive processes. For example, although the concept of cognitive "accessibility" as a metric of processing difficulty is useful in accounting for both lexical frequency and discourse mention effects, because high-frequency and second mention words are more accessible than low-frequency and first mention words, these phenomena presumably rely on different kinds of accessibility – the former on lexical accessibility and the latter on discourse or referential accessibility. These different types of accessibility may exhibit different effects on processing in different contexts or exhibit different sensitivity to other cognitive or linguistic constraints, which individual differences research could help uncover. Given that the role of individual cognitive differences in speech processing in the neurotypical population is relatively poorly understood, our work in this area represents only a very preliminary step toward

unpacking the potential interactions in this domain, but our initial findings suggest that individual differences may be an important component to understanding phonetic reduction processes.

2.5 Conclusions

We propose that a more complex view of phonetic reduction processes is necessary to account for these observed patterns of variation. As suggested above, this complexification must minimally involve a gradient notion of processing difficulty combined with an allowance for nonlinear relationships between the linguistic factors, the processing difficulty continuum, and phonetic reduction processes. These nonlinear relationships could allow us to capture the apparent limits on phonetic reduction that are observed in some contexts, as well as the variation in the magnitude of phonetic reduction that is observed across acoustic domains and linguistic contexts. This complexification may also involve the differentiation of different kinds of processing demands, including the costs associated with accessing different kinds of linguistic information.

The necessary research to identify the nature of the processing demands that impact phonetic reduction is also likely to help distinguish among the talker-oriented, listener-oriented, and passive evolutionary approaches. Conceptually, all three accounts can be adapted to accommodate the proposed requirements for a gradient notion of processing difficulty that is nonlinearly related to both the linguistic factors and phonetic reduction processes and that differs across acoustic domains. From a listener-oriented perspective, the estimation of potential listener difficulty simply involves more complex computations of processing costs and the appropriate degree of phonetic reduction given the context. From a talker-oriented perspective, processing costs from different levels of representation (e.g., discourse and lexical) must be combined nonlinearly to drive the observed variation in production. From an evolutionary perspective, the exemplar space of potential production targets must be defined based on a large set of weighted contributing factors so that the selected production target reflects the nonlinear combination of the contextual effects that are experienced over time.

The general pattern of interactions between dialect variation and phonetic reduction can also be accommodated in any of the three approaches under the assumption that the dialect-specific variants are the truly native variants for the talker and are therefore easier for the talker to produce. In a listener-oriented account, the talker provides more dialect information by producing the easy, dialect-specific variants when the listener is likely to understand the message. That is, under easy processing conditions, talkers can afford to provide additional

information indexing social information about themselves. However, under more difficult processing conditions, talkers produce more effortful, standard variants in an attempt to make processing easier for the listener.[9] In a talker-oriented account, when processing is relatively easy, dialect-specific variants are activated most quickly because they are the native variants, but when processing is more difficult, more time is available to allow standard variants to be accessed. Similarly, in an evolutionary account, words that are easy to process can be produced and perceived with greater dialect variation and are therefore represented with more variable distributions than words that are harder to process. Thus, for example, high-frequency words will be represented not only by distributions containing more reduced forms but also by distributions containing more dialect-specific forms, leading to the selection of more extreme dialect-specific variants for high-frequency words than for low-frequency words in production.

Nevertheless, all three approaches also face challenges from some of the findings reported in the literature. The listener-oriented account is challenged by findings such as those obtained by Bard et al. (2000), which show that talkers do not always take the needs of their listeners into account. One proposed solution to this apparent problem for the listener-oriented account is to assume a simpler computation of listener need (e.g., Galati and Brennan's 2010, "one-bit" model of audience design), but this kind of simplification is clearly at odds with the evidence we have presented, suggesting the need for a more complex relationship between processing difficulty and phonetic reduction processes. In contrast, the talker-oriented account cannot easily accommodate the speaking style data, which reveal similar phonetic effects arising from explicit instructions about listener needs. That is, the nature of the speaking style manipulation is difficult to reconcile with the talker-oriented account. One obvious solution to this problem would be to treat speaking style as a distinct phenomenon that is separate from phonetic reduction processes, but the acoustic-phonetic realizations of the two phenomena are so similar that this solution seems to violate the goal

9 This account critically relies on the assumption that standard variants are more intelligible than nonstandard variants. Although standard varieties are more intelligible than nonstandard varieties, regardless of the listener's native dialect (e.g., Clopper and Bradlow 2008; Floccia et al. 2006; Sumner and Samuel 2009), nonstandard varieties are also highly intelligible to native speakers of those varieties (e.g., Floccia et al. 2006; Mason 1946; Sumner and Samuel 2009). Thus, the listener-oriented account may lead to different predictions depending on whether the talker and the listener share a dialect. In particular, when a nonstandard dialect is shared by the talker and the listener, dialect-specific information may be enhanced in difficult processing contexts to maximize intelligibility, contrary to the patterns observed in our data that were collected under conditions in which the dialect of the imagined interlocutor was unspecified.

of parsimony in theoretical accounts of speech production. Similarly, the evolutionary account was developed with a focus on lexical frequency effects. The extension of the model to other linguistic factors contributing to phonetic reduction therefore presents the most significant challenge to this approach. Whereas lexical frequency is straightforwardly represented in an exemplar model by the number of experienced tokens, the implementation of a model that can account for other lexical, discourse, and stylistic factors is less straightforward. Finally, Turnbull's (2015a) individual differences data present a challenge to all three approaches because they reveal different patterns of interaction between the cognitive AQ measure and phonetic reduction across linguistic factors, suggesting that different underlying cognitive processes are at play.

In the same way that different phenomena present challenges to the different approaches, some phenomena may be best accounted for by one of the three approaches. For example, the talker-oriented mechanism provides a strong account for discourse mention as in Bard et al.'s (2000) study, whereas evolutionary mechanisms provide a compelling account of lexical frequency as in Pierrehumbert's (2002) model, and a listener-oriented approach is the most obvious account of speaking style as an explicit adjustment in response to task instructions. These intuitions that different approaches provide compelling accounts of different results, together with the evidence for mixed results across linguistic factors and acoustic domains, have led some researchers to abandon a single account of phonetic reduction in favor of a hybrid approach. For example, Watson (2010) proposed a hybrid account in which temporal reduction reflects talker-oriented processing costs, but reduction in f0 reflects listener-oriented processing costs. Similarly, Turnbull (2015a) argued for a hybrid account of his individual cognitive differences data in which lexical effects on phonetic reduction reflect an exemplar lexicon as in the evolutionary perspective, but contextual effects on phonetic reduction reflect a talker-oriented model of the common ground. Although this kind of hybrid approach is less parsimonious than a single account of phonetic reduction, the complexity of the interactions among linguistic, social, and cognitive factors in the realization of phonetic reduction may ultimately require a model of multiple different processes across linguistic factors and/or acoustic domains.

The extent to which phonetic reduction processes are under conscious control is another area of investigation which may help distinguish among these approaches. For example, it is intuitively clear that some speaking style effects are controlled directly by the talker, whereas lexical frequency effects appear to be largely unconsciously controlled. However, care must be taken in the design and interpretation of such investigations, as research in social cognition suggests that a volitional action is not necessarily a consciously controlled action, and vice

versa (see, e.g., Dijksterhuis and Aarts 2010; Moors and De Houwer 2006). A more explicit understanding of the processing demands associated with the relevant linguistic contexts, potentially through careful individual differences research, may provide insight into the locus or loci of the phonetic reduction phenomenon.

Phonetic reduction must also be examined more carefully in interaction with other domains. Our research has revealed interactions with other linguistic factors (see also Gahl 2015, on segmental effects and lexical neighborhood density), with dialect variation (see also Hay, Jannedy, and Mendoza-Denton 1999; Munson 2007), and with individual cognitive factors. These factors all contribute to the phonetic realization of linguistic units and therefore cannot be completely controlled in any analysis of phonetic reduction. Segmental and prosodic structure have a substantial impact on word and vowel duration (de Jong 2004; Klatt 1976), as well as spectral vowel information (de Jong 1995; Fourakis 1991), adding considerable variability to comparisons across words (as in the lexical frequency and lexical neighborhood density analyses) or comparisons of the same words in different contexts (as in analyses involving spontaneous speech). Dialect variation has a substantial impact on spectral vowel information (Labov, Ash, and Boberg 2006), as well as prosody and timing (Clopper and Smiljanic 2011, 2015), adding variability to comparisons across talkers. Our individual differences research (Turnbull 2015a, 2015b) shows that the implementation and magnitude of phonetic reduction also vary across talkers within social groups, adding further variability to our data. Recent advances in automatic phonetic alignment and acoustic analysis, as well as more powerful statistical modeling tools, give us the opportunity to embrace these complex interactions in the search for a more complete understanding of phonetic reduction and its relationship to other speech processing phenomena.

Acknowledgments: This work was supported by the National Science Foundation (BCS-1056409) and a Presidential Fellowship from the Ohio State University Graduate School. We are grateful to Francesco Cangemi, Benjamin Munson, and two anonymous reviewers for comments on a previous draft.

References

Anderson, A. H., M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, et al. 1991. The HCRC map task corpus. *Language and Speech* 34. 351–366.

Arnold, J. E., J. M. Kahn & G. C. Pancani 2012. Audience design affects acoustic reduction via production facilitation. *Psychonomic Bulletin and Review* 19. 505–512.

Arnon, I. & U. Cohen Priva 2013. More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech* 56. 349–371.

Aylett, M. 2000. *Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and care of articulation in spontaneous speech*. PhD thesis, University of Edinburgh.

Aylett, M. & A. E. Turk 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47. 31–56.

Aylett, M. & A. E. Turk 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America* 119. 3048–3058.

Baese-Berk, M. & M. Goldrick 2009. Mechanisms of interaction in speech production. *Language and Cognitive Processes* 24. 527–554.

Baker, R. E. & A. R. Bradlow 2009. Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech* 52. 391–413.

Balota, D. A. & J. I. Chumbley 1985. The locus of word frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language* 24. 89–106.

Bard, E. G., A. H. Anderson, C. Sotillo, M. Aylett, G. Doherty-Sneddon & A. Newlands 2000. Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language* 42. 1–22.

Baron-Cohen, S., A. M. Leslie & U. Frith 1985. Does the autistic child have a 'theory of mind'? *Cognition* 21. 37–46.

Baron-Cohen, S., S. Wheelwright, R. Skinner, J. Martin & E. Clubley 2001. The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders* 31. 5–17.

Beattie, G. W. & B. L. Butterworth 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* 22. 201–211.

Beddor, P. S., J. D. Harnsberger & S. Lindemann 2002. Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics* 30. 591–627.

Bell, A. 1984. Language style as audience design. *Language in Society* 13. 145–204.

Bell, A., J. M. Brenier, M. Gregory, C. Girand & D. Jurafsky 2009. Predictability effects on durations of content and function words in conversational speech. *Journal of Memory and Language* 60. 92–111.

Bezerra, B. M., A. S. Souto, A. N. Radford & G. Jones 2010. Brevity is not always a virtue in primate communication. *Biology Letters* 7. 23–25.

Blevins, J. & A. Wedel 2009. Inhibited sound change: An evolutionary approach to lexical competition. *Diachronica* 26. 143–183.

Bouavichith, D. & L. Davidson 2013. Segmental and prosodic effects on intervocalic voiced stop reduction in connected speech. *Phonetica* 70. 182–206.

Broadbent, D. E. 1967. Word-frequency effect and response bias. *Psychological Review* 74. 1–15.

Brouwer, S., H. Mitterer & F. Huettig 2013. Discourse context and the recognition of reduced and canonical spoken words. *Applied Psycholinguistics* 34. 519–539.

Browman, C. P. & L. Goldstein 1992. Articulatory phonology: An overview. *Phonetica* 49. 155–180.

Bullock-Rest, N., A. Cerny, C. Sweeney, C. Palumbo, K. Kurowski & S. E. Blumstein 2013. Neural systems underlying the influence of sound shape properties of the lexicon on spoken word production: Do fMRI findings predict effects of lesions in aphasia? *Brain and Language* 126. 159–168.

Burdin, R. S. & C. G. Clopper 2015. Phonetic reduction, vowel duration, and prosodic structure. *Proceedings of the 18th International Congress of Phonetic Sciences. 378.*

Burdin, R. S., S. Phillips-Bourass, R. Turnbull, M. Yasavul, C. G. Clopper & J. Tonhauser 2015. Variation in the prosody of focus in head- and head/edge-prominence languages. *Lingua* 165. 254–276.

Burdin, R. S., R. Turnbull & C. G. Clopper 2015. Interactions among lexical and discourse characteristics in vowel production. *Proceedings of Meetings on Acoustics* 22. 060005.

Byrd, D. 1994. Relations of sex and dialect to reduction. *Speech Communication* 15. 39–54.

Calhoun, S. 2010a. The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language* 86. 1–42.

Calhoun, S. 2010b. How does informativeness affect prosodic prominence? *Language and Cognitive Processes* 25. 1099–1140.

Casserly, E. D. & D. B. Pisoni 2010. Speech perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science* 1. 629–647.

Chafe, W. 1974. Language and consciousness. *Language* 50. 111–133.

Clopper, C. G. & A. R. Bradlow 2008. Perception of dialect variation in noise: Intelligibility and classification. *Language and Speech* 51. 175–198.

Clopper, C. G., J. F. Mitsch & T. N. Tamati 2017. Effects of phonetic reduction and dialect variation on vowel production. *Journal of Phonetics* 60. 38-59.

Clopper, C. G., K. L. Rohrbeck & L. Wagner 2012. Perception of dialect variation by young adults with high-functioning autism. *Journal of Autism and Developmental Disorders* 42. 740–754.

Clopper, C. G. & J. B. Pierrehumbert 2008. Effects of semantic predictability and regional dialect on vowel space reduction. *Journal of the Acoustical Society of America* 124. 1682–1688.

Clopper, C. G. & R. Smiljanic 2011. Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics* 39. 237–245.

Clopper, C. G. & R. Smiljanic 2015. Regional variation in temporal organization in American English. *Journal of Phonetics* 49. 1–15.

Clopper, C. G. & T. N. Tamati 2014. Effects of local lexical competition and regional dialect on vowel production. *Journal of the Acoustical Society of America* 136. 1–4.

Clopper, C. G., R. Turnbull, and R. S. Burdin in press. Assessing predictability effects in connected read speech. *Linguistics Vanguard*.

Cohen Priva, U. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology* 6. 243–278.

de Jong, K. J. 1995. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America* 97. 491–504.

de Jong, K. 2004. Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *Journal of Phonetics* 32. 493–516.

Deng, L., D. Yu & A. Acero 2006. A bidirectional target filtering model of speech coarticulation: Two-stage implementation for phonetic recognition. *IEEE Transactions on Audio and Speech Processing* 14. 256–265.

Dijksterhuis, A. & H. Aarts 2010. Goals, attention, and (un)consciousness. *Annual Review of Psychology* 61. 467–490.

Doherty, M. E., K. M. Shemberg, R. B. Anderson & R. D. Tweney 2013. Exploring unexplained variation. *Theory & Psychology* 23. 81–97.

Eckert, P. & J. R. Rickford 2001. *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Engelhardt, P. E. & F. Ferreira 2014. Do speakers articulate over-described modifiers differently from modifiers that are required by context? Implications for models of reference production. *Language, Cognition and Neuroscience* 29. 975–985.

Ernestus, M. 2014. Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua* 142. 27–41.

Ferguson, S. H. & D. Kewley-Port 2007. Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research* 50. 1241–1255.

Ferrer-i-Cancho, R. & B. Elvevåg 2010. Random texts do not exhibit the real Zipf's law-like rank distribution. *Plos One* 5. e9411.

Ferrer-i-Cancho, R., A. Hernández-Fernández, D. Lusseau, G. Agoramoorthy, M. J. Hsu & S. Semple 2013. Compression as a universal principle of animal behavior. *Cognitive Science* 37. 1565–1578.

Ferrer-i-Cancho, R. & F. Moscoso del Prado 2011. Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment*. L12002.

Flemming, E. 2010. Modeling listeners: Comments on Pluymaekers et al. and Scarborough. In C. Fougeron, B. Kühnert, M. D'Imperio & N. Vallée (eds.), *Laboratory Phonology 10*, 587–605. Berlin: De Gruyter Mouton.

Floccia, C., F. Girard, J. Goslin & G. Konopczynski 2006. Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance* 32. 1276–1293.

Fourakis, M. 1991. Tempo, stress, and vowel reduction in American English. *Journal of the Acoustical Society of America* 90. 1816–1827.

Fowler, C. A. 1988. Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech* 31. 307–319.

Fowler, C. A. & J. Housum 1987. Talkers' signalling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language* 26. 489–504.

Fox, N. P., M. Reilly & S. E. Blumstein 2015. Phonological neighborhood competition affects spoken word production irrespective of sentential context. *Journal of Memory and Language* 83. 97–117.

Gahl, S. 2015. Lexical competition in vowel articulation revisited: Vowel dispersion in the Easy/Hard database. *Journal of Phonetics* 49. 96–116.

Gahl, S. & S. M. Garnsey 2004. Knowledge of grammar, knowledge of usage: Syntactic probability affects pronunciation variation. *Language* 80. 748–775.

Gahl, S., Y. Yao & K. Johnson 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66. 789–806.

Galati, A. & S. E. Brennan 2010. Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language* 62. 35–51.

Ganong, W. F. 1980. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* 6. 110–125.

Garrett, A. & K. Johnson 2012. Phonetic bias in sound change. In A. C. L. Yu. (ed.), *Origins of Sound Change: Approaches to Phonologization,* 51–97. Oxford: Oxford University Press.

Goldinger, S. D. 1998. Echoes of echoes?: An episodic theory of lexical access. *Psychological Review* 105. 251–279.
Goldrick, M., C. Vaughn & A. Murphy 2013. The effects of lexical neighbors on stop consonant articulation. *Journal of the Acoustical Society of America* 134. EL172–EL177.
Happé, F. G. E. & U. Frith 2006. The weak coherence account: Detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders* 36. 5–25.
Hawkins, J. 2014. *Cross-Linguistic Variation and Efficiency*. Oxford: Oxford University Press.
Hay, J., S. Jannedy & N. Mendoza-Denton 1999. Oprah and /ay/: Lexical frequency, referee design, and style. *Proceedings of the 14th International Congress of Phonetic Sciences,* 1389–1392.
Hoetjes, M., R. Koolen, M. Goudbeek, E. Krahmer & M. Swerts 2015. Reduction in gesture during the production of repeated references. *Journal of Memory and Language* 79/80. 1–17.
Hoetjes, M., E. Krahmer & M. Swerts 2012. Do repeated references result in sign reduction? *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, 461–466.
Holliday, J. J. & R. Turnbull 2015. Effects of phonological neighborhood density on word production in Korean. *Proceedings of the 18th International Congress of Phonetic Sciences*. 891.
Howes, D. 1957. On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America* 29. 296–305.
Hunnicutt, S. 1987. Acoustic correlates of redundancy and intelligibility. *Quarterly Progress Status Report, Speech Transmission Lab* 28 (2–3). 7–14.
Ito, K. & S. R. Speer 2006. Using interactive tasks to elicit natural dialogue. In Sudhoff, S. et al. (eds.), *Methods in Empirical Prosody Research,* 229–257. Berlin: Mouton de Gruyter.
Jaeger, T. F. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61. 23–62.
Jaeger, T. F. 2013. Production preferences cannot be understood without reference to communication. *Frontiers in Psychology* 4. 230.
Järvinen-Pasley, A., J. Pasley & P. Heaton 2008. Is the linguistic content of speech less salient than its perceptual features in autism? *Journal of Autism and Developmental Disorders* 38. 239–248.
Johnson, K. 1997. Speech processing without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (eds.), *Talker Variability in Speech Processing,* 145–165. San Diego: Academic Press.
Johnson, K. 2004. Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (eds.), *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium,* 29–54. Tokyo: The National Institute for Japanese Language.
Johnson, K., E. Flemming & R. Wright 1993. The hyperspace effect: Phonetic targets are hyperarticulated. *Language* 69. 505–528.
Jurafsky, D., A. Bell, M. Gregory & W. D. Raymond 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper. (eds.), *Frequency and the Emergence of Linguistic Structure,* 229–254. Amsterdam: John Benjamins.
Kahn, J. M. & J. E. Arnold 2012. A processing-centered look at the contribution of givenness to durational reduction. *Journal of Memory and Language* 67. 311–325.
Kahn, J. M. & J. E. Arnold 2015. Articulatory and lexical repetition effects on durational reduction: Speaker experience vs. common ground. *Language, Cognition and Neuroscience* 30. 103–119.

Kaiser, E., D. C.-H. Li & E. Holsinger 2011. Exploring the lexical and acoustic consequences of referential predictability. In I. Hendrickx, A. Branco, S. L. Devi & R. Mitkov (eds.), *Anaphora Processing and Applications,* 171–183. Heidelberg: Springer.

Kaland, C., M. Swerts & E. Krahmer 2013. Accounting for the listener: Comparing the production of contrastive intonation in typically-developing speakers and speakers with autism. *Journal of the Acoustical Society of America* 134. 2182–2196.

Kalikow, D. N., K. N. Stevens & L. L. Elliott 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America* 61. 1337–1351.

Keysar, B. 2008. Egocentric processes in communication and miscommunication. In I. Kecskes & J. Mey. (eds.), *Intention, Common Ground and the Egocentric Speaker-Hearer,* 277–296. New York: Mouton de Gruyter.

Keysar, B. & D. J. Barr 2005. Coordination of action and belief in communication. In J. C. Trueswell & M. K. Tanenhaus. (eds.), *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions,* 71–94. Cambridge, MA: MIT Press.

Keysar, B., D. J. Barr, J. A. Balin & J. S. Brauner 2000. Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science* 11. 32–38.

Kirov, C. & C. Wilson 2012. The specificity of online variation in speech production. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, 587–592.

Klatt, D. H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America* 59. 1208–1221.

Kleinman, J., P. L. Marciano & R. L. Ault 2001. Advanced theory of mind in high-functioning adults with autism. *Journal of Autism and Developmental Disorders* 31. 29–36.

Labov, W., S. Ash & C. Boberg 2006. *Atlas of North American English*. New York: Mouton de Gruyter.

Ladd, D. R. 2011. Phonetics in phonology. In J. Goldsmith, J. Riggle & A. C. L. Yu (eds.), *The Handbook of Phonological Theory,* 348–373. Oxford: Wiley.

Lam, T. Q. & D. G. Watson 2010. Repetition is easy: Why repeated referents have reduced prominence. *Memory and Cognition* 38. 1137–1146.

Lam, T. Q. & D. G. Watson 2014. Repetition reduction: Lexical repetition in the absence of referent repetition. *Journal of Experimental Psychology: Learning, Memory and Cognition* 40. 829–843.

Lehiste, I. 1971. The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America* 51. 2018–2024.

Levy, R. & T. F. Jaeger 2007. Speakers optimize information density through syntactic reduction. *Proceedings of the 20th Annual Conference on Neural Information Processing Systems,* 849–856.

Lieberman, P. 1963. Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech* 6. 172–187.

Lin, S., P. S. Beddor & A. W. Coetzee 2014. Gestural reduction, lexical frequency, and sound change: A study of post-vocalic /l/. *Laboratory Phonology* 5. 9–36.

Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling,* 403–439. Dordrecht: Kluwer.

Luce, P. A. & J. Charles-Luce 1985. Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *Journal of the Acoustical Society of America* 78. 1949–1957.

Luce, P. A. & D. B. Pisoni 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19. 1–36.
Luo, B., T. Jiang, Y. Liu, J. Wang, A. Lin, X. Wei & J. Feng 2013. Brevity is prevalent in bat short-range communication. *Journal of Comparative Physiology A* 199. 325–333.
Mason, H. M. 1946. Understandability of speech in noise as affected by region of origin of speaker and listener. *Speech Monographs* 13 (2). 54–68.
Miller, G. A. & S. Isard 1963. Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior* 2. 217–228.
Mitterer, H. 2006. On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics* 68. 1227–1240.
Mitterer, H. & K. Russell 2013. How phonological reductions sometimes help the listener. *Journal of Experimental Psychology: Learning, Memory and Cognition* 39. 977–984.
Moon, S.-J. & B. Lindblom 1994. Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America* 96. 40–55.
Moore-Cantwell, C. 2013. Syntactic predictability influences duration. *Proceedings of Meetings on Acoustics* 19. 060206.
Moors, A. & J. De Houwer 2006. Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin* 132. 297–326.
Mooshammer, C. & C. Geng 2008. Acoustic and articulatory manifestations of vowel reduction in German. *Journal of the International Phonetic Association* 38. 117–136.
Munson, B. 2007. Lexical characteristics mediate the influence of sex and sex typicality on vowel-space size. *Proceedings of the 16th International Congress of Phonetic Sciences,* 885–888.
Munson, B. 2013. The influence of production latencies and phonological neighborhood density on vowel dispersion. *Proceedings of Meetings on Acoustics* 19. 060192.
Munson, B., J. Edwards, S. K. Schellinger, M. E. Beckman & M. K. Meyer 2010. Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of *Vox Humana*. *Clinical Linguistics and Phonetics* 24. 245–260.
Munson, B. & N. P. Solomon 2004. The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research* 47. 1048–1058.
Myers, J. & Y. Li 2009. Lexical frequency effects in Taiwan Southern Min syllable contraction. *Journal of Phonetics* 37. 212–230.
Nusbaum, H. C., D. B. Pisoni & C. K. Davis 1984. Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report No. 10,* 357–376. Bloomington, IN: Speech Research Laboratory, Indiana University.
Pardo, J. S. 2012. Reflections on phonetic convergence: Speech perception does not mirror speech production. *Language and Linguistics Compass* 6. 753–767.
Pate, J. K. & S. Goldwater 2011. Predictability effects in adult-directed and infant-directed speech: Does the listener matter? *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 1569–1574.
Peramunage, D., S. E. Blumstein, E. B. Myers, M. Goldrick & M. Baese-Berk 2011. Phonological neighborhood effects in spoken word production: An fMRI study. *Journal of Cognitive Neuroscience* 23. 593–603.
Peterson, G. E. & I. Lehiste 1960. Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* 32. 693–703.
Picheny, M. A., N. I. Durlach & L. D. Braida 1985. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research* 28. 96–103.

Picheny, M. A., N. I. Durlach & L. D. Braida 1986. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research* 29. 434–445.
Pierrehumbert, J. 2001a. Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure,* 137–158. Amsterdam: John Benjamins.
Pierrehumbert, J. 2001b. Stochastic phonology. *Glot International* 5. 195–207.
Pierrehumbert, J. 2002. Word-specific phonetics. In C. Gussenhoven & N. Warner (eds.), *Laboratory Phonology VII,* 101–139. Berlin: Mouton de Gruyter.
Pierrehumbert, J. 2003a. Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech* 46. 115–154.
Pierrehumbert, J. 2003b. Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay & S. Jannedy (eds.), *Probabilistic Linguistics,* 177–228. Cambridge, MA: MIT Press.
Pitt, M. A., L. C. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume & E. Fosler-Lussier 2007. *Buckeye Corpus of Conversational Speech*. Columbus, OH: Department of Psychology, Ohio State University.
Pluymaekers, M., M. Ernestus & R. H. Baayen 2005a. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62. 146–159.
Pluymaekers, M., M. Ernestus & R. H. Baayen 2005b. Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America* 118. 2561–2569.
Qian, T. & T. F. Jaeger 2012. Cue effectiveness in communicatively efficient discourse production. *Cognitive Science* 36. 1312–1336.
Ramscar, M. & R. H. Baayen 2013. Production, comprehension, and synthesis: A communicative perspective on language. *Frontiers in Psychology* 4. 233.
Raymond, W. D., R. Dautricourt & E. Hume 2006. Word-internal /t, d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change* 18. 55–97.
Rooth, M. 1992. A theory of focus interpretation. *Natural Language Semantics* 1. 75–116.
Sasisekaran, J. & B. Munson 2012. Effects of repeated production on vowel distinctiveness within nonwords. *Journal of the Acoustical Society of America* 131. EL336–EL341.
Scarborough, R. 2010. Lexical and contextual predictability: Confluent effects on the production of vowels. In C. Fougeron, B. Kühnert, M. D'Imperio & N. Vallée (eds.), *Laboratory Phonology 10,* 557–586. Berlin: De Gruyter Mouton.
Scarborough, R. 2013. Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. *Journal of Phonetics* 41. 491–508.
Scarborough, R., J. Brenier, Y. Zhao, L. Hall-Lew & O. Dmitrieva 2007. An acoustic study of real and imagined foreigner-directed speech. *Proceedings of the 16th International Congress of Phonetic Sciences,* 2165–2168.
Scarborough, R. & G. Zellou 2013. Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *Journal of the Acoustical Society of America* 134. 3793–3807.
Schober, M. F. 1993. Spatial perspective-taking in conversation. *Cognition* 47. 1–24.
Schuppler, B., M. Ernestus, O. Scharenborg & L. Boves 2011. Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics* 39. 96–109.
Semple, S., M. J. Hsu & G. Agoramoorthy 2010. Efficiency of coding in macaque vocal communication. *Biology Letters* 6. 469–471.

Semple, S., M. J. Hsu, G. Agoramoorthy & R. Ferrer-i-Cancho 2013. The law of brevity in macaque vocal communication is not an artefact of analysing mean call durations. *Journal of Quantitative Linguistics* 20. 209–217.

Shields, L. W. & D. A. Balota 1991. Repetition and associative context effects in speech production. *Language and Speech* 34. 47–55.

Silverman, D. 2012. *Neutralization*. Cambridge: Cambridge University Press.

Smiljanic, R. & A. R. Bradlow 2005. Production and perception of clear speech in Croatian and English. *Journal of the Acoustical Society of America* 118. 1677–1688.

Sonderegger, M. & A. C. L. Yu 2010. A rational account of perceptual compensation for coarticulation. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, 375–380.

Stewart, M. E. & M. Ota 2008. Lexical effects on speech perception in individuals with "autistic" traits. *Cognition* 109. 157–162.

Sumner, M. & A. G. Samuel 2009. The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language* 60. 487–501.

Tenpenny, P. L. 1995. Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review* 2. 339–363.

Tily, H. & V. Kuperman 2012. Rational phonological lengthening in spoken Dutch. *Journal of the Acoustical Society of America* 132. 3935–3940.

Tupper, P. F. 2014. Exemplar dynamics models of the stability of phonological categories. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 1628–1633.

Turk, A. 2010. Does prosodic constituency signal relative predictability? A smooth signal redundancy hypothesis. *Laboratory Phonology* 1. 227–262.

Turnbull, R. 2015a. *Assessing the listener-oriented account of predictability-based phonetic reduction*. PhD dissertation, Ohio State University.

Turnbull, R. 2015b. Patterns of individual differences in reduction: Implications for listener-oriented theories. *Proceedings of the 18th International Congress of Phonetic Sciences. 106*.

Turnbull, R. 2017. The role of predictability in intonational variability. *Language and Speech*. 60. 123-153

Turnbull, R. in press. Effects of lexical predictability on patterns of phoneme deletion/reduction in conversational speech in English and Japanese. *Linguistics Vanguard*.

Turnbull, R., R. S. Burdin, C. G. Clopper & J. Tonhauser 2015. Contextual predictability and the prosodic realisation of focus: A cross-linguistic comparison. *Language, Cognition and Neuroscience* 30. 1061–1076.

Turnbull, R. & C. G. Clopper 2013. Effects of semantic predictability and dialect variation on vowel production in clear and plain lab speech. *Proceedings of Meetings on Acoustics* 19. 060028.

van Bergem, D. R. 1993. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication* 12. 1–21.

van Son, R. & L. C. W. Pols 2003. Information structure and efficiency in speech production. *Proceedings of Eurospeech 2003*, 769–772.

van Son, R. & J. P. H. van Santen 2005. Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication* 47. 100–123.

Vitevitch, M. S. 2002. The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28. 735–747.

Vitevitch, M. S. & P. A. Luce 1998. When words compete: Levels of processing in perception of spoken words. *Psychological Science* 9. 325–329.
Vitevitch, M. S. & P. A. Luce 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40. 374–408.
Wagner, M. & J. Klassen 2015. Accessibility is no alternative to alternatives. *Language, Cognition and Neuroscience* 30. 212–233.
Wagner, P., J. Trouvain & F. Zimmerer 2015. In defense of stylistic diversity in speech research. *Journal of Phonetics* 48. 1–12.
Warner, N. & B. V. Tucker 2011. Phonetic variability of stops and flaps in spontaneous and careful speech. *Journal of the Acoustical Society of America* 130. 1606–1617.
Watson, D. G. 2010. The many roads to prominence: Understanding emphasis in conversation. *Psychology of Learning and Motivation* 52. 163-183.
Watson, D. G., J. E. Arnold & M. K. Tanenhaus 2008. Tic tac TOE: Effects of predictability and importance on acoustic prominence in language production. *Cognition* 106. 1548–1557.
Wedel, A. 2006. Exemplar models, evolution and language change. *The Linguistic Review* 23. 247–274.
Wright, R. 2004. Factors of lexical competition in vowel articulation. In J. Local, R. Ogden & R. Temple (eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI,* 75–86. Cambridge: Cambridge University Press.
Wightman, C. W., S. Shattuck-Hufnagel, M. Ostendorf & P. J. Price 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91. 1707–1717.
Yu, A. C. L. 2010. Perceptual compensation is correlated with individuals' "autistic" traits: Implications for models of sound change. *Plos One* 5 (8). 1–9.
Yu, A. C. L. & H. Lee 2014. The stability of perceptual compensation for coarticulation within and across individuals: A cross-validation study. *Journal of the Acoustical Society of America* 136. 382–388.

Wim A. van Dommelen

3 Reduction in native and non-native read and spontaneous speech

Abstract: This chapter views phonetic reduction phenomena from both a first and a second language perspective. The specific aim of this contribution is to explore phonetic reduction of stop consonants in read and spontaneous speech produced by native and non-native (Norwegian) speakers of English. The following three research questions are investigated. First, to which degree native versus non-native English shows differences in segmental reduction. Second, whether reduction patterns vary with speaking style and, third, whether speaking style effects differ between native and non-native speech production.

Read speech material used in the study consisted of BBC news transcripts read by 10 native speakers each of British English and Norwegian. Spontaneous speech material produced by those speakers comprised dialogues elicited by means of a picture replication task performed by pairs of speakers sharing the native language.

Evaluation of this speech material involved auditory analysis as well as acoustic measurements enabling specification of relevant segmental and subsegmental acoustic details. In spite of revealing a complex picture of reduction behaviour, some consistent trends emerged from the two types of analysis. In both native and non-native speakers, segmental reduction was frequent. At the same time, in native productions tendencies of stronger reduction were observed. The effect of speaking style appeared to vary across parameters investigated, but was generally not very strong. Interactions between language background and speaking style were largely non-significant, thus indicating similarly augmented reduction in native and non-native spontaneous speech.

Keywords: stops, English, Norwegian, native, non-native, spontaneous, read speech

3.1 Introduction

This chapter views phonetic reduction phenomena from both a first and a second language speech perspective. Acquiring one's first language (L1) implies being exposed from birth to a multitude of phonetic realisations of the

Wim A. van Dommelen, Department of Language and Literature, Norwegian University of Science and Technology

https://doi.org/10.1515/9783110524178-003

speech sounds from the language's sound inventory. One could assume that most of the words typically spoken in infant-directed speech can be regarded as what might be called canonical forms, featuring non-reduced sound realisations. Still, recent research has shown that highly reduced pronunciation variants may appear frequently in infant-directed speech (Lahey and Ernestus 2014). In adult speech, phonetic reductions are very common. In his analysis of a large American English database with conversational speech, Johnson (2004) counted segmental deviation rates of 25% for content words and 40% for function words. A native speaker of a language will thus repeatedly encounter word forms that are to some degree reduced, genuinely canonical realisations possibly representing the exception rather than the rule. Conditions for non-native learners of a second language (L2) will usually be different. Learning an L2 in a formal setting will often entail being confronted with canonical word forms. The learner encounters such forms demonstrated in class and written in textbooks' introductory chapters presenting vowel and consonant inventories of the pertinent language accompanied by single words as illustration. Certainly, in the early stage of learning, the main focus is on optimal realisation of speech sounds in single words. As a rule, connected speech phenomena have been given only low priority in L2 pronunciation teaching (e.g., Davidsen-Nielsen 1975; Kuiper and Scott Allan 2010). For speakers being immersed in an L2 environment, learning conditions are more favourable. Nevertheless, the number of word form exemplars will be substantially smaller than for native speakers, although of course dependent on age of learning. It has been shown that younger age of learning will make L2 speech more native-like (cf. Flege 1995; Flege, Munro, and Mackay 1995).

The goal of the present study is to investigate to what degree different exposure to canonical forms will affect non-native speakers' reduction behaviour. To that end, we will compare native and non-native behaviour in two different speaking styles, read and spontaneous speech. Previous studies have revealed abundant phonetic variability in L1 speech both within and between speaking styles. Comparing more formal (often read) speech with some form of more casual speech, some consistent tendencies have emerged. Typical characteristics of casual speech are centralised vowel qualities (Koopmans-van Beinum 1980; Laan 1997), reduced values of spectral measures in consonants (Nakamura, Iwano, and Furui 2008; van Son and Pols 1999), and weakening and elision of consonants (Barry and Andreeva 2001). Reduced energy of consonant relative to surrounding vowels was observed in van Son and Pols (1999) and Warner and Tucker (2011). On the other hand, findings on temporal organisation diverge. Shorter segment durations in spontaneous speech were reported for Russian by Bondarko et al. (2003) and Bolotova (2003), for Dutch by van

Son and Pols (1999), and for Finnish by de Silva et al. (2003). In contrast, in the latter study, longer segment durations were observed in spontaneous versus read Russian and Dutch. Also, reported tendencies for f0 differ. De Silva et al. (2003) measured lower mean f0 values in spontaneous Dutch but got the opposite result for Finnish. So, reduction processes due to different speaking styles are complex and may be difficult to predict.

Previous studies on L2 speech have presented relatively scarce empirical evidence on reduction phenomena, focussing mainly on suprasegmentals. A well-established characteristic of L2 speech is a generally slower speech rate than in native speech production. This has been mostly demonstrated for L2 speakers of English (e.g., Bradlow et al. 2011; Guion et al. 2000; Mackay and Flege 2004; Trofimovich and Baker 2006) but also for L2 users of Finnish (Toivola et al. 2010), German (Gut 2009), and Norwegian (van Dommelen 2007). Less is known about the effect of speaking style on speech rate in L2 production. Comparing read and spontaneous speech from different groups of L2 users of Dutch, Cucchiarini, Strik, and Boves (2002) observed similar articulation rates for both speaking styles. Faster articulation rates in read non-native German as well as English have been found by Gut (2009) but, conversely, generally slower speech rates for read versus conversational speech by Toivola et al. (2010). Also reduction phenomena in L2 speech have until now only marginally been topic of acoustic-phonetic investigations (cf. the overview of phonologically oriented studies in Gut 2009). Less frequent syllable reduction in Mandarin-accented English versus native English has been reported by Bradlow et al. (2011). Extensive measures of rhythmic properties at the syllable level in Gut's (2009) study demonstrated consistent native–non-native differences for German and English. Spilková (2014) investigated both segmental and rhythmical aspects of reduction in native and non-native (Czech and Norwegian) speakers of English. The study showed that realisations of three function words were influenced by speakers' L1 background, although the patterns of effects differed between the function words. Her analysis of rhythmical aspects of repeated mentions led her to the conclusion that native and non-native speakers display similar tendencies to phonetic reduction.

During the two last decades or so, models have been developed to predict and explain deviant phoneme realisations by L2 users. To that end, speech sounds in learners' native language phonological systems are compared with those occurring in L2. Examples of such an approach are Best's Perceptual Assimilation Model (PAM: Best 1995; PAM-L2: Best and Tyler 2007), Flege's (1995) Speech Learning Model, and Polka and Bohn's (2011) Natural Referent Vowel framework. Common to the above-mentioned models is that they take canonical representations of speech sounds as their point of departure. To the best of our knowledge,

until now no attempts have been made to include reduction processes in modelling L2 pronunciation.

In view of our present standard of knowledge of reduction phenomena particularly in L2 speech, this investigation seeks to gain more insight through an analysis of read and spontaneous English speech material produced by native and non-native (Norwegian) speakers. Using materials collected by Spilková (2014), we will focus on the phonetic properties of stop consonants, largely at the subphonemic level. The phoneme inventories of English and Norwegian share two series of stop consonants, /p, t, k/ and /b, d, g/ (for Norwegian, see Kristoffersen 2000). Norwegian has two further stop pairs, palatal /c, ɟ/ and retroflex /ʈ, ɖ/. Glottal stops (/ʔ/) occur as phonetic variants of phonologically voiceless stops in English (e.g., Ladefoged and Maddieson 1996). According to informal observations, also in Norwegian, glottal stops are not uncommon. Different from English, they are not used for glottal reinforcement or substitution of oral stops but mostly as indicators of syntactic boundaries. For the present purposes, we will specify the stop series /p, t, k/ versus /b, d, g/ as phonologically voiceless versus voiced, although views on specification diverge for different languages (cf. Ringen and van Dommelen 2013). From evidence presented for English and Norwegian, it can be concluded that the implementation of the voicing contrast differs, at least in intervocalic stop consonants. Edwards (1981) reported an average of 81% closure voicing in English lenis stops in intervocalic position, and Docherty (1992) a degree of 58%. A much higher proportion (93%) was observed in intervocalic Norwegian lenis stops by Ringen and van Dommelen (2013). Norwegian fortis stops in intervocalic position featured throughout (voiceless) preaspiration. In fortis stops in utterance-initial position, voicing lag of on average 52 ms was observed. Halvorsen (1998) reported a mean value of 65 ms for comparable Norwegian stops. Similar values of around 68 ms for voiceless English stops in intervocalic position were found by Edwards (1981). The data available from previous studies thus seem to suggest that the main difference between phonetic implementations of /p, t, k/ and /b, d, g/ in English and Norwegian is concerned with the degree of phonetic voicing of stop closure.

For the present investigation, we will adopt canonical stop realisation as a virtual point of reference. For both phoneme series /p, t, k/ and /b, d, g/, this would imply the building of a complete consonantal closure followed by an abrupt opening. A canonical closure would contain only some passive voicing for the former series and voicing of substantial duration and amplitude for the latter. Generally, we would expect the release burst of a phonologically voiceless stop to be stronger than that of a voiced one. Reduction is a gradual process on a scale ranging from canonical realisation to complete deletion of a stop. Some degree

of reduction might imply failure to build a complete closure, shortened closure duration, and/or a weaker release burst. Strong reduction could manifest itself as realisation as a flap or approximant.

As to actual reduction behaviour in native speakers, based on results from the studies cited above one would predict generally weakened spectral and intensity measures in conversational speech. It seems that for segment durations no specific predictions can be made. Compared with L1 speakers' conditions, non-native reduction behaviour can be assumed to be formed by relatively high exposure to canonical forms and much lower exposure to reduced forms. Therefore, it is hypothesised that the present Norwegian subjects will exhibit a lesser degree of stop reduction in general and also a weaker effect of speaking style than the native speakers. In addition, differences between language-specific ways of realising the stop series /p, t, k/ and /b, d, g/ can be expected to influence stop production.

3.2 Method

3.2.1 Analysis

In contrast to the acoustic analysis of vowel reduction, at present no widely used methods for similar analysis of stops exist. The present study consists of two parts, a qualitative analysis based on auditory and visual inspection of speech signals and a quantitative analysis of segment durations and intensity contours. The qualitative methodology is inspired by the investigation of /t/ reduction in Dutch by Schuppler et al. (2012). In that study, human observers specified stop closures among other things as containing voicing, nasal murmur, or some form of friction, and stop releases as containing a single or multiple burst that could be strong or weak. The present analysis went in similar detail, but only the most conspicuous properties were selected for presentation of results. The quantitative method adopted here is similar to the analysis of stops and flaps in Warner and Tucker (2011) who measured consonant duration, intensity, voicing, and formant structure during consonants appearing between vowels or sonorants. While our selection of tokens mostly comprised stops appropriate for duration and energy measurements, only about 3% were flaps. Therefore, no formant measurements were made. For each of the two analysis methods, a set of variables was established. To enhance the following descriptions' transparency, methodological details are given separately in each of the corresponding sections.

3.2.2 Speech corpus material and speakers

The speech material used for this investigation was selected from two types of recordings made by Spilková (2014) comprising read speech and task-elicited spontaneous speech. To obtain the former type of speech, subjects read one page of a BBC news transcript. Spontaneous speech dialogues were elicited by recording pairs of speakers performing a picture replication task. In this task, one speaker (A) had to describe a cartoony picture to the other speaker (B) who had to draw a replica of the picture on a sheet of paper without any visual information.

Groups of subjects in the Spilková (2014) corpus were native speakers of English and Norwegian speakers of English as a foreign language. Recordings of the natives took place in a sound-attenuated booth at the Department of Experimental Psychology, University of Bristol, using a Shure WH20 headset. Non-natives were recorded in a sound-treated studio at the Department of Language and Communication Studies, NTNU, Trondheim, using an MILAB LSR-1000 microphone. In both cases, recordings were stored with a sampling frequency of 44.1 kHz and 16-bit quantisation. From the native picture task material, we selected five dialogue recordings excluding recordings with occasional technical artefacts (mainly wind noise; Table 3.1). Durations of the recordings varied between 27 and 64 minutes. As the non-native dataset, all the corpus' five collected dialogues between Norwegian speakers were chosen, varying in duration between 48 and 73 minutes. One speaker (AM) was excluded from analysis because of his Norwegian-English bilingual background. For the reading task, a total of 11 different BBC news transcripts were used. As a consequence, the lexical material available for phonetic analysis differed within and between speaker groups. Mean duration of the news transcript recordings was 3.5 minutes for the natives and 3.9 minutes for the non-natives.

Both native and non-native subjects were recruited from student populations. As can be seen from Table 3.1, while six out of the 10 native speakers spoke Standard Southern British English, two subjects had a Yorkshire, and one each a South Wales and a Lancashire dialect background. All Norwegian speakers were considered sufficiently proficient in English with a first exposure to English no later than at the age of 10. Two of them had spent two years on an English school (speakers NFH and MBE, the latter having lived in the USA from 3.5 to 7 years of age). Spilková (2014) assessed non-native fluency in her corpus by measuring articulation rate expressed as the number of syllables per second (including repairs and false starts but excluding pauses). She reports for the 10 Norwegian speakers a mean articulation rate of 4.4 syllables/s. Nine speakers had articulation rates varying between 3.2 and 5.0 syllables/s; for bilingual subject AM a value of 5.4 syllables/s was found (Table 3.1). In view of the absence of similar information for the native speakers in the existing corpus, we measured

Table 3.1: Speaker details including role as speaker A or B in dialogues and number of transcripts used in reading task. AOL = Age of Learning (first exposure to English at age in years). Art. rate = articulation rate in spontaneous speech (syllables/s). Qualitative analysis of spontaneous speech involved only speaker A recordings, and quantitative analysis both speaker A and B recordings.

L1	Pair #	Speaker	Age	Sex	Dialect/AOL	Transcript #	Art. rate
English	1A	JE	25	F	SSBE	1	4.6
	1B	GMH	29	F	SSBE	2	5.1
	2A	SG	26	F	SSBE	3	5.3
	2B	AW	27	F	SSBE	4	4.6
	3A	PD	21	M	SSBE	5	4.8
	3B	VS	32	F	SSBE	6	4.7
	4A	RB	22	M	Lancashire	7	5.9
	4B	CW	26	M	Yorkshire	8	4.3
	5A	RA	27	F	South Wales	1	5.6
	5B	IM	23	F	Yorkshire	3	5.9
Norwegian	1A	EA	28	M	10	6	3.6
	1B	AM*	26	M	Bilingual	10	5.4
	2A	AH	22	M	5	9	5.0
	2B	MBG	23	F	7	2	4.3
	3A	NFH	19	F	10	3	4.6
	3B	MBE	19	M	3	7	4.6
	4A	JOO	22	M	8	8	4.0
	4B	EV	23	M	8	5	3.2
	5A	IET	25	F	10	4	4.4
	5B	MSE	25	F	10	11	4.4

* Bilingual speaker AM was excluded from analysis.

their speech rate using the same method. Based on samples of approximately 150 syllables per speaker, a mean native speech rate of 5.1 syllables/s was measured. Thus, native speakers delivered a higher speech rate than non-natives, although not dramatically so (on average 0.7 syllables/s faster).

3.3 Qualitative analysis

3.3.1 Method – selected material and definition of stop properties

Annotation of the speech material was performed by the author using Praat (Boersma and Weenink 2014). Qualitative evaluation involved auditory and

visual inspection of speech waveform and spectrogram. Generally, Praat's default values were used for the spectrogram (frequency range 0–5,000 Hz). In a number of cases, for the analysis of friction noise and release bursts, a frequency range of up to 12,000 Hz was chosen. This analysis of the dialogue material involved only speaker A productions, that is, material from five native and five non-native speakers (see Table 3.1). Beyond time constraints, the reason for this limitation was the fact that speaker B had a more passive role in the picture replication task and, as a consequence, fewer and often shorter contributions. Analysis of the read news transcripts, however, comprised recordings from both speaker classes A and B. Results for speakers A and B will be presented in separate sections.

Due to the heterogeneity of the speech material, particularly of the spontaneous speech, it was not feasible to take phonetic context into account as a systematic factor for the speech sounds to be investigated (/p/, /t/, /k/ and /b/, /d/, /g/). Therefore, stop consonant tokens were collected from the start of a recording until a representative number was reached. For the read news transcripts, the number of annotations was at least 60 for each of the 10 natives and 65 for each of the 9 non-natives. From the dialogue material, 70 cases were analysed for each of the native and non-native speakers. The total number of analysed speaker A tokens was 1,282 (less than the maximum possible number due to excluded cases, e.g., where stops were completely deleted).

For the analysis, a canonical stop was assumed to possess a complete constriction of the vocal tract (silent for /p, t, k/ and potentially filled with voicing for /b, d, g/) and a release burst. The following stop properties were specified in a binary way (present vs. absent; strong vs. weak), briefly presented in the following overview. They will be specified in more detail in the description of the results (cf. illustrations in Figure 3.1):

(a) Glottalisation – substitution of oral closure by some form of glottal constriction (classified as present/absent)
(b) Complete closure – a complete constriction of the vocal tract, to the exclusion of incomplete constrictions filled with some form of friction, for example, nasal friction (present/absent)
(c) Voicing – voicing during (part of) the closure having a relatively strong amplitude (present/absent)
(d) Burst – brief period of noise from consonantal release (present/absent)
(e) Strong burst – specifying relative strength of the consonantal release burst (strong/ weak)
(f) Local friction – presence versus absence of local friction at the release (present/absent)
(g) Friction voicing – voicing of local friction (present/absent)

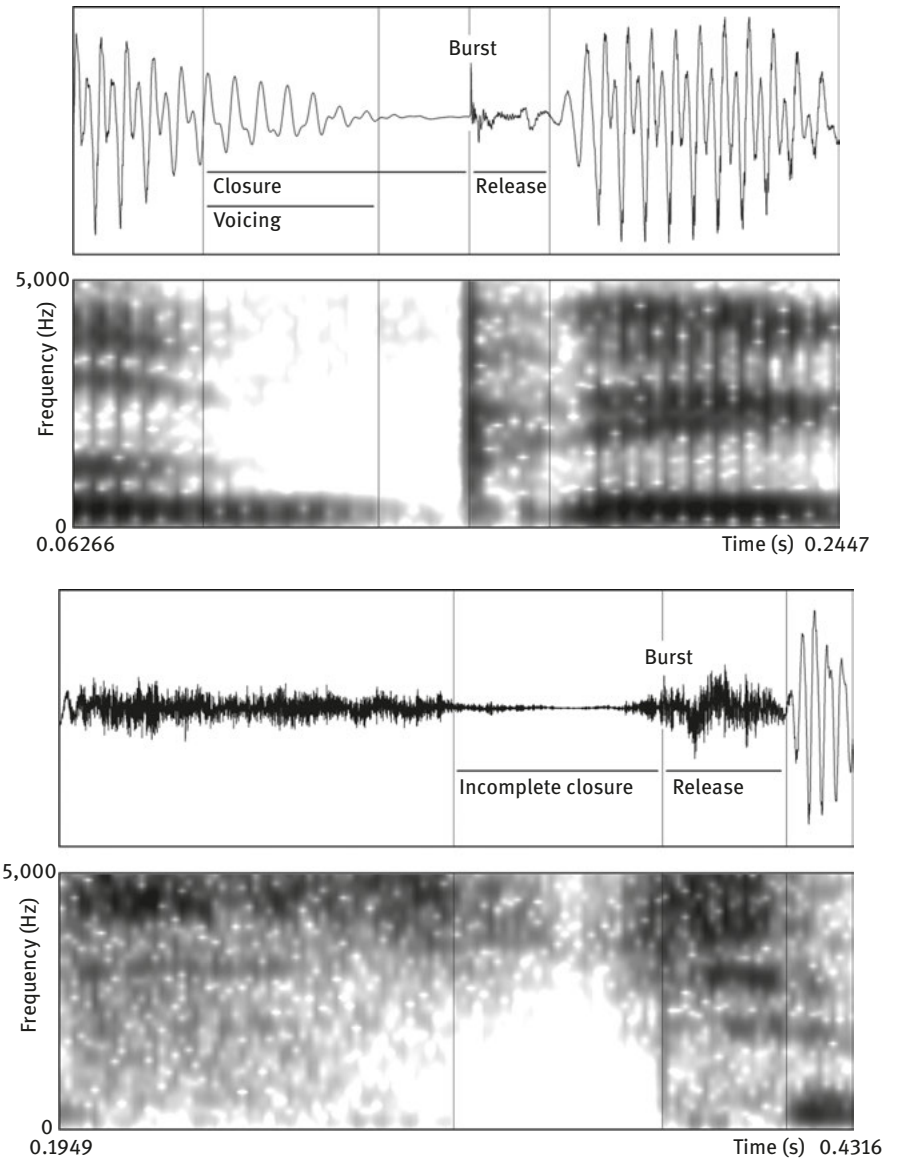

Figure 3.1: Top: Realisation of closure with voicing and release of /p/ in <the picture>. Strong burst with local friction. Bottom: Realisation of closure and release of /t/ in <seems to>. Weak burst (below the letter <u>).
Both tokens are from spontaneous speech produced by each of two non-native speakers.

3.4 Results of qualitative analysis

We shall start describing the results for read and spontaneous speech produced by speaker A subjects (five natives and five non-natives). After that, analysis results on read speech from speaker B productions will be presented (five native and four non-native speakers). Chi-square tests will be used to test the two experimental hypotheses formulated above. First, differences between native and non-native values pooled across the two speaking styles will be analysed. Second, for each stop property, we will test the interaction between the factors speaking style and L1, that is, whether the degree of change from read to spontaneous is different for the two groups of speakers.

3.4.1 Speaker A results

Main results from the qualitative evaluation are presented in Figure 3.2. First of all, let us have a look at stop glottalisation. This category comprises full glottal stops as well as tokens where the consonantal constriction contained one or several glottal pulses. All stops categorised as glottalised represent substitution of a canonically oral constriction by a glottal one; occasional occurrences of glottalisation due to the presence of a boundary or utterance-final word position are not included. It can be seen that stop glottalisation was much more frequent in native than in non-native production (pooled across read and spontaneous speech 26.0% vs. 3.8%; $\chi^2(1) = 93.2$; $p < 0.001$). Further, there was a significant interaction between speaking style and L1 due to the increase in glottalisation from read to spontaneous for the English speakers (ENG-A read: 21.4% vs. spontaneous 30.0%; correspondingly NOR-A: 3.8% vs. 3.7%; $\chi^2(1) = 9.05$; $p = 0.003$).

The next property investigated was the type of consonantal constriction which could be produced as a complete closure or be incomplete with some form of noise (e.g., nasal airstream or friction noise in the stop of an /st/ cluster). A complete closure could be voiceless or filled with voicing. While not more than 28.2% of all native productions contained this form of constriction, the number of occurrences for the non-natives was significantly higher (35.8%; $\chi^2(1) = 4.04$; $p = 0.045$). There was no significant L1 by speaking style interaction (speaking style effects of 5.9% for ENG-A vs. 1.4% for NOR-A; $\chi^2(1) = 2.82$; $p = 0.093$).

The criterion for assigning the label *Voicing* to a segment was the presence of relatively strong periodicity during a closure or part of it. Segments with only short periodicity of small amplitude that could be interpreted as passive voicing were labelled as voiceless. Also with regard to closure voicing, differences

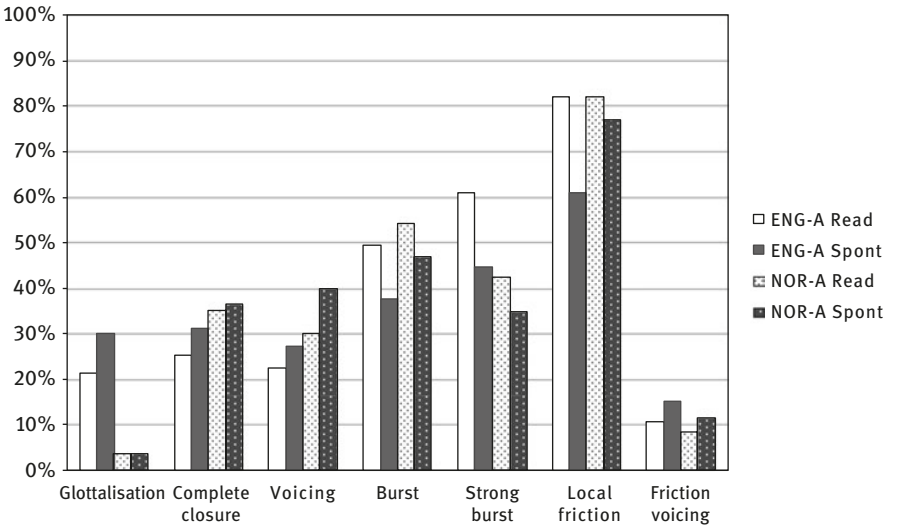

Figure 3.2: Stop properties (occurrences in %) in read and spontaneous speech according to qualitative evaluation (see text). ENG-A/NOR-A = 5 English/5 Norwegian speakers (A in dialogue).

between the two speaker groups were observed. In native speech, 24.7% of the stop productions were classified as voiced versus 35.3% in non-native tokens (a significant difference: $\chi^2(1) = 8.48$; $p = 0.004$). Both groups showed speaking style effects and, although for the Norwegian speakers, the effect was larger, there was no significant L1 by speaking style interaction (spontaneous vs. read for ENG-A: 4.6%, for NOR-A: 9.8A%; $\chi^2(1) = 1.84$; $p = 0.175$). Closer inspection of the data split up into fortis and lenis stops revealed that the amount of voiced fortis stops in spontaneous non-native productions was unexpectedly high (37.8% vs. 19.8% for the natives). For read speech, the native–non-native difference was substantially smaller (NOR-A: 13.8% vs. ENG-A: 12.0%). Statistical analysis of voicing in fortis stops only revealed significant effects of both L1 and its interaction with speaking style ($\chi^2(1) = 12.6$; $p < 0.001$ and $\chi^2(1) = 8.19$; $p = 0.004$, respectively).

Indication of a release burst was the presence of brief broadband noise in the spectrogram and one or more small spikes in the speech waveform. Not only in stops preceding vowels or heterorganic consonants but in many cases also before homorganic fricatives release bursts could be separated from the following segment, for example, in frequently occurring <it's>. Different manifestations as single or multiple burst and also glottal burst were subsumed under the property *Burst*. The data for this property showed that also in this respect stop

consonant manifestations often differed from the prototypical form. In approximately half of all native as well as non-native tokens bursts were observed (ENG-A: 43.1%; NOR-A: 50.2%), the between-group difference being non-significant ($\chi^2(1) = 2.33$; $p = 0.127$). Additionally, both speaker groups had larger amounts of bursts in read versus spontaneous speech resulting in a non-significant L1 by speaking style interaction (ENG-A: 11.8%, NOR-A: 7.3%; $\chi^2(1) = 1.03$; $p = 0.310$).

A further classification criterion related to consonantal release noise was the coding of a burst as described in the previous paragraph as strong or weak. Indicators for this coding were relative duration and amplitude in both speech waveform and spectrogram. Pooled across speaking styles there was a significant L1 effect (ENG-A: 53.2% vs. NOR-A: 38.6%; $\chi^2(1) = 4.83$; $p = 0.028$). Although the effect of speaking style was larger for the native than for the non-native group, its interaction with the factor L1 did not reach statistical significance (ENG-A: 16.3% vs. NOR-A: 7.8%; $\chi^2(1) = 2.96$; $p = 0.086$).

In addition to the specification of a stop release as strong or weak, it was investigated if the release burst was accompanied by friction noise generated at the place of articulation or not. It turned out that pooled across speaking styles this *local friction* showed similar values for both speaker groups (ENG-A: 70.7%, NOR-A: 79.4%; $\chi^2(1) = 1.85$; $p = 0.174$). Only for the native speakers its occurrence was strongly dependent on speaking style, therefore giving rise to a significant L1 by speaking style interaction (speaking style effects of 20.9% for ENG-A vs. 5.0% for NOR-A; $\chi^2(1) = 9.71$; $p = 0.002$).

Finally, the local friction was classified as voiceless or voiced. Voicing could be seen to occur both in lenis and fortis stop consonants, for example, in reduced /t/ in intervocalic or post-nasal position. In native and non-native tokens, this type of voicing was equally rare (ENG-A: 12.7%, NOR-A: 10.1%; $\chi^2(1) = 1.28$; $p = 0.258$). Also, there was no significant L1 by speaking style effect (ENG-A: 4.3% vs. NOR-A: 3.0%; $\chi^2(1) = 0.23$; $p = 0.633$).

3.4.2 Speaker B results

The qualitative analysis presented thus far involved data collected from both read and spontaneous speech produced by five native and five non-native subjects. By analysing the same speakers in both speaking style conditions, any potentially confounding speaker idiosyncrasies have thus been excluded. All those 10 subjects performed as speaker A in the picture task. Additionally, the BBC news transcripts read by speaker B subjects were analysed (five natives and

four non-natives, the fifth Norwegian subject being excluded due to his bilingual background). The results for those nine speakers may give some indication of how consistent the production patterns found thus far are. Due to the absence of data on spontaneous speech for those speakers it will, of course, not be possible to compare L1 by speaking style interaction effects.

It can be seen from Figure 3.3 that for both subgroups of native and non-native speakers, the proportion of glottalised stops is very similar to what was found for the speaker A subjects (ENG-B: 19.5% vs. NOR-B: 5.1%; cf. Figure 3.2). The difference between the two groups was significant ($\chi^2(1) = 17.4$; $p < 0.001$). This is also true for the frequency of occurrence of complete closures in stop production. Although percentages are somewhat higher than for speakers A, also here complete closure occurred more often in non-native than in native read speech (ENG-B: 29.4% vs. NOR-B: 45.3%; $\chi^2(1) = 6.16$; $p = 0.013$). Also as to voicing, speaker B subjects behaved similar to speaker A subjects: 20.5% of the constrictions in native productions were classified as voiced versus 27.2% in non-native stops. This difference did, however, not reach statistical significance. In view of the conspicuous number of voiced fortis stops for speakers A mentioned above, the data were split up into fortis and lenis. For speakers B, however, natives and non-natives had similar numbers of voiced fortes (ENG-B: 11.2% vs. NOR-B: 16.2%; $\chi^2(1) = 1.58$; $p = 0.209$; cf. ENG-A: 19.8% vs. NOR-A: 37.8%).

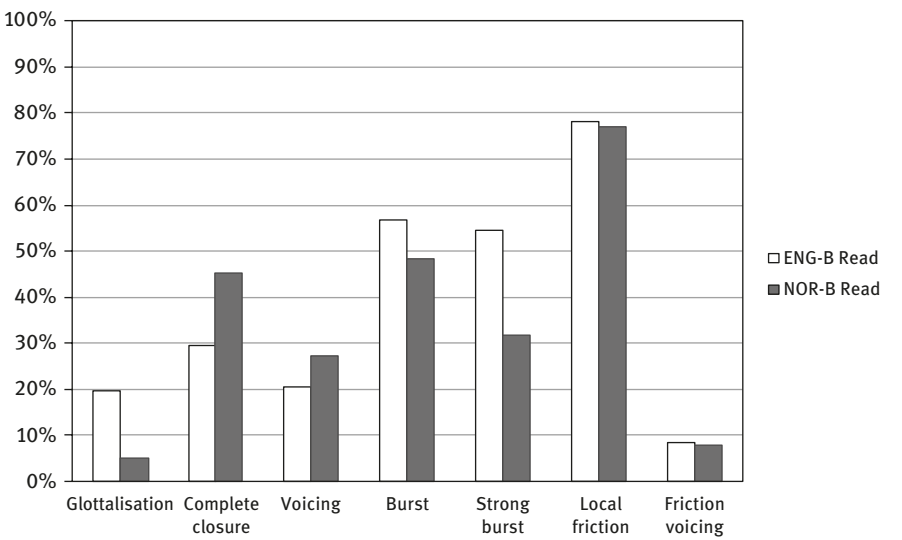

Figure 3.3: Stop properties (occurrences in %) in read speech according to qualitative evaluation (see text). ENG-B/NOR-B = 5 English/4 Norwegian speakers (B in dialogue).

An exception to the emergence of similar tendencies for speakers A–B observed thus far is the frequency of the occurrence of a burst. While a stop burst was present somewhat more often in native than non-native speaker B tokens (ENG-B: 56.8% vs. NOR-B: 48.4%), the opposite pattern was found for speaker A subjects (ENG-A: 49.5% vs. NOR-A: 54.2%). Note, however, that both these two tendencies were non-significant. The results for the remaining three properties were in line with speaker A observations. While strong bursts were more common in native than in non-native stops (ENG-B: 54.6% vs. NOR-B: 31.7%; $\chi^2(1) = 5.30$; $p = 0.021$), local friction occurred almost equally often (ENG-B: 78.2% vs. NOR-B: 77.2%; $\chi^2(1) < 1$). Voicing of local friction was, again, not very frequent and occurred equally often in native and non-native stops (ENG-B: 8.3% vs. NOR-B: 7.8%; $\chi^2(1) < 1$).

3.4.3 Conclusions from the qualitative analysis

The qualitative analysis presented above has revealed relatively consistent patterns in the reduction of stop consonants. To a large degree, the data could be interpreted as showing different reduction behaviour for natives compared to non-natives. Non-native production patterns were characterised by less glottalisation, more frequent occurrences of complete consonantal closure and closure voicing (the latter particularly in fortis stops). Further, non-native bursts were generally less often classified as strong. With respect to the remaining three properties, presence of a release burst, local release friction, and voicing of local friction, native and non-native stops did not differ. Interpretation of the observed tendencies in terms of phonetic reduction is not straightforward. The observation of less frequent complete closure in native versus non-native stops seems to comply with stronger reduction. In contrast, the simultaneous higher frequency of strong bursts in native stops is counterintuitive. Further, it is not clear beforehand if the absence of voicing during a stop should be taken as an indication of reduction or not.

Most of the L1 by speaking style interactions appeared to be non-significant, thus indicating similar style-dependent reduction behaviour for the English and Norwegian speakers. Exceptions were significant interactions for the properties glottalisation and local release friction. Additionally, substantial differences in reduction behaviour were found for voicing in fortis stops. Non-native fortis stops produced spontaneously were characterised by particularly frequent closure voicing. It remains to be seen if the quantitative analysis will confirm these results.

3.5 Quantitative analysis

3.5.1 Method – measurements and variables

A fundamental problem in a quantitative analysis of speech material like the type used for this investigation is its heterogeneity and the lack of reference points. Due to varying segmental and prosodic context conditions, absolute segment durations, for example, of the pertinent stops, have only limited information value. Measures involving signal intensity are not meaningful in an absolute sense but should be defined relative to reference points in neighbouring context.

Annotated segments described in Section 3.3.1 were taken as a point of departure for performing different types of measurements. The first type involved the intensity contour of the speech waveform. Using a Praat script, intensity was calculated with a time step of 1 ms and, by setting minimum pitch to 600 Hz, an analysis window of 5.3 ms in order to obtain sufficient time resolution. From the intensity contour in the annotated segments, the following intensity variables were calculated (cf. Figure 3.4): slope of intensity fall from the beginning of the consonantal closure to 10 ms into the closure in dB/cs (*Slope fall*) and slope of intensity rise from 10 ms before consonantal release to point of release in dB/cs (*Slope rise*). The unit of cs (= 10 ms) was chosen to facilitate interpretation of measurement results as illustrated in the figure. For closures shorter than 20 ms, end of fall and beginning of rise coincided in the point of minimum intensity in the closure. In addition, three intensity difference measures were calculated. The first one, *MaxMean*, was defined as the difference between the intensity maximum of the segment preceding the closure and mean intensity of the closure. *MaxMeanMax* represents the average difference between mean intensity of the closure and the intensity maximum of the segment preceding the closure and that of the consonantal release, respectively. The third measure, *MeanMax*, quantifies the relative intensity of the consonantal release, that is, the difference between mean intensity of the closure and maximum intensity of the release.

The second type of measurement was in the temporal domain and involved closure and release duration. In addition, to counteract local speaking rate and other confounding effects, a relative measure of release duration was calculated as the ratio of release duration divided by closure + release duration expressed in per cent (*Release%*).

Thirdly, proportion of voicing during the consonantal closure was measured by the following procedure involving two steps. In the first step, the duration of periodicity during the closure was expressed as percentage of closure duration (*Voicing%*). In the second step, in order to account for varying voicing amplitude, the proportion of voicing was given a weight factor. Weight factors were

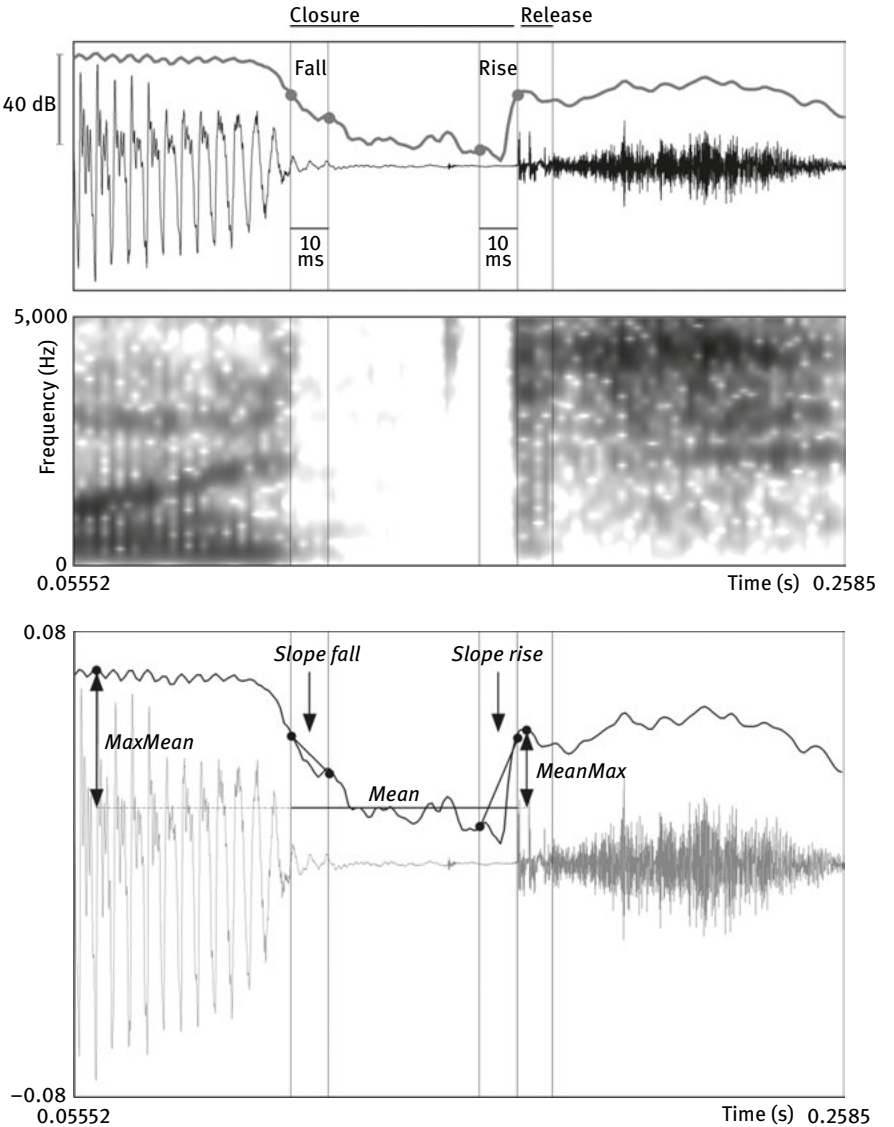

Figure 3.4: Top: Waveform and spectrogram of [ɒtʃ] in <watching> pronounced by female native speaker. Overlaid intensity contour illustrates points of measurements for *Slope fall* and *Slope rise* (see below and text). Bottom: Derived intensity measures. *MaxMean* and *MeanMax* = intensity difference between mean of closure phase and maximum of preceding and following segment, respectively. These were used to calculate *MaxMeanMax* = (*MaxMean* + *MeanMax*)/2.

calculated by running a Praat script. First, to exclude extraneous noise in the closure, the signal was band-pass filtered between 50 and 500 Hz. Subsequently, intensity of closure voicing and the preceding segment was calculated with a time step of 1 ms and an analysis window of 5.3 ms. Finally, a weight factor was defined as mean intensity of the voicing divided by intensity of the preceding segment. Weighted proportion of closure voicing (*VoicingWT*) was calculated as the product of the weight factor and the variable *Voicing%*.

3.6 Results of quantitative analysis

3.6.1 Selection of material and statistical treatment

For the quantitative analysis, speech material from all 10 English native speakers and nine Norwegian speakers was used (cf. Section 3.2.2). The total number of observations was 2,376. All calculations were based on stops that had a release. In addition, to ensure equal conditions across samples only cases with *Slope fall* < 0 and *Slope rise* > 0 were included. These selection criteria were used for all intensity variables. While both for these variables and for the duration variables tokens with consonantal release were selected, only criterion for selection of the voicing variables was the presence of genuine closure voicing (e.g., excluding cases of voiced nasal airstream).

Statistical analysis was carried out using the R program's package lme4 (R Core Team, 2012) to calculate Linear Mixed Effects Models (Barr et al. 2013; see also Baayen 2008). Fixed factors were L1 (English, Norwegian), speaking style (read, spontaneous), and stop type (fortis, lenis) with by-speaker random slopes and intercepts for the latter two factors. Each analysis involved comparison of a model with those factors with a model without the factor under scrutiny. Similarly, each of three two-way interactions (L1 × speaking style, L1 × stop type, and stop type × speaking style) was analysed. To assess the significance of single factors and interactions, likelihood ratio tests were performed. As a rule, only statistically significant ($\alpha = 0.05$) results are reported.

3.6.2 Unexpected technical artefacts

During the process of quantitative data analysis, a serious technical problem was discovered. Exploring the distribution of spectral energy in stop closures,

native and non-native stop productions appeared to differ consistently, the Norwegian speakers' productions showing more energy below 500 Hz than native speakers' stops. The effect seemed too beautiful to be true. It was investigated if it possibly could be due to differences in technical equipment rather than L2-related phenomena. While all recordings with native speakers in Bristol were made using a Shure WH20 headset, microphones in the Norwegian studio were MILAB LSR-1000. To examine possible effects of microphone frequency characteristics, silent portions from both types of recordings were collected and analysed with Praat, comparing the amount of energy below and over 500 Hz. Measurements revealed that in the MILAB LSR-1000 recordings, the relative level of energy below 500 Hz was on average approximately 9 dB higher than in the Shure WH20 signals. A further step in learning more about technical conditions involved recording a speaker via two channels simultaneously, using a Sennheiser HS 2-1 headset and a Shure KSM44 microphone in the Norwegian studio. Comparison of the energy distribution in the two recordings showed a higher low-frequency energy level for the Shure microphone, also here amounting to about 9 dB.

Inspection of technical data for the above-mentioned microphones suggested that the intensity bias effects were due to a combination of effects. Whereas the MILAB LSR-1000 has a generally very flat frequency response, the Shure WH20 headset drops increasingly energy from 200 Hz downwards. This would explain the differences found in the Bristol versus Trondheim recordings. On the other hand, the similar effects measured using Sennheiser HS 2-1 and Shure KSM44 cannot be due to different specifications because they have very similar frequency response curves in the low-frequency region. Here, the reason must be a phenomenon occurring in pressure-gradient microphones known as proximity effect, that is, increasingly boosted lower frequencies at decreasing speaking distances. Also, MILAB LSR-1000's frequency response is affected by this phenomenon.

It was of paramount importance for the interpretation of the outcomes of the measurements to explore the possible bias introduced by the technical artefacts. Therefore, an attempt was made to simulate the boosting effect found in the recordings using the graphic equaliser function in Adobe Audition (Version 3.0). All 10 BBC news transcripts spoken by native speakers were chosen as test material. In each recording, frequencies in the region 80–500 Hz were amplified with a maximum amount of 8 dB at 125 Hz gradually decreasing to 3 dB at 500 Hz. Subsequently, the manipulated recordings were analysed in the same way as the originals. Investigation of the intensity measures revealed a bias of 2–3 dB towards lower values, a result that must be taken into account in the following discussion of the results. The effect of the signal manipulation on the intensity weighted voicing variable (*VoicingWT*) appeared to be approximately 1.5%. As we will see below, this bias is relatively small in comparison to the size of the present effects.

3.6.3 Intensity measurement results

Results of the intensity measurements are presented in Table 3.2. For each variable, we shall first have a look at the trends as they emerged from our analysis and subsequently discuss how they can be assumed to be affected by the difference in recording conditions. Since none of the investigated interactions (L1 × speaking style, L1 × stop type, and stop type × speaking style) reached statistical significance, they won't be mentioned in the following description of the results.

For both speaker groups and both stop types, the slope of intensity fall into the stop closure was less steep in spontaneous than in read speech. Pooled across conditions, the speaking style factor appeared to be significant ($\chi^2(1) = 6.18$; $p = 0.013$). While this was not the case for the factor stop type ($\chi^2(1) = 3.29$; $p = 0.070$), the effect of L1 reached statistical significance ($\chi^2(1) = 9.97$; $p = 0.002$). Pooled across conditions, native versus non-native *Slope falls* were 2.6 dB/cs steeper. Given the biasing influence of recording conditions, however, this value cannot be considered to be realistic. Our investigation of this factor showed that boosted frequency components under 500 Hz would artificially reduce the value of a variable like *Slope fall*. This means, for example, that the tabled value for *Slope fall* of –9.0 dB/cs in the read non-native material *ceteris paribus* would have approached the value of –12.6 dB/cs observed for the natives. Since a realistic estimate of the size of the bias is about 2–3 dB/cs it seems reasonable to assume that the apparent effect of language background is nullified. Importantly, recording conditions were the same for the two speaking styles within each of the two speaker groups. Therefore, the factor speaking style as well as its interaction with L1 was not affected by the recording artefacts.

The pattern for the *Slope rise* values is less consistent. For this variable, only the steeper rise in native read versus spontaneous stops (2.3 dB/cs) seems to indicate a speaking style induced difference. However, neither this nor any of the other experimental factors reached statistical significance. Only the factor stop type was marginally significant ($\chi^2(1) = 3.08$; $p = 0.079$). The overall difference between non-native and native values amounted to –2.6 dB/cs. Presumably due to relatively large individual variation, the native language factor did not reach significance. Anyway, adding the estimated bias value to the apparent difference of –2.6 dB/cs implies annihilation of any potential effect of language background.

As to the difference between mean intensity during stop closure and intensity maximum in the preceding segment, *MaxMean*, only the stop type factor was significant ($\chi^2(1) = 21.7$; $p < 0.001$). *MaxMean* values were on average 1.3 dB higher in read versus spontaneous speech but the effect did not reach statistical significance. Calculation of the overall native–non-native

Table 3.2: Intensity measures of stops in read and spontaneous speech. ENG/NOR = 10 English/9 Norwegian speakers. *Slope fall*/*rise* = fall/rise of intensity in dB/cs. Difference measures in dB: *MaxMean* = maximum previous segment – mean closure; *MeanMax* = maximum release – mean closure; *MaxMeanMax* = (*MaxMean* + *MeanMax*)/2 (see text). Standard deviations in italics.

Fortis

		Slope fall		***Slope rise***		***Max-Mean***		***MaxMean-Max***		***Mean-Max***		***n***
ENG	Read	−12.6	*7.0*	19.8	*11.5*	31.7	*9.8*	27.5	*9.4*	23.3	*10.5*	321
	Spont	−10.6	*5.9*	17.5	*11.1*	29.9	*10.0*	24.7	*8.5*	19.5	*9.3*	322
NOR	Read	−9.0	*5.3*	15.4	*10.2*	29.7	*7.9*	24.0	*7.6*	18.2	*9.4*	251
	Spont	−8.4	*5.1*	16.1	*10.5*	28.3	*9.3*	22.5	*8.4*	16.8	*9.6*	367

Lenis

		Slope fall		***Slope rise***		***Max-Mean***		***Max-MeanMax***		***Mean-Max***		***n***
ENG	Read	−9.2	*5.6*	15.5	*13.3*	21.0	10.6	17.0	*10.2*	13.1	*10.9*	47
	Spont	−8.7	*5.4*	15.9	*12.3*	22.8	*9.8*	16.4	*8.5*	10.0	*8.8*	55
NOR	Read	−9.5	*8.5*	13.1	*9.7*	20.8	*9.5*	14.7	*9.2*	8.7	*9.9*	28
	Spont	−6.6	*5.2*	12.2	*6.9*	20.2	*7.2*	13.6	*7.0*	6.9	*7.9*	19

difference showed a somewhat larger value for the native realisations (1.3 dB; n.s.). Adjusting this value in order to compensate for technical artefacts would change the polarity of the difference. In view of the present effect sizes, however, it seems unlikely that the actual native language effect would turn out to be significant.

Inspection of the *MaxMeanMax* variable, average stop closure intensity related to preceding and following intensity maximum, revealed some consistent effects. *MaxMeanMax* was on average 2.0 dB larger in read versus spontaneous speech and 8.7 dB larger in fortis versus lenis stops. Both factors were statistically significant ($\chi^2(1) = 4.59$; $p = 0.032$ and $\chi^2(1) = 23.2$; $p < 0.001$, respectively). Particularly with regard to the effect of speaking style, it is important to keep in mind that it would basically remain the same after adjusting for the influence of microphone differences. Therefore, the speaking style effect of 2.0 dB can be regarded as reliable. The overall effect of L1 was 2.3 dB for the English versus Norwegian productions and statistically non-significant. Estimating the unbiased effect, it seems probable that this result would remain unchanged.

Intensity maximum of consonantal release relative to average stop closure energy (*MeanMax*) was significantly stronger in read than in spontaneous speech (2.8 dB; $\chi^2(1) = 5.92$; $p = 0.015$) and in fortis versus lenis stops (9.1 dB;

$\chi^2(1) = 23.2$; $p < 0.001$). In the native versus non-native material, larger *MeanMax* values were observed (3.3 dB; $\chi^2(1) = 6.19$; p = 0.013), but also the value of this variable should be adjusted to compensate for different recording conditions. Again, it seems safe to assume that the actual L1 effect would turn out to be non-significant.

3.6.4 Summary of intensity measurement results

The pattern of results on intensity was relatively clear. As to the effect of L1, the data seem to suggest that there is no difference between native and non-native behaviour. Taking into account the bias introduced by different recording conditions, it can be argued that the significant effects found for the variables *Slope fall* and *MeanMax* were artificial. Neither for any of the other intensity variables, a significant L1 effect could be found or assumed to exist. In contrast, speaking style could be shown to affect intensity values. *Slope fall*, *MaxMeanMax*, and *MeanMax* were all reliably larger in read speech in comparison with spontaneous tokens. Although speaking style effects were generally somewhat larger for native than non-native stops, all L1 × speaking style interactions were statistically non-significant. Therefore, the results did not allow the conclusion of different native versus non-native reduction behaviour in spontaneous versus read speech. Not surprisingly, the well-established effect of stop type (fortis/lenis) was observed in the present data in the form of significant or marginally significant statistical results.

3.6.5 Duration and voicing measurement results

Duration measurements were not affected by different recording conditions, so that results can be taken at their face value. The only variable sensitive to that factor was the intensity weighted voicing variable *VoicingWT*. For that variable, we will have to consider to what degree the results could be invalidated by recording artefacts. None of the investigated interactions (L1 × speaking style, L1 × stop type, and stop type × speaking style) were statistically significant, with the L1 × speaking style interaction for *Release*% as the only exception (see below).

Inspection of the results presented in Table 3.3 shows that closure durations were similar for the two groups of speakers, the overall difference of 6 ms being non-significant. This result is due to the deviating pattern of similar values for non-native fortis in the two speaking styles (46 ms). Running statistical analyses

separately for fortis and lenis showed that the between-group difference for the latter stop type was significant (9 ms longer closure for the non-natives; $\chi^2(1) = 6.82$; $p = 0.009$), in contrast to the former (4 ms). Also the factor speaking style reached only significance for the lenis stops (on average 7 ms longer closure duration in spontaneous vs. read lenis stops; $\chi^2(1) = 4.99$; p = 0.026). Finally, stop closures were generally longer in fortis than lenis consonants (6 ms; $\chi^2(1) = 6.84$; $p = 0.009$).

Stop release was significantly longer in native than non-native realisations (on average 11 ms; $\chi^2(1) = 21.7$; $p < 0.001$). As can be seen from the table, mean release durations varied only little or not at all with speaking style (n.s.). Stop category was a strong determiner of release duration, fortis release being 15 ms longer than lenis release ($\chi^2(1) = 22.4$; $p < 0.001$).

The *Release*% variable quantifies relative duration of the release in the stop consonant. For native stops, its value was considerably higher than for non-native productions (11.1%; $\chi^2(1) = 25.6$; $p < 0.001$), both in fortis (12.0%) and in lenis stops (9.8%). The latter result is confirmed by the non-significant L1 × stop type interaction ($\chi^2(1) < 1$). It is also apparent that relative release duration was longer in fortis versus lenis stops (7.2%; $\chi^2(1) = 13.6$; $p < 0.001$). The effect of speaking style was relatively small (pooled across conditions 3.1%) and only marginally significant ($\chi^2(1) = 3.44$; $p = 0.064$). It was different for the two groups of speakers

Table 3.3: Segment durations and voicing in stops in read and spontaneous speech. ENG/NOR = 10 English/9 Norwegian speakers. *Closure* and *Release* duration in ms. *Release%* = release duration relative to closure + release duration in per cent. *Voicing%* = proportion of closure voicing in per cent of closure duration. *VoicingWT* = proportion of voicing weighted by relative intensity (see text). Standard deviations in italics.

Fortis

		Closure		*Release*		*Release%*		*n*	*Voicing%*		*VoicingWT*		*n*
ENG	Read	36	*23*	38	*29*	47.6	*23.9*	188	31.8	*39.6*	25.0	*32.6*	171
	Spont	45	*24*	38	*30*	41.6	*21.2*	275	32.6	*38.8*	26.0	*32.4*	271
NOR	Read	46	*21*	25	*25*	31.5	*20.1*	159	45.4	*36.2*	35.6	*29.8*	162
	Spont	46	*23*	24	*21*	32.4	*19.3*	297	48.6	*35.1*	40.2	*31.0*	332

Lenis

		Closure		*Release*		*Release%*		*n*	*Voicing%*		*VoicingWT*		*n*
ENG	Read	29	*17*	19	*13*	39.5	*18.4*	62	65.7	*39.9*	53.6	*35.1*	102
	Spont	38	*22*	19	*23*	30.1	*20.0*	63	74.1	*31.7*	60.9	*29.6*	84
NOR	Read	40	*29*	13	*12*	26.5	*19.7*	50	78.5	*36.7*	69.2	*35.7*	69
	Spont	46	*27*	11	*10*	22.5	*18.7*	31	73.4	*41.3*	63.6	*36.8*	40

(ENG: 6.2%; NOR: –1.2% across fortis and lenis stops), as confirmed by the significant L1 × speaking style interaction ($\chi^2(1) = 4.36$; $p = 0.037$).

Measurements of proportion of voicing during stop closure (*Voicing*%) revealed generally stronger voicing in the non-native stops. Pooled across conditions the between-group difference was 9.5%, according to statistical analysis marginally significant ($\chi^2(1) = 3.54$; $p = 0.060$). Although no significant L1 × stop type interaction was found ($\chi^2(1) = 1.25$; $p = 0.263$), separate analyses for fortis and lenis indicated different tendencies for the two stop types. Fortis stops produced by non-natives featured 15.3% more closure voicing than those from natives ($\chi^2(1) = 4.94$; $p = 0.026$). In contrast, the corresponding L1 effect for the lenis category was not significant (7.2%; $\chi^2(1) < 1$). As could be expected, lenis stops contained significantly more voicing than their fortis counterparts (31.8%; $\chi^2(1) = 29.8$; $p < 0.001$). The factor speaking style did not reach significance.

The outcomes for the weighted voicing variable (*VoicingWT*) were similar to those observed for *Voicing*%. The overall effect of L1 turned out to be significant, with higher values for non-native productions (9.0%; $\chi^2(1) = 4.72$; $p = 0.030$). Again, the effect was stronger for fortis (13.1%; $\chi^2(1) = 4.60$; $p = 0.032$) than for lenis stops (10.3%; $\chi^2(1) < 1$). The size of these effects seemed large enough to justify the conclusion that they were robust and not endangered by the bias due to technical artefacts. Speaking style affected weighted voicing marginally with somewhat higher values in spontaneous versus read speech (1.7%; $\chi^2(1) = 3.47$; $p = 0.063$). Stop category was a significant factor, lenis values surpassing fortis values by 28.2% ($\chi^2(1) = 31.3$; $p < 0.001$).

3.6.6 Summary of duration and voicing measurement results

Different from what was found for intensity-related variables, analysis of duration and voicing parameters revealed differences between native and non-native behaviour. Non-natives had longer closure durations in lenis stops, and their stop release was reliably shorter in both fortis and lenis stops. For this speaker group, also proportion of closure voicing was found to be larger than for the native speaker group. This was particularly conspicuous in fortis stop tokens. The well-investigated effect of fortis versus lenis on segment durations and proportion of closure voicing was apparent in the results. In contrast, the effect of speaking style on stop realisations was less consistent and generally rather weak. Further, with the exception of the significant L1 × speaking style interaction for relative release duration (*Release*%), all interactions were non-significant. Therefore, it can be concluded that the two speaker groups behaved similarly with regard to the effect of speaking style on stop reduction.

3.7 General discussion

The goal of the study presented in this contribution was to explore phonetic reduction of stop consonants in read and spontaneous speech produced by native and non-native (Norwegian) speakers of English. Through the investigation of a number of parameters, from the qualitative and quantitative analyses a complex picture emerged. Glottalisation of stops was virtually limited to native production and seemed to occur more often in spontaneous than in read speech. As to the whole complex of other parameters investigated, differences between native and non-native behaviour were gradual rather than of an absolute nature. The qualitative analysis revealed that for both speaker groups around 70% of the stops had an incomplete closure. At the same time, this type of reduction occurred more frequently in native production. Release bursts were present in about half of the tokens from both speaker groups but were more often weaker in non-native realisations. On the whole, the present qualitative observations are in congruence with Schuppler et al. (2012), who reported a non-canonical realisation occurrence of Dutch word-final /t/ in conversational speech of approximately 88%. Results from the quantitative analysis suggested that intensity-related phenomena were largely unaffected by the factor of language background. In contrast, that factor appeared to play an important role in the temporal organisation of the speech tokens. In non-native tokens, lenis closure durations were longer and release durations generally shorter than in native speech. Consistently larger proportions of voicing were measured in stops produced by the Norwegian speakers, especially after weighting stop voicing proportion with relative intensity.

The effect of speaking style on production patterns was apparent in the qualitative results on stop properties and more clearly in the intensity and duration parameters. Characteristics for spontaneous compared with read speech were less steep fall of intensity into the closure and less intense consonantal release. Both the direction and the size of these effects were in line with the intervocalic sound energy differences in stops produced by a Dutch speaker in van Son and Pols (1999) and in American stops produced in spontaneous and careful (connected and isolated) read speech in Warner and Tucker (2011). At the same time, closure durations were generally longer and only native release durations (expressed as proportion of total stop duration) shorter in spontaneous speech. Further, the already large proportion of voicing in non-native fortis stops was increased in spontaneous production. For lenis stops, opposite patterns were found for natives and non-natives.

A remarkable outcome was the widespread absence of significant interactions between the factors language background and speaking style. This means that natives and non-natives had similar reduction patterns or at least that the present

method has not succeeded in revealing them. Given the type of material in this study, it is clear that one should be careful not to jump to conclusions. Due to the read speech materials' heterogeneity and the use of spontaneous recordings, conditions such as surrounding phonetic context, stress, and speech rate varied. These conditions thus caused the measurements to be affected by experimental noise. Spilková (2014) used the same recordings in her investigation of the three function words *to*, *of*, and *in* produced by the present native speakers of English, Norwegian and (in her work) Czech. Her analysis of the effects of neighbouring segment types on different measures revealed diverging tendencies. While vowel proportions in the words *in* and *to* were found to be affected by right context type, this was not the case in *of*. Further, neither left nor right context type had a significant effect on the proportion of voicing in the fricative in the preposition *of*, and the proportion of release in the plosive in the word *to*. Warner and Tucker (2011) demonstrated that phrase frequency, stress conditions, and to a lesser degree, segmental context (stops before or after /r/, before /l/, before full vowel or schwa) are confounding factors affecting speaking style effects. It seems thus difficult to estimate the possible impact of such effects unless virtually all potentially relevant factors are experimentally controlled. Obviously, in the investigation of spontaneous speech ambitions in that direction would be in diametric contradiction to the goal of such work. In the present study, an attempt was made to counteract the influence of uncontrolled factors by analysing relatively large amounts of tokens. In this way, the influence of such random effects should to a certain degree be cancelled out. Informal testing of the potential influence of varying phonetic context suggested that the present effects are relatively robust. Testing was done by limiting the selection of intensity data to stop tokens followed by a vowel. For the category of fortis stops, this reduced the number of observations from 1,261 to 727 (corresponding to a reduction by 42%). Still, the patterns in the intensity results remained basically the same. The same selection criterion of a following vowel caused a reduction of 54% in the available closure voicing data for fortis stops (from 936 to 435). Here, too, the effects of L1 and speaking style remained in principle unchanged. Since almost all tokens (99%) in the material selected for stop voicing measurement were preceded by voiced segments, also unsystematic variation of preceding context can be ruled out as a biasing factor. These observations seem to suggest that the confounding effect of varying phonetic context in the material used for the quantitative analysis must have been rather limited.

An apparently robust aspect in non-native stop consonant production was the relatively large proportion of closure voicing. Particularly in fortis stops, proportion of voicing was consistently larger than in native English. The question is if this phenomenon can be explained by differences in stop consonant realisation in Norwegian versus English. In light of the evidence from previous studies

presented in the Introduction, it can be concluded that the use of stop voicing is more extensive in Norwegian than in English. However, there doesn't seem to be a straightforward explanation of Norwegian speakers' tendency to particularly voice fortis stops in English. It is clear that it cannot be a question of a simple transfer of native language patterns. Different from the voicing opposition in stops, apart from the pair /f/ – /v/ voiceless fricatives do not have voiced counterparts in the Norwegian phonological system. Possibly this difference in the exploitation of the voicing opposition makes this feature salient to Norwegian speakers and causes some form of overgeneralisation. As regards predicting L2 pronunciation behaviour, it is obvious that the observed phenomenon is hard to model.

A fruitful aspect of this study, at least for the author, was the discovery of the potentially dramatic effects of recording conditions on intensity measures. Microphone frequency characteristics crucially determine intensity contours, as was the case for the present measurement of release intensity and slopes of intensity fall and rise. This insight is not only important for recording speakers at different locations using different microphones. A microphone with a cardioid frequency characteristic will boost low-frequency components relative to higher parts of the spectrum depending of speaking distance. Therefore, within-speaker variation of microphone distance will distort speech signal intensity. To avoid wrong interpretations, it is important to keep such technical conditions under control.

References

Baayen, R. H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Barr, D. J., R. Levy, C. Scheepers & H. J. Tily 2013. Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68. 255–278.

Barry, W. J. & B. Andreeva 2001. Cross-language similarities and differences in spontaneous speech patterns. *Journal of the International Phonetic Association* 31. 51–66.

Best, C. T. 1995. A direct realist view of cross-language speech perception. In W. Strange (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research,* 171–203. Timonium: York Press.

Best, C. T. & M. D. Tyler 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (eds.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, 13–34. Philadelphia: John Benjamins Publishing Company.

Boersma, P. & D. Weenink 2014. Praat: Doing phonetics by computer (Version 5.3.82) [Computer program]. Retrieved 29 July 2014, from http://www.praat.org/.

Bolotova, O. 2003. On some acoustic features of spontaneous speech and reading in Russian (quantitative and qualitative comparison methods). *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 3–9 August 2003, vol. 1, 913–916.

Bondarko, L. V., N. B. Volskaya, S. O. Tananaiko & L. A. Vasilieva 2003. Phonetic properties of Russian spontaneous speech. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 3–9 August 2003, vol. 3, 2973–2976.

Bradlow, A. R., L. Ackerman, L. A. Burchfield, L. Hesterberg, J. Luque & K. Mok 2011. Language- and talker-dependent variation in global features of native and non-native speech. *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 17–21 August 2011, 356–359.

Cucchiarini, C. H. Strik & L. Boves 2002. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America* 111. 2862–2873.

Davidsen-Nielsen, N. 1975. *English Phonetics*. Oslo: Gyldendal Norsk Forlag.

de Silva, V., A. Iivonen, L. V. Bondarko & L. C. W. Pols 2003. Common and language dependent phonetic differences between read and spontaneous speech in Russian, Finnish and Dutch. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 3–9 August 2003, vol. 3, 2977–2980.

Docherty, G. J. 1992. *The timing of voicing in British English obstruents*. Berlin/New York: Foris Publications.

Edwards, T. J. 1981. Multiple feature analysis of intervocalic English plosives. *Journal of the Acoustical Society of America* 69. 535–547.

Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. In W. Strange (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 233–277. Timonium: York Press.

Flege, J. E., M. J. Munro & I. R. A. Mackay 1995. Effects of age of second-language learning on the production of English consonants. *Speech Communication* 1. 1–26.

Guion, S. G., J. E. Flege, S. H. Liu & G. H. Yeni-Komshian 2000. Age of learning effects on the duration of sentences produced in a second language. *Applied Psycholinguistics* 21. 205–228.

Gut, U. 2009. *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Frankfurt am Main: Peter Lang.

Halvorsen, B. 1998. *Timing relations in Norwegian stops.* Ph.D. dissertation, University of Bergen.

Johnson, K. 2004. Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (eds.), *Spontaneous Speech: Data and Analysis*, 29–54. Tokyo, Japan: The National International Institute for Japanese Language.

Koopmans-van Beinum, F. J. 1980. *Vowel contrast reduction: an acoustic and perceptual study of Dutch vowels in various speech conditions.* Amsterdam: Academische Pers B.V.

Kristoffersen, G. 2000. *The phonology of Norwegian.* Oxford: Oxford University Press.

Kuiper, K. & W. Scott Allan 2010. *An introduction to English language: Word, sound and sentence.* Basingstoke: Palgrave Macmillan, 3rd edition.

Laan, G. P. M. 1997. The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication* 22. 43–65.

Ladefoged, P. & I. Maddieso 1996. *The sounds of the world's languages.* Oxford, UK and Cambridge, MA: Blackwell Publishers.

Lahey, M. & M. Ernestus 2014. Pronunciation variation in infant-directed speech: Phonetic reduction of two highly frequent words. *Language Learning and Development* 10. 308–327.

Mackay, I. R. A. & J. E. Flege 2004. Effects of the age of second language learning on the duration of first and second language sentences: The role of suppression. *Applied Psycholinguistics* 25. 373–396.

Nakamura, M., K. Iwano & S. Furui 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech and Language* 22. 171–184.

Polka, L. & O.-S. Bohn 2011. Natural Referent Vowel (NRV) framework: An emerging view of early phonetic development. *Journal of Phonetics 39*. 467–478.

R Core Team 2012. R: A language and environment for statistical computing. Retrieved from https://ww.R-project.org/.

Ringen, C. & W. A. van Dommelen 2013. Quantity and laryngeal contrasts in Norwegian. *Journal of Phonetics* 41. 479–490.

Schuppler, B., W. A. van Domme len, J. Koreman & M. Ernestus 2012. How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. *Journal of Phonetics* 40. 595–607.

Spilková, H. 2014. *Phonetic reduction in spontaneous speech: An investigation of native and non-native production*. PhD Thesis, Trondheim.

van Son, R. J. J. H. & L. C. W. Pols 1999. Acoustic description of consonant reduction. *Speech Communication* 28. 125–140.

Toivola, M., M. Lennes, J. Korvala & E. Aho 2010. A longitudinal study of speech rate and pauses in non-native Finnish. *Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech, New Sounds 2010*, Poznań, Poland, 1–3 May 2010, 499–504.

Trofimovich, P. & W. Baker 2006. Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition* 28. 1–30.

van Dommelen, W. A. 2007. Temporal patterns in Norwegian as L2. In Ulrike Gut & Jürgen Trouvain (eds.), *Non-Native Prosody: Phonetic Description and Teaching Practice*, 121–143. Berlin/New York: Mouton de Gruyter.

Warner, N. & B. Tucker 2011. Phonetic variability of stops and flaps in spontaneous and careful speech. *Journal of the Acoustical Society of America* 130. 1606–1617.

Martine Adda-Decker and Lori Lamel

4 Discovering speech reductions across speaking styles and languages

Abstract: In spontaneous and casual speech, word productions may exhibit strong deviations from their canonical or citation form pronunciations, with a large proportion of speech reductions. Reductions are observed even in prepared speech, such as journalistic speech by professionals, although their proportions remain low and they tend to occur in highly predictable word contexts. Speech reductions may be due to underarticulation or due to shortened pronunciations. Speech reduction results in either different segments (centralized vowels, lenited consonants), fewer segments, or even fewer syllables (Adda-Decker et al. 2005; Duez 2003; Ernestus 2000; Johnson 2004; Van Son & Pols 2003). Reductions seem to mostly affect the least informative speech portions (Jurafsky et al. 2001), i.e., function words, morphological items, discourse markers, and repeated information.

There has been a growing use of automatic speech recognition tools as an aid to carry out empirical studies on very large speech corpora (Nguyen & Adda-Decker 2013; VLSP workshop, Philadelphia 2011) as they facilitate investigation of pronunciation variants. Forced alignments of speech with a canonical (citation form) pronunciation dictionary can reveal temporally reduced speech regions (Adda-Decker and Lamel 1999; Adda-Decker and Snoeren 2011). If some segments are merged or simply not uttered (e.g., the unstressed syllable of "student" in "student athletes" pronounced as "stu'n athletes"), the forced alignment will produce very short segments in the corresponding region (the unstressed syllable "-dent").

In this chapter, we propose some large-scale investigations of speech reduction phenomena in journalistic and conversational, casual speech in two languages, French and English. We address reduction from the perspective of temporal speech reduction. We are looking for minimal duration segments (as automatically aligned by the system) to hypothesize which speech portions are most prone to reduction. In particular, we will look for phone sequences that tend to undergo temporal reduction in various speaking styles. We address the question of within-word reduction as opposed to word juncture reduction. We will sum up our descriptive work by proposing language-specific and more universal trends in temporal speech reduction. We will present representative examples and figures of typically reduced items as a function of the different speech genres in

Martine Adda-Decker, LPP, CNRS, Université Paris 3 Sorbonne Nouvelle
Lori Lamel, LIMSI, CNRS, Université Paris Saclay

https://doi.org/10.1515/9783110524178-004

French and English. The proposed method and results may contribute to gaining deeper insight into the characteristics of speech reduction and to increasing our understanding of the mechanisms underlying pronunciation variation.

Keywords: English, French, speaking style, temporal reduction, forced alignment

4.1 Introduction

In this chapter, we propose a dual investigation of speech reduction, embracing both technological and linguistic aspects. This double-sided approach aims at combining our experience in automatic speech recognition (ASR) with efforts to relate the observed variation to linguistic structure and processes. The study of speech reduction is attracting increasing interest in linguistic and psycholinguistic research as witnessed by the Nijmegen 2008 workshop on this topic (Ernestus and Warner 2011). Speech reduction is also acknowledged to be a major challenge in automatic speech processing. The last decades have witnessed major progress in ASR, largely due to the widespread use of statistical models combined with the availability of very large speech and language corpora. ASR systems require a large volume of spoken and written data to estimate models of spoken language, with acoustic models that model the elementary speech sounds and so called language models for ordering speech sounds into meaningful word sequences. These models capture the average properties of phoneme realizations and include statistics about word and pronunciation frequencies. Speech recognition research has progressively addressed more challenging speech data, moving from well-prepared speech to spontaneous conversations. Studies of ASR transcription performance reveal important differences across speaking styles, attributable to lexical choices, wording, and phrasing, as well as to the acoustic realization of a given word.

Since the eighties of the last millennium, the ASR research community has progressively addressed the difficulties of processing more varied and less controlled speech. A major bottleneck was the lack of written material truly reflecting spontaneous speech. Another observation was that part of the difficulty was related to pronunciations: acoustic models estimated using read speech data performed poorly on spontaneous speech – the same word, when read aloud and when occurring in natural discourse is not uttered quite the same. With respect to the former, important initiatives were launched to manually transcribe various sources of natural, more or less spontaneous speech, and hundreds of hours of transcribed speech have been made for improved ASR modeling. These data are also very interesting for large-scale studies to get a better view of acoustic

realization differences between speaking styles, and since such large corpora are only available for a few languages, various efforts are underway to apply the findings across languages. In this chapter, we develop some observations highlighting commonalities and differences as a function of language and speaking style.

In the following, speech reduction is approached as a temporal reduction or duration shortening. Measured shortening may reflect a variety of processes, such as vowel reductions, consonant cluster reductions, assimilation and lenition processes, segmental and syllabic deletions and restructuring. Many speech scientists share the belief that much knowledge can be gained from studying characteristics of casual speech (Greenberg and Chang 2000; Greenberg et al. 2003; Nakamura, Furui, and Iwano 2006; Strik et al. 2006; Schuppler et al. 2014). For instance, Greenberg et al. (2000, 2002) investigated syllabic structures in casual speech from the Switchboard data. Nakamura et al. (2006) compared spectral properties of careful and casual speech on large Japanese corpora, thereby highlighting spectral reduction. Strik et al. (2006) studied reduction phenomena in Dutch, with a focus on the problem of disappearing sounds, especially in multiword expressions.

Speech reductions seem to first affect the least informative speech portions (Jurafsky et al. 2001), for example, function words that are predictable from the context, idioms, morphological items (in particular endings), and discourse markers. Speech reduction can be manifest itself in various ways, such as producing different (e.g., centralized) phonemes, fewer phonemes, or even fewer syllables (Adda-Decker et al. 2005; Duez 2003; Ernestus 2000; Van Son and Pols 2003).

As far as phonemic segmentation and labeling is concerned, it is far from obvious that an automatic speech recognizer will prefer the same options as a human expert. A human listener cannot always tell for sure whether a phoneme is deleted since some of the missing phoneme's acoustic features may be present in adjacent phonemes, and may even be perceived. Moreover, it is well known that human speech perception may sometimes be biased by higher level language knowledge and understanding (see, e.g., Elman and McClelland 1988; Ganong 1980; Samuel and Pitt 2003). By contrast, an ASR system, for a given parameterization, will consistently make the same decisions over the entire corpus.

In this contribution, we investigate temporal reduction via a cross-lingual study in English and French. The basic idea is using a forced speech alignment tool (Adda-Decker and Lamel 1999) based on the LIMSI speech recognition system (Gauvain et al. 1994) to identify speech regions that are prone to reduction. We extend the methods proposed in Adda-Decker and Snoeren (2011) to provide evidence of temporal speech reduction with the help of automatic speech alignment using full-form pronunciation dictionaries and global descriptors, such as distributions of phone segment durations. Increasing proportions of short segments are often indicative of a higher degree of temporal reduction. The forced alignments

are used to quantify speech reduction in large corpora of various speaking styles, ranging from broadcast news to telephone and face-to-face conversations. By studying large speech corpora, the extent of speech reduction can be quantified as a function of various factors, such as speaking style and language, broad phonemic category and syllable position, or social variables including gender, age, or status.

The study of such speech regions may lead to deeper insight into the complexity of reduction phenomena specific to spontaneous speech, increase our understanding of the general mechanisms underlying pronunciation variation, and, last but not least, contribute to better acoustic speech models for ASR in the future. Before presenting our corpus-based study, we provide a brief introduction to speech reduction, illustrated by a few examples. This is followed by a short overview of speech modeling of reduction in ASR systems. The remainder of this chapter presents results and discusses the implications of the outcomes of the corpus-based study.

4.2 Temporal speech reduction

A considerable amount of research has been devoted to the study of speech reduction phenomena, including consonant lenition, consonant cluster simplification, vowel reduction, and syllable restructuring (see, e.g. Dilley and Pitt 2007; Duez 2003; Ernestus 2000; Tseng 2005; Van Son and Pols 2003, Gahl 2008, Torreira and Ernestus 2012, Whalen 1991). Frequent "phonological words" reflecting temporal structure reduction can be found in written English (e.g., *isn't*, *it's*, *gonna* in informal writing) and in written German (*ins* instead of *in das*, 'in the'). In French, similar reduction phenomena occur (*ça* instead of *cela*, 'it'). In this chapter we are interested in temporal reduction phenomena, in particular those that are not reflected in written language.

Reduced pronunciations are often observed in common word sequences which usually are easily predictable from the context. Some examples in French are: *il y a* [ilija] 'there is' which is most often uttered as *y a* [ja], and *je ne sais pas*, [ʒənəsɛpa] 'I don't know' which may have an acoustic realization close to [ʃɛpa] or even [ʃpa], where the *ne* in the negative form *ne ... pas* is completely omitted and /ʒə/ and /s/ are merged to form a single fricative segment that is [ʃ]-like. The /ɛ/-vowel may also become devoiced and merge with the preceding fricative segment.

Similar examples can be cited for English, where some reductions have even been widely adopted in written language. The word sequence *I do not know* is generally written as *I don't know*, and is often further reduced in speech to simply

I'd know or *dunno*. From an ASR perspective, this problem has been addressed by adding reduced forms such as *wanna, dunno,* and *gonna* as lexical items or including "multiwords" (sequences of words that tend to frequently co-occur) in the pronunciation dictionary as a single entry (see Strik and Cucchiarini 1999; Strik, Binnenpoorte, and Cucchiarini 2005). Multiwords will have multiple pronunciations ranging from a concatenation of canonical forms to strong reductions. In English, *want to* can match a range of pronunciation variants from [wantu] to [wʌnə].

Figure 4.1 shows examples of typical reductions as observed in spontaneous speech. The left-hand side example in Figure 4.1 is taken from a casual conversational and the right-hand one from a broadcast interview with politicians. These examples illustrate that the scope of sequential reductions often surpasses word boundaries, typically involving one or more short function words.

Strong temporal reductions may also be observed in content words. This type of reduction often occurs with words that are highly predictable in a given context. For example, in news reports, polysyllabic words such as *(prime) minister, president* may be uttered very quickly with only parts of the underlying form recognizably uttered in the surface form, especially when they are followed by the person's name. Figure 4.2 illustrates strong temporal reductions for content words in prepared speech taken from a French broadcast news recording, where the reduction is particularly strong in the excerpt shown in the right part.

Figure 4.3 shows an English example with two different realizations of the word sequence *President Zardari*, occurring twice in the same conversation. While all of the phones of the word president are clearly articulated the first time and also

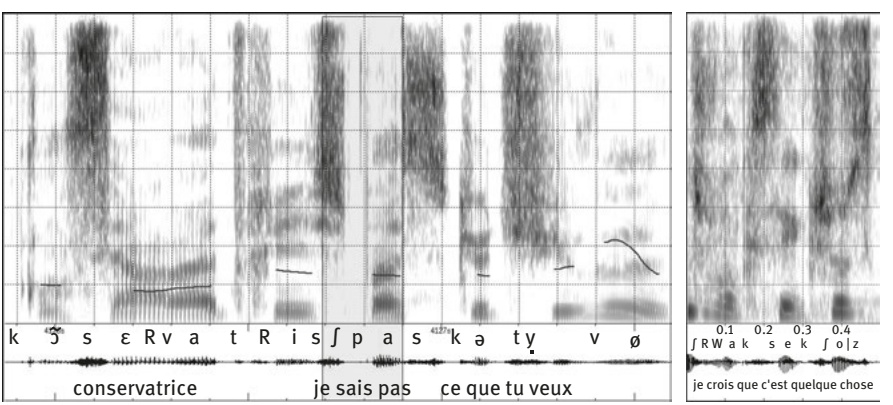

Figure 4.1: Speech signals of common reduction phenomena in French. Left: *je sais pas* 'I don't know' /ʒəsɛpa/ in the context *conservatrice, je sais pas, ce que tu veux* 'conservative, dunno, what you want' is approximately produced as [ʃpɑ]. Right: *je crois que c'est quelque chose* 'I believe it is something' /ʒəkʀwɑkəsɛkɛlkəʃoz/ is approximately produced as [ʃʀwɑksekʃoz].

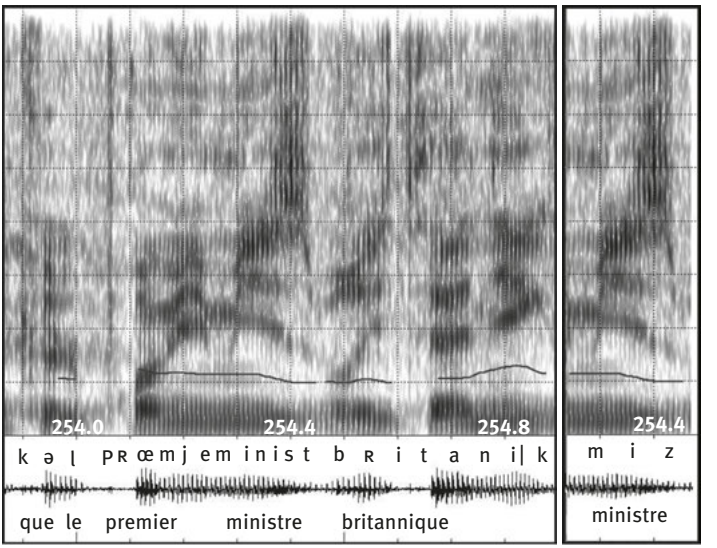

Figure 4.2: Speech signal illustrating French content word reduction in *ministre*. Left: word in context: *que le premier ministre britannique* 'that the British prime minister' aligned as [mInIst], the system's shortest variant. Right: focus on the word *ministre* /ministʁə/ approximately produced as [miz].

clearly visible in the left spectrogram of Figure 4.3, the later production shown in the right panel is severely reduced, in particular the two word-final unstressed syllables of *president* (produced as [pʁɛzn]).

It is noteworthy that also the president's name is shortened with deletion of at least the two consonants /ʁ/ and /d/ and most probably also the preceding vowel /ɑ/ as the nucleus of an unstressed syllable.

In the following, the word "stress" is used to refer to accented parts in speech, be they due to lexical stress (as in English) or to phrase-final accentuation (as typical for French which has no lexical stress) or due to the utterance's focus structure. We thus use the word "stress" with its general English definition corresponding to prominent regions in both languages. However, the interested reader should keep in mind that the most appropriate prosodic terminology for French would be "accent" (Jun and Fougeron 2002). Whenever necessary, and when speaking more specifically about French, we will use the term "accent" instead of "stress." Our hypothesis is that although any stretch of speech may be shortened and altered, reduction is considered to affect most often the unstressed segments, whereas the prominent or stressed parts tend to remain more clearly articulated. Prominent or stressed regions may be considered as major anchor points attracting perceptual focus. These stressed regions enable or at least ease

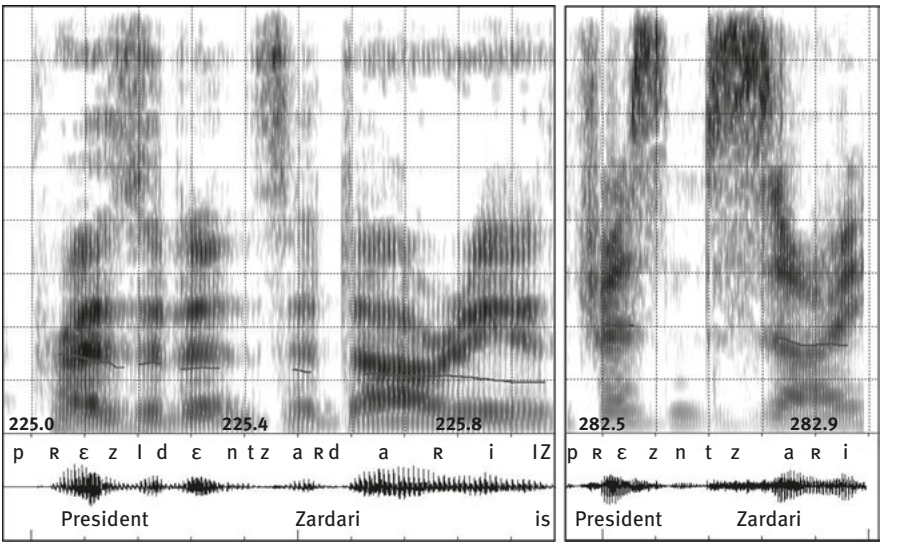

Figure 4.3: Speech signal illustrating content word reduction in *president* in English. Left: *President Zardari* clearly articulated. Right: *President Zardari* strongly reduced, approximately uttered as [pʀɛzntzɑʀi].

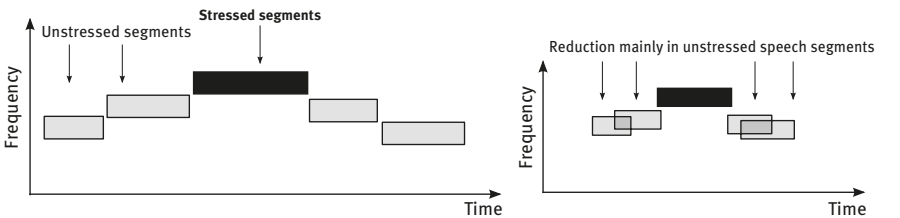

Figure 4.4: Schematic representation of spontaneous speech with prominent or stressed (black boxes) and unstressed (gray boxes) parts. Left: clearly uttered speech with both prominent and unstressed parts realized. Right: temporally reduced speech mainly affecting unstressed parts.

the restoration of the reduced, unstressed portions. Figure 4.4 gives a schematic view of our understanding of temporal speech reduction.

This view or "model" is compatible with temporal reductions as shown in spectrograms in Figures 4.1 and 4.2. In the first example, *je suis d'accord* 'I agree,' reductions mainly involve unstressed parts *je suis*. Corpus-based investigations can contribute to the validation of this hypothesis.

Before turning to our corpus-based study and the related methodology, we first give a brief overview of speech modeling in ASR and discuss the effects of speech reduction on ASR performance if it is not appropriately dealt with in a system's acoustic speech and pronunciation models.

4.3 Automatic speech processing as tools for linguistic studies

In this section, we present some basic processing and modeling steps in ASR systems. In particular, we focus on temporal modeling aspects in order to demonstrate how temporally reduced speech may be detected by the system. In our opinion, it is essential to grasp these basic steps to understand the potential contributions and limits of forced alignment and correspondingly labeled data. We also briefly address the issues of segmentation accuracy and of segmentation labels as compared to those produced by human experts.

4.3.1 Speech modeling

A first processing step corresponds to the conversion of the acoustic signal to a sequence of acoustic parameter vectors used by the alignment system. Figure 4.5 (left) illustrates this conversion from the speech signal (bottom) to parameter vectors (top): a time window with a length of several pitch periods is required to compute meaningful spectral coefficients in voiced speech. The window size is of fixed length of typically 30 ms, and it is shifted by a fixed step of usually 10 ms to produce a steady flow of acoustic parameter vectors. This window duration guarantees the inclusion of at least two pitch cycles in a deep male voice. The frame-based processing implies that automatically determined segment boundaries are no longer placed on a continuous time axis, but on a discrete grid with a regular (10 ms) spacing. Furthermore, fine details in the speech signal or the corresponding spectrograms that may be essential cues for human experts are not available for boundary location. The segment boundaries of a given word are placed to globally optimize the location of the predicted segments (via the pronunciation dictionary) with respect to the observed signal. Although not used here, shorter steps of 5 ms have also been experimented within the literature (Bartkova and Jouvet 2015) especially to address variant selection of temporally reduced speech variants.

Hidden Markov models (HMMs) (Rabiner and Juang 1986) are widely used to model the sequences of acoustic feature vectors, with acoustic units corresponding to phones as shown on the right side of Figure 4.5. Although Figure 4.5 shows only one single model, a given phoneme is typically modeled by a large set of context-dependent phone (allophone) models, as context strongly influences acoustic realizations. Figure 4.6 illustrates the speech modeling and alignment process. Each acoustic vector becomes part of a single phone segment

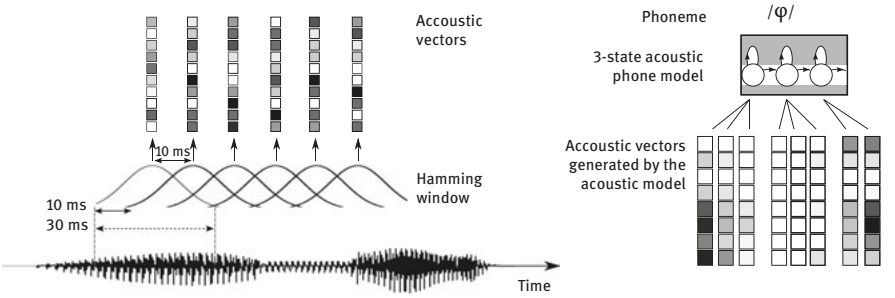

Figure 4.5: Left: Speech parameterization: audio signal is converted to acoustic vectors with a 10-ms frame rate. Right: outline of acoustic phone modeling: 3-state HMM phone model linking the abstract phoneme level to an acoustic realization.

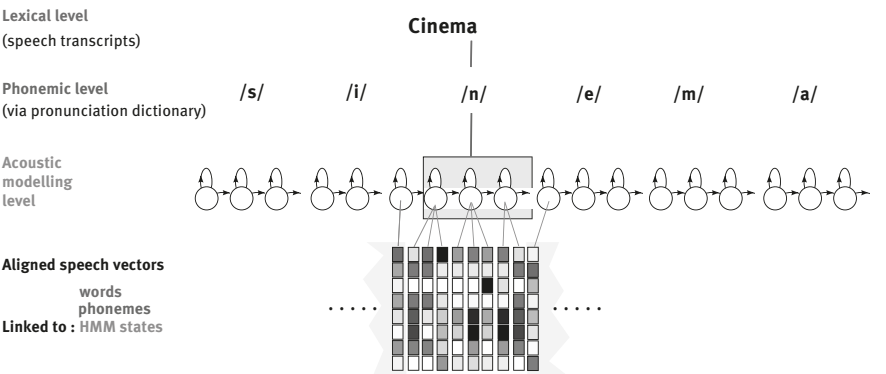

Figure 4.6: Multilevel speech modeling for automatic speech alignment: the lexical level with the written words links to a phonemic level with canonical pronunciations. Each phonemic symbol is associated with allophonic (context-dependent) acoustic model to account for contextual variation. The central state of each 3-state HMM corresponds to the center of the phone segment, the outer states to the phone's start and ending, with potential transitional frames to the neighboring segments.

(or a silence segment at word or phrase boundaries). Segment boundaries are typically located in transitional parts. Various studies have reported boundary location accuracy under 20 ms (Di Canio et al. 2012). Our experience with boundary location is in line with this result. The output of forced speech alignment is a sequence of contiguous phone segments with labels predicted by the pronunciation dictionary.

The quality of pronunciations included in the alignment system is of crucial importance in the production of automatically aligned speech data. However, this leads to a host of questions about how to determine the pronunciation(s) that will be useful for further linguistic investigations? Should they reflect surface forms or underlying forms? Should they reflect phonetic or phonemic labels? A canonical pronunciation dictionary typically includes the full-form pronunciations, which supposes that all possible segments are pronounced. For many languages, there is a strong correspondence between orthographic and spoken forms. Canonical pronunciations thus tend to produce phonemic labels of the underlying form. By introducing pronunciation variants in the dictionary, the alignment system can choose among different options to produce labels that are closest to the actually produced sounds. In this case, the automatic labeling may become closer to what is considered as a broad phonetic labeling as it tends to adjust to the observed production. However, even if pronunciation variants are added, the labels and segmentation options remain constrained to the actual options foreseen by the alignment system. How to ensure that major variants are included in the pronunciation dictionary? Generally speaking, the automatic system needs to learn the pronunciation variants, which consists of providing it with audio samples with corresponding transcripts, and cannot reliably make very fine level distinctions.

4.3.2 ASR errors and temporal structure

In this section, we briefly address ASR errors as these are one of the means of revealing mistakes and missing variants in the pronunciation dictionary. Previous studies for the French language reported word error rates of about 10% for careful (i.e., journalistic) speech and above 15% for casual telephone speech, using large corpora for system training (hundreds of hours of appropriate casual speech data) and complex system combinations (see Lefèvre, Gauvain, and Lamel 2005; Prasad et al. 2005; Gauvain et al. 2005). Approximately 30–40% of the errors in automatic transcriptions of careful speech consist of homophone or near-homophone errors without temporal reduction. Several studies have focused on close homophone substitutions in terms of ASR errors and human perception (Vasilescu et al. 2009, 2011).

Table 4.1 gives some examples of near-homophone errors with temporal reduction in prepared journalistic speech. The pronunciations of the hypothesized word sequences are shorter than those of the reference transcription. When analyzing casual, conversational speech, however, the proportion of errors due to temporal reduction increases significantly. Temporally reduced speech, corresponding to sequences of short words, such as discourse markers (*tu sais, tu*

Table 4.1: Near-homophone errors with temporal reduction: the system's *Hypothesis* pronunciation is shorter than the pronunciation of the *Reference* transcription. The comment column suggests phonological processes underpinning the observed variation.

Reference	Pronunciation	Hypothesis	Pronunciation	Comment
ça avait	/saavɛ/	savaient	/savɛ/	V#V merger
semble que	/sɑ̃bləkə/	somme que	/sɔmkə/	word-final CC deletion
parce que	/paʁsəkə/	ce que	/səkə/	atone syll. deletion
près de Paris	/pʁɛdəpaʁi/	préparé	/pʁepaʁe/	clitic deletion

vois 'you know, you see') and markers of reported speech (*il m'a dit, je lui ai dit* 'he told me, I told him'), is particularly frequent in these kinds of data. Consequently, these sequences are often prone to recognition errors, unless specific shortened pronunciations are included in the pronunciation dictionary used for both acoustic model training and for decoding.

During forced alignment, full pronunciation models tend to be a poor match with temporally reduced speech. In such regions, the segmentation is characterized by several contiguous small segments of minimal duration, which can be automatically detected by looking for minimal duration phone segments in the forced alignments. These regions of short segments tend to reflect a mismatch of the system's speech model when a short surface form needs to be aligned with a longer underlying form. This situation may result in ASR transcription errors as can be illustrated by the following example: *quai de Seine* 'bank of the Seine' /kɛdəsɛn/ was uttered in two syllables (without the schwa vowel) and misrecognized as *quête saine* 'health quest' /kɛtsɛn/. The two sequences are almost homophonic, where the differences can be explained by a combination of French phonological processes such as schwa elision and regressive voice assimilation. French compound nouns are typically built as <noun>-*de*-<noun> sequences. In such constructs, the schwa of *de* 'of' is typically deleted before a consonant (here the /s/ of *Seine*) when preceded by an open syllable (here the /kɛ/ of *quai*). The /d/ may then become a devoiced [t] due to the following unvoiced /s/ (cf. Snoeren, Hallé, and Segui 2006).

In casual speech, it is common to find complex combinations of various reduction processes. Using large corpora provides the opportunity to elaborate a synthetic overview of the various reduction processes (cf. Schuppler et al. 2008). As a first step in this direction, we propose to quantify temporal reductions using forced alignment. This allows us to measure deviations from canonical temporal structures in terms of their phone segment duration distributions as well as in terms of minimum duration sequences. This approach is further explained in the methodology section (see Section 4.5).

4.4 Speech corpora

Several large speech corpora were used in these studies, containing different styles of data in French and English: public broadcast speech including both news (BN-news) and less formal conversations (BN-conv)., conversational telephone speech (CTS), and face-to-face conversations. The careful speech data set stems from French BN and corresponds to 360 hours of various radio and TV shows that were used for the *Technolangue* ESTER (Galliano et al. 2005) campaign distributed by ELDA (European Language Data Agency, http://www.elda.fr). Similar data for English are widely available, in particular the BN data produced for the DARPA *Rich Transcription 2004 Broadcast News* evaluation (Nguyen et al. 2004), distributed by LDC (Linguistic Data Consortium, http://www.ldc.upenn.edu/Catalog/). Some of the broadcast data were classified as broadcast conversations (BN-conv). These data are more spontaneous (less prepared) than BN-news, and are often interviews or debates.

The casual speech data set is comprised of about 120 hours of LIMSI internal French telephone conversations. These conversations are mostly between friends and/or family members, so the corpus therefore contains a highly casual speaking style. The casual speech data set for English comes from the Switchboard (Godfrey, Holliman, and McDaniel 1992) and Fisher data (distributed by LDC) including thousands of hours of speech. In these corpora, the telephone callers do not know each other and are supposed to speak about assigned topics. Therefore, the speech, although spontaneous, is less casual here than the speech in the French corpus. Each corpus includes hundreds of male and female speakers.

Furthermore, we studied a French corpus of face-to-face conversations between friends, the NCCFr – Nijmegen corpus of casual French – (Torreira, Adda-Decker, and Ernestus 2010), available at Nijmegen for research purposes. Table 4.2 provides a summary of the studied material.

Table 4.2: Corpora used in this study: BN careful, prepared (news) and conversational (conv) speech, and casual telephone (tel) and face-to-face (f2f) speech. French (left panel) and English (right panel).

	# word tokens	duration		# word tokens	duration
BN-news	3,600 k	360 h	*BN-news*	7,200 k	720 h
BN-conv	600 k	44 h	*BN-conv*	1,500 k	124 h
Casual-tel	1,000 k	100 h	*Casual-Tel*	25,000 k	2,300 h
Casual-f2f	350 k	31 h			

4.5 Methodology

Figure 4.7 illustrates how an ASR system can be used as an instrument for linguistic studies. The system can be used to align a word-level transcription with the speech signal, given the pronunciations for each word. The provided pronunciations can allow the investigation of linguistic phenomena. Some previous investigations showed that major linguistic trends (e.g., vowel reduction, French liaison, voice assimilation, and regional accent specificities) could be validated using automatically aligned speech data (Adda-Decker, Gendrot, and Nguyen 2008; Woehrling 2009). Previous work also compared segmenting various data types with full-form canonical pronunciations as well as variants designed to detect vowel and consonant changes or deletions due to reductions (Adda-Decker and Lamel 1999). Building upon work done by Adda-Decker and Lamel (2005) and Adda-Decker and Snoeren (2011), in this work the methodology is applied to highlight temporal reduction tendencies based on measures of phone durations (distributions, durations by phone classes, or phone sequences).

The basic idea exploited is that when temporally reduced speech is aligned against full-form pronunciations, there will generally be several contiguous phone segments of minimal duration (i.e., 30 ms here). An example of reduced speech together with an illustration of its automatic alignment using a full-form pronunciation model is shown in Figure 4.8. Many of the aligned segments are of minimal duration. It is worth noting that in this case most segments are neither correctly located nor correctly labeled. However, a sequence of minimal duration segments highlights temporal reduction, which is the point of interest here, and is investigated by tracking such sequences of minimal duration in our corpora.

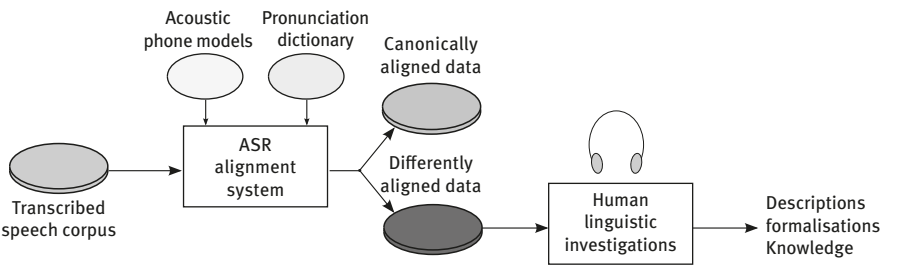

Figure 4.7: The automatic speech recognizer as an instrument to automatically select canonically and differently aligned subsets of speech deviating from expected representation. These subsets are of interest for more in-depth linguistic investigations.

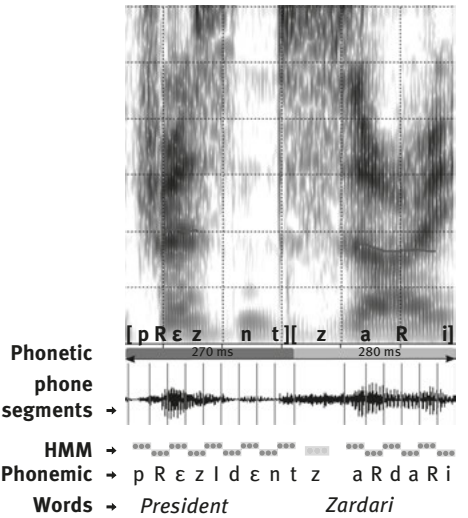

Figure 4.8: Minimal duration segment sequence in temporally reduced English words President Zardari: /pʀɛzɪdɛnt zɑʀdɑʀi/, approximately produced as [pʀɛzntzɑʀi] /. As a result, the automatically aligned segments (except segment [z]) are of minimal duration (1 acoustic vector per state which results in 30-ms segments in our 3-state HMMs).

4.6 Results

A first investigation examines segmental durations produced by automatic speech alignments using full-form pronunciation dictionaries in order to localize temporal reduction in fluent speech. After comparing global segment duration distributions across speaking styles and languages, we will focus on duration variation fixing either the phone identity or the word identity.

The second line of investigation aims at qualifying and quantifying the observed temporal reduction phenomena beyond the segment level. To this end, we introduced shorter pronunciation variants into the dictionary. During forced alignment, the best matching variant was chosen. Our hypothesis is that this chosen variant not only provides information about the presence of temporal reduction but also uncovers possible clues about the reduction processes involved.

In the context of automatic speech processing, known temporal reduction phenomena may be accounted for in the pronunciation dictionary by adding pronunciation variants which are shorter than the canonical form (Lamel and Adda 1996; Lamel and Gauvain 2005; Karanasou and Lamel 2011).[1] Assessing their usage during alignment can give an indication of the importance (frequency) of the phenomenon. However, our belief is that some of the temporal reduction

1 Note that the aligned pronunciation of government /gʌvəmənt/ in Figure 4.10 is already reduced.

phenomena still escape our inventory of explicit knowledge as they tend to be unnoticed by native speakers of the language. Several phone segments in a row that are of minimal duration (30 ms) with respect to the forced alignment procedure tend to reveal such temporally reduced regions in the speech data. Reduced sequences also tend to cause trouble to foreign language speakers who struggle to follow given the blatant mismatch between their learnt full-form pronunciations and the reduced ones produced by native speakers. This also raises interesting cognitive processing questions that are beyond the scope of this chapter. What is perceived by listeners stems partly from the acoustic input and partly from the representations in their brain, reflecting among other things their past language experience and their current contextual situation.

4.6.1 Segmental duration

In the following, we provide a bird's-eye view of segmental duration variation, before detailing some illustrative examples at segmental and lexical levels.

4.6.1.1 Segment duration distributions

To provide a synthetic view of segment durations, Figure 4.9 shows the phone segment duration distributions of aligned data in French and English for both prepared BN and spontaneous telephone speech. The speech alignments relied on full-form pronunciations with only a small number of exceptions with shorter variants to account for the well-known reduction phenomena (e.g., in English *hundred*: /hʌndrəd, /hʌnrəd/, /hʌnəd/; in French *autre* 'other': /otrə, /otr/, /ot/). The alignments tend to find the best match between the proposed acoustic word models and the speakers' productions. As highlighted in Figure 4.8, strongly reduced productions will result in sequences of minimal duration (30 ms) segments. High rates of short durations are thus indicative of temporal reduction.

The top part of Figure 4.9 provides a histogram of proportions of segments in French as a function of segment duration, with corresponding histograms for English on the bottom. To save space on the abscissa, durations are given in centiseconds (cs) in the figures and not in milliseconds (ms) as in the text. The results are broken down into prepared (left) and conversational (right) speech styles. Concerning prepared speech, the largest number of segments (>14% in French, 13% in English) have a duration of 60 ms in French and 50 ms in English. With respect to the spontaneous telephone speech, the French distribution has by far the most segments (>18%) in the shortest duration bin of 30 ms, with almost one-third of the segments having a duration of 30 or 40 ms. As for

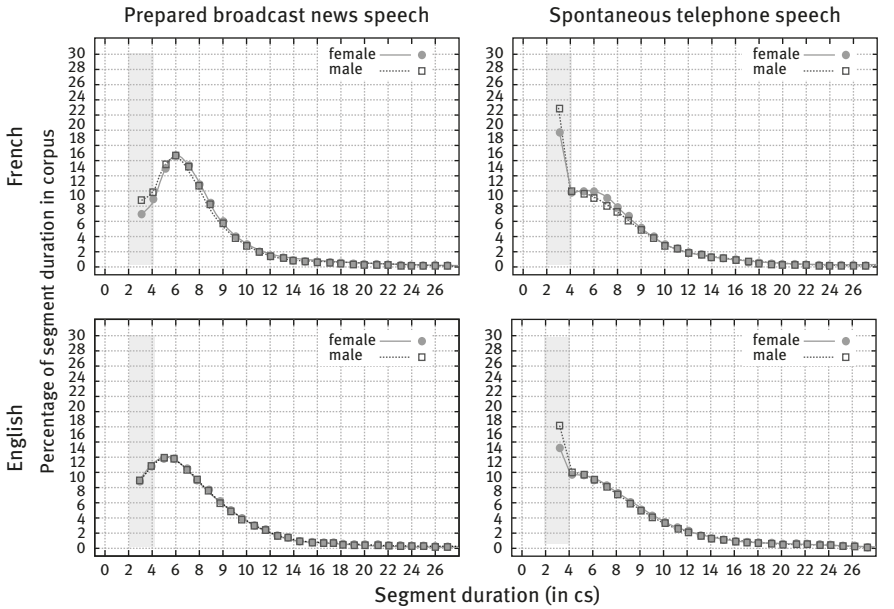

Figure 4.9: Phone segment duration distribution (all phone segments pooled). Comparison between broadcast news (left panels) and spontaneous telephone speech (right panels). The highlighted region (3–4 cs) corresponds to potentially reduced segments.

French, English spontaneous speech also has the highest proportion of segments (15%) in the minimal duration bin and 25% of the segments have a duration of up to 40 ms. The same trends are observed for male and female speakers. The high proportion of short-duration segments (highlighted in pink in Figure 4.9) in spontaneous speech suggests that temporal reduction is an important issue to address in order to improve our knowledge of native pronunciations and related phonological processes in spontaneous speech and to increase the acoustic modeling accuracy in ASR. Even though the proportion of minimal duration segments is much lower in prepared speech, there still are 8% of all segments with 30 ms duration and 18% with 30 or 40 ms duration in both languages. If the minimal duration segments highlight temporal reduction, the distributions of Figure 4.9 reveal their strong presence in spontaneous speech. They also reveal that carefully prepared speech is also concerned, although to a lesser extent.

4.6.1.2 Position-dependent analysis of the English plosives /t/ and /k/

While the duration histograms indicate overall trends, hereafter we examine the realizations of some English consonants in more detail. How do segment durations

vary with their position in a word or a phrase? or as a function of lexical stress? Table 4.3 reports figures for some typical English words highlighting variation in duration of /t/ and /k/ consonants. Words are chosen so as to illustrate the influence on duration of word-initial and medial positions and of changing lexical stress. For each word type, the number of tokens in the corpus and the percentage of minimal (up to 40 ms) duration segments are shown. The chosen words occur rather frequently in both English corpora (BN and CTS). It can be seen that the duration of a phoneme's realization depends on its position in the word. For example, in lexical stress-bearing syllables (marked (*) in Table 4.3; [t] in *talking, trying, nineteen, hotel*) or /k/ in *coming, because*), average durations of syllable-initial consonants are all higher than 80 ms, and rates of minimum duration segments remain low. In contrast, these rates tend to increase for consonants in syllable coda positions and in atone syllables in general. Similarly, average segment durations tend to decrease, more strongly for /t/ than for /k/. It can be observed that /t/ is more likely to undergo temporal reduction than /k/ in the shown examples. The highest minimum duration rates are observed for a [t] in

Table 4.3: Position-dependent analysis of /t/ and /k/ in some typical English words in conversational broadcast data (left) and in the telephone Switchboard and Fisher conversations (right). Average phone durations are given in ms together with standard deviations. (*) indicates that /C/ is syllable-initial in a lexical stress position.

	/C/ position	Broadcast			SWB/Fisher Conversations		
	/t/	#tkn	avrg. dur. stdev	% min. dur.	#tkn	avrg. dur. stdev	% min. dur.
talking	w-init (*)	814	95 42	5	4,898	80 34	11
trying	w-init (*)	684	95 43	6	4,464	85 38	11
nineteen	w-mid (*)	560	80 23	7	706	89 21	8
hotel	w-mid (*)	105	118 29	0	178	126 32	1
ninety	w-mid	323	70 27	20	821	43 26	22
getting	w-mid	803	59 33	52	5,692	39 21	86
little	w-mid	1,041	59 31	41	9,379	37 28	91
exactly	w-mid	387	54 29	43	6,328	39 29	85
	/k/						
coming	w-init (*)	825	108 52	4	2,301	96 35	2
conversation	w-init	110	92 28	3	610	88 27	6
doctor	w-mid	85	84 34	9	649	65 25	21
focus	w-mid	125	83 28	6	254	69 20	10
because	w-mid (*)	5,342	80 35	16	32,062	88 38	9
basically	w-mid	399	52 30	53	2,499	64 28	30

a consonantal environment [k_l] (in *exactly*). In general, it can be seen that segmental durations are lower for telephone conversation data than for broadcast data, exception made for /t/ in *hotel* and /k/ in the two last lines in Table 4.3. The latter need some additional comments: *because* and *basically* had proven to be often shortened in spontaneous English and thus had additional reduced pronunciations (/kɔz/ and /besIkli/) in the CTS pronunciation dictionary: 40% of the *because* tokens were thus aligned with the 3-phone pronunciation (deletion of atone syllable *be-*) and all *basically* tokens preferred the shortest variant. The option of shorter pronunciations in CTS data during forced alignment thus resulted in somewhat longer segment durations than in BN data where these shorter pronunciations were not provided in the dictionary. This study suggests that it is also interesting to consider including such reduced variants in the BN pronunciation dictionary.

4.6.1.3 Word-internal duration variation

Another way of examining position dependence in the phonetic realization of segments is illustrated in Figure 4.10, which depicts the temporal realization of sample polysyllabic words in English and French. Similar words were selected (*government*, *governments*, *gouvernement*, and *gouvernementale*) in both English and French. The words were extracted from the BN corpora, pooling occurrences in all phrasal positions and were frequent enough so as to consider the measurements as speaker independent. The hypothesis is that the average durations of unstressed segments are much shorter than those of stressed ones and are also shorter than the overall average segment duration. Some of them may even fall in the minimum duration zone (shown in red) for which adding shortened pronunciations to the pronunciation dictionary could be envisioned. It is interesting to see that the duration profiles are very similar for identical lemmas, even though the number of occurrences differs importantly. The longer words tend to have shorter segment durations in unstressed parts. It is also nice to observe that English words tend to have longer segments due to lexical stress in the word-initial part, and French words tend to have longer segments in the word-final part which often co-occurs with phrase-final position. Some of the final lengthening may also be due to pre-pausal position, which is not explicitly denoted in our data.

4.6.2 Alignments using reduced variants

After having observed that many speech stretches are aligned with minimal durations, a sensible solution then consists of anticipating shorter pronunciation

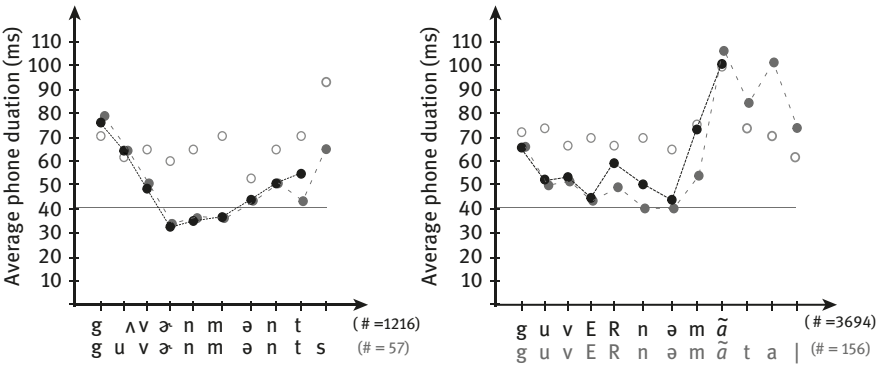

Figure 4.10: Average segment durations of polysyllabic content words as obtained by automatic alignment of BN data. Left: English *government* (in black), *governments* (in blue). Right: French: *gouvernement* (in black), *gouvernemental* (in blue). The abscissa shows the phone labels (and # of word occurrences) and the ordinate the average phone duration in ms. For comparison, the empty circles show the overall average phone durations.

variants in the dictionary. In this part, we thus move away from segmental duration investigations to study the usage of these shorter variants during forced alignment.

Temporally reduced stretches of speech may correspond to relatively known phenomena (e.g., *dunno* in English). The use of "multiwords" that merge potentially reduced word sequences into one single pronunciation dictionary entry enables the introduction of shorter pronunciations accounting for cross-word phenomena. Multiwords were introduced in ASR (Stolcke et al. 2000; Strik et al. 2005) to tackle this spontaneous or fluent speech-specific reduction problem. The rationale of "multiwords" is to limit the proposed reductions to these and only these word sequences preventing their broader usage in a general cross-word situation. Hence, they generally correspond to highly frequent word bigrams. Analysis of ASR errors in spontaneous speech reveals that temporally reduced stretches of speech may also occur in less frequent word bigrams (see Table 4.1: *semble que* 'seems that' recognized as *somme que* 'sum that' due to word-final CC cluster dropping in prosodic word-internal position).

4.6.2.1 English variants

The effectiveness of shorter pronunciation variants in multiwords was studied via forced alignment in a subset of the English CTS Switchboard corpus (185 hours of speech). A total of 250 multiwords were introduced for a vocabulary

of about 25k word types. For this experiment, common reduced written forms (e.g., *didn't, you've*) of the manual transcripts were matched and pooled with the corresponding full multiword form *did-not, you-have*. Table 4.4 shows some typical examples with the different pronunciations proposed in the dictionary: full form and shorter variants. To limit the number of variants to be displayed in the table, equal-length variants that differ only in vowel quality (e.g., [tu], [tə]) were merged. The variants are ordered by decreasing length, with the most frequently aligned one shown in boldface. For each multiword, the table indicates the frequency of each variant (i.e. the number of times it was used during alignment (#Align) and its corresponding percentage $\left(\frac{\#Align}{\#Total}\right)$ in the total number of tokens of the multiword (#Total). It is interesting to describe the different transformations when moving from the full-form pronunciation to the more reduced ones. It can be observed that the widely studied vowel reduction process (change of vowel quality of peripheral vowels toward a more central schwa) often accompanies pronunciation shortening, unless the vowel completely disappears from the variant.

Table 4.4: Examples of ASR multiwords with shortened pronunciation variants to deal with temporal reductions in spontaneous English Switchboard data. For each multiword type are given: the total number of tokens, the different pronunciation hypotheses (full form and variants) of the dictionary along with the number of tokens aligned with each one, and the corresponding ratio (#Align/#Total).

Multi-word	**#Total**	***Full form*** **+ Variants**	**#Align**	$\frac{\#Align}{\#Total}$	**Comments**
English spontaneous speech – Switchboard data					
did-not	2,559	dɪd nɑt	103	4.0	full form
		+ dɪdn̩t	275	10.7	n(ɑ → ə)
		+ dɪdn̩	1175	45.9	+ final-/t/ deletion
		+ dɪn̩	1006	39.3	+ coda /d/ deletion
we-have	3257	wihæv	1500	46.1	full form
		+ wiəv	205	6.3	onset /h/ del. + (æ → ə)
		+ wiv	1552	47.7	+ V-deletion
going-to-be	750	gɔɪŋtubi	73	9.7	full form
		+ **gɔnəbi**	432	57.6	complex: ɪŋt → n
		+ gəbi	245	32.7	+ complex: nɔə → ə
wants-to	157	wɔntstu	15	9.6	full form
		+ wɔnstu	78	49.7	coda C-cluster simplification
		+ wɔntsə	7	4.5	onset /t/-deletion
		+ wɔnsə	57	36.3	both /t/-deletions

Syllabic consonants reflect the merging of a schwa with following consonant (n,m,l,r). A metathesis can result in a /ɹ/-vowel sequence (such as in the word *hundred*) being realized as /ɚ/. The consonants /t,d/ are easily deleted especially in homorganic consonant clusters and in coda positions. An intervocalic /h/ appears to be elidible even in onset position.

4.6.2.2 French variants

For French, we have not yet investigated the use of multiwords for ASR. Unlike English, French standard writing does not tend to provide reduced written forms even though they may appear in oral productions (*je ne sais pas* 'I don't know' may be written as *je sais pas* in less formal writing, but never as *chais pas* even though this is a most common pronunciation in spontaneous French). For the NCCFr casual face-to-face speech, shortened pronunciations were introduced in the pronunciation dictionary to test whether they would be selected for temporally reduced words during speech alignment Table 4.5 shows examples of spontaneous speech reductions in single French words taken from the automatic alignments of the NCCFr corpus.

Different phonological processes are seen to be active in reduced pronunciations in French. Among these, schwa vowel deletion in final but also word-internal position is certainly the most pervasive one. As a result, the rhythmic pattern changes with a smaller number of more complex syllables. The French schwa is typically considered as optional: whether or not it is realized depends on the speaker, his/her regional origins, his/her speaking rate, the embedding context, the length of the prosodic word, etc. Consonant clusters in syllable coda positions are often simplified. In particular, liquids (/R/ and /l/) tend to disappear not only in word-final plosive-liquid clusters (*être → êt'* 'to be'; *montre → mont'* 'show'), but also in syllable coda position before another consonant (*parce que → pa'ce que* 'because'; *quelque → que'que* 'some'; *film → fi'm* 'movie'). In contrast, the schwa vowel, although often very short, is more systematically produced in English.

Table 4.5 also exemplifies /t/ deletions (in *main(te)nant* 'now'), however they tend to occur in homorganic consonant neighborhoods. Beyond schwa vowel deletion, we can observe that vowels in unaccented positions may disappear. For example, the frequent French word *peut-être* 'maybe' tends to be pronounced [ptɛt] in casual speech with a loss of the central rounded /ø/ vowel besides the simplification of the final consonant cluster. Another important process contributing to temporal reduction in spontaneous French (but not examined here) corresponds to vowel deletion (be it V1 or V2) in V#V contacts in cross-word

Table 4.5: Examples of words with shortened pronunciation variants introduced to handle temporal reductions in spontaneous French NCCFr data. For each entry the total number of tokens and the different pronunciation hypotheses (full form and variants) of the dictionary are given. The number of tokens aligned with each variant and the corresponding ratio (#Align/#Total) are specified. The most popular variant is shown in bold.

Word	#Total	Full form + Variants	#Align	#Align / #Total	Comments
French spontaneous speech – NCCFr data					
parce que	2590	pɑʀsə	4	0.2	full form
'because'		+ pɑʀs	45	1.7	no final schwa
		+ **pas**	1309	50.6	+ C-cluster simplification
		+ ps	1232	47.6	+ vowel deletion
peut-être	636	pøtɛtʀə	18	2.8	full form
'maybe'		+ pøtɛtʀ	28	6.0	no final schwa
		+ p(ø|ə)tɛt	109	17.1	final cluster simplification
		+ **ptɛt**	481	75.6	+ unaccented vowel deletion
maintenant	352	mɛ̃tənɑ̃	8	2.3	full form
'now'		+ mɛ̃tnɑ̃	114	32.4	no internal schwa
		+ mɛ̃nɑ̃	230	65.3	+ /t/-deletion
quelques	56	kɛlkə	14	25	full form
'some'		+ **kɛkə**	28	50	+ /l/-deletion
		+ kɛ(k|g)	14	25	+ schwa deletion

situations. A typical example here is the *t'as* 'you've' production instead of *tu as* 'you have.' ASR error analysis often pointed out such cases, be they located in highly frequent words such as *tu as* or in less frequent ones. V#V contacts are good candidates to undergo reduction with either vowel deletion or vowel merging. Temporal reduction may hence become more or less severe, depending on the cascade of phonological processes involved. We hope that the proposed descriptive work and methodology to spot temporal reductions in spontaneous speech will contribute to better disentangle the complexities of speech production and perception.

4.7 Discussion

In this chapter, we introduced the idea of using forced alignments to locate temporally reduced sequences in fluent speech. When used with full-form

pronunciation dictionaries, sequences of minimal duration segments reveal potentially reduced productions. Although the exact phone labels and time stamps of the aligned segments of the temporally shortened regions should not be taken as ground truth, the detected sequence is certainly pronounced differently than the predicted full form. The larger the number of contiguous minimum duration segments, the stronger the hypothesis of an actual temporal reduction.

Whereas reduction, and more specifically temporal reduction, is often considered to be specific to casual or at least spontaneous speech, our comparative investigations of both prepared and spontaneous speech in English and French reveal that temporal reduction exists in both speech styles, although to a lesser extent in the former as can be expected. We believe that similar mechanisms underlie the production and processing of pronunciation variants in the different speaking styles. Temporal reduction involves unstressed stretches of speech more often than regions of focus. When examining the words most frequently included in minimal duration sequences, it is not surprising to find high-frequency function words. Frequency might thus be one of the factors explaining such reduction. However, considering reduced sequences in low-frequency words, major factors seem to be related to repetition (which is equivalent to a local boost in frequency) and to a prosodic grouping in an unstressed position. In all of the examined cases, reduced sequences are embedded in unstressed or nonemphasized portions of speech.

Prepared, formal, noninteractive speech is generally uttered in a relatively steady tempo, whereas spontaneous speech undergoes substantial fluctuations in tempo. Interactive speech also includes more discourse markers, which are particularly prone to temporal reduction. Many words and phrases may serve as discourse markers. For example, the high frequency of *I_don't_know* in English or *je_sais_pas* in French with their variously reduced surface forms in spontaneous speech is more often related to a discourse marker function, than to the expression of a lack of knowledge.

Temporally reduced speech is challenging for current ASR systems and for nonnative listeners, resulting in misrecognitions. A practical approach taken for ASR is the introduction of multiword expressions, which allow shorter pronunciations to be associated with the word sequence. The English expression *sort of*, which is frequently observed in both BN and CTS corpora and tends to be produced very quickly, is a good candidate for a multiword expression. Our investigations for French confirmed that the examples shown at the beginning of the chapter (*je crois que*, 'I believe that' and *je ne sais pas*, 'I don't know') are good candidates for multiword modeling in ASR. They generally have a very low average phone duration and are discourse marker-like

expressions comparable to multiwords in English. These expressions and corresponding audio samples which can help improve the performance of ASR systems could also be helpful for L2 training to better survive in a native speakers' environment.

By using forced alignment to quantify temporal reduction phenomena we have tried to demonstrate how ASR systems may serve as a tool to systematically investigate variations across different speaking styles and languages. We hope that the present results will shed some new light on the intrinsically complex nature of temporal processes in speech. In future work, we plan to refine the present approach and to further extend the analysis of the alignment results, with the aim of using this approach to discover new pronunciation variants attributable to temporal reduction. Studying linguistic phenomena from an ASR perspective using large corpora might also give us some clues about the encoding of information in speech. The speech signal is endowed with many fine phonetic details and features that the human listener is somehow able to rely on even in the face of ambiguity and noise. The perspectives available through an ASR approach are manifold. For researchers working in the domain of ASR, the ultimate goal is to uncover rules to improve pronunciation modeling. These rules can be applied to rarely observed or unobserved words, for which pronunciation variants cannot be estimated statistically. The framework developed should help to describe and quantify more or less well-known linguistic phenomena on the phonemic and lexical levels, which is of relevance to linguists and cognitive scientists alike.

Acknowledgments: Parts of the research reported in this chapter have been funded by grants from the CNRS, the French Investissements d'Avenir – Labex EFL program (ANR-10-LABX-0083), and ANR Vera and Quaero. We would like to thank Jean-Luc Gauvain and Gilles Adda for their help.

References

Adda-Decker, Martine & Natalie Snoeren 2011. Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics* 39. 261–270.

Adda-Decker, Martine, Cédric Gendrot & Noël Nguyen 2008. Contributions du traitement automatique de la parole à l'étude des voyelles orales du français. *Traitement Automatique des Langues* 49 (3). 13–46.

Adda-Decker, Martine, Philippe Boula de Mareüil, Gilles Adda & Lori Lamel 2005. Investigating syllabic structures and their variation in spontaneous French. *Speech Communication* 46. 119–139.

Adda-Decker, Martine & Lori Lamel 2005. Do speech recognizers prefer female speakers? *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Lisbon, 2205–2208.

Adda-Decker, Martine & Lori Lamel 1999. Pronunciation variants across system configuration, language and speaking style. *Speech Communication* 29. 83–98.

Bartkova, Katarina & Denis Jouvet 2015. Impact of frame rate on automatic speech-text alignment for corpus- based phonetic studies. *Proceedings of the 18th ICPhS*, Glasgow, paper no 667 (5 pages).

Di Canio, Christian, Hosung Nam, Douglas H. Whalen, Timothy Bunnell, Jonathan D. Amith & Rey C. Garcia 2012. Assessing agreement level between forced alignment models with data from endangered language documentation corpora, *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Portland.

Dilley, Laura C. & Mark Pitt 2007. A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition. *Journal of the Acoustical Society of America* 122. 2340–2353.

Duez, Danielle, 2003. Modelling Aspects of Reduction and Assimilation in Spontaneous French Speech. In *Proceedings of the IEEE-ISCA Workshop on Spontaneous Speech Processing and Recognition*, Tokyo.

Elman, Jeffrey L. & James L. McClelland 1988. Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language* 27. 143–165.

Ernestus, Mirjam 2000. Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface. Utrecht: LOT.

Ernestus, Mirjam & Natasha Warner (Eds.). 2011. Speech reduction [Special Issue]. *Journal of Phonetics* 39. Fougeron, Cécile, Jean-Philippe Goldman & Ulli H. Frauenfelder 2001. Liaison and schwa deletion in French: an effect of lexical frequency and competition. *Proceedings of ESCA Eurospeech*, Aalborg, 639–642.

Gahl, Susanne 2008. "Time" and "Thyme" are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84 (3). 474–496.

Galliano, Sylvain, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre & Guillaume Gravier. 2005. The Ester phase II evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Lisbon, pp. 1149–1152.

Ganong, William F. 1980. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* 6 (1). 110–125.

Gauvain, Jean-Luc, Lori Lamel, Gilles Adda & Martine Adda-Decker 1994. Speaker-independent continuous speech dictation. *Speech Communication* 15. 21–37.

Gauvain, Jean-Luc., Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Véronique Gendner, Lori Lamel & Holger Schwenk 2005. Where are we in transcribing French broadcast news? *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Lisbon, 1655–1658.

Godfrey, John J., Edward Holliman & Jane McDaniel 1992. Switchboard: telephone speech corpus for research and development. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE-ICASSP)*, San Francisco, 517–520.

Greenberg, Steven & Shuangyu Chang 2000. Linguistic dissection of Switchboard-corpus automatic speech recognition systems. *Proceedings International Speech Communication Association ISCA-ITRW Workshop on ASR*, Paris, 195–202.

Greenberg, Steven, Hannah Carvey & Leah Hitchcock 2002. The relation between stress accent and pronunciation variation in spontaneous American English discourse. In Proceedings of Speech Prosody, Aix-en-Provence, France, 351–354.

Greenberg, Steven, Hannah Carvey, Leah Hitchcock & Shuangyu Chang 2003. Temporal properties of spontaneous speech – a syllable-centric perspective. *Journal of Phonetics* 31. 465–485.

Johnson, Keith. 2004. Massive reduction in conversational American English. Proceedings Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium.

Jun, Sun-Ah & Cécile Fougeron 2002. The realizations of the accentual phrase in French intonation. *Probus (special issue on Intonation in the Romance Languages)*, J. Hualde, ed., vol. 14, 147–172.

Jurafsky, Daniel, Alan Bell, Michelle Gregory & William D. Raymond 2001. Probabilistic relations between words: Evidence from reduction in lexical production in *Frequency and the Emergence of Linguistic Structure*, Bybee and Hopper eds. John Benjamins, pp. 229–254.

Karanasou, Panagiota & Lori Lamel 2011. Pronunciation variants generation using SMT-inspired approaches. *36th International Conference on Acoustics, Speech and Signal Processing, IEEE-ICASSP)*, Prague, 4908– 4911.

Lamel, Lori & Gilles Adda 1996. On designing pronunciation lexicons for large vocabulary, continuous speech recognition. In *Proceedings of International Conference on Speech and Language Processing (ICSLP)*, Philadelphia, 6–9.

Lamel, Lori & Jean-Luc Gauvain 2005. Alternate phone models for conversational speech. In Proceedings of *IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE-ICASSP)*, vol. 1, Philadelphia, 1005–1008.

Lefèvre, Fabrice, Jean-Luc Gauvain & Lori Lamel 2005. Genericity and portability for task-dependent speech recognition. *Computer Speech and Language* 19. 345–363.

Nakamura, Masanobu, Sadaoki Furui & Koji Iwano 2006. Acoustic and linguistic characterization of spontaneous speech. *International Speech Communication Association (ISCA) workshop on Speech Recognition and Intrinsic Variation*, Toulouse.

Nguyen, Long, Sherif Abdou, Mohamed Afify, John Makhoul, Spyros Matsoukas, Richard Schwartz, Bing Xiang, Lori Lamel, Jean-Luc Gauvain, Gilles Adda, Holger Schwenk & Fabrice Lefèvre. 2004 The 2004 BBN/LIMSI 10XRT English broadcast news transcription system. *Proceedings DARPA Rich Transcription Workshop (RT04)*, Palisades.

Nguyen, Noel & Martine Adda-Decker (Eds.) 2013. Méthodes et outils pour l'analyse phonétique des grands corpus oraux. Traité IC2, série Cognition et traitement de l'information. Hermes-Lavoisier, ISBN 978-2-7462-4530-3.

Prasad, Rohit, Spyros Matsoukas, Chia-Lin Kao, Jeff Ma., Dong-Xin Xu Thomas Colthurst, Owen Kimball, Richard Schwartz, Jean-Luc Gauvain, Lori Lamel, Holger Schwenk, Gilles Adda & Fabrice Lefèvre 2005. The 2004 BBN/LIMSI 20xRT English conversational telephone speech recognition system. *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Lisbon, 1645–1648.

Rabiner, Lawrence R. & Biing-Hwang Juang 1986. An introduction to hidden Markov models. *IEEE Acoustics Speech and Signal Processing Magazine* ASSP-3 (1). 4–16. January.

Samuel, Arthur G. & Mark A. Pitt 2003. Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language* 48. 416–434.

Schuppler, Barbara, Mirjam Ernestus, Odette Scharenborg and Louis Boves 2008. Preparing a Corpus of Dutch spontaneous dialogues for automatic phonetic analysis. *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Brisbane, 1638–1641.

Schuppler, Barbara, Martin Hagmüller, Juan A. Morales-Cordovilla & Hannes Pessentheiner. 2014. GRASS: The Graz corpus of read and spontaneous speech. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, 1465–1470.

Schuppler, Barbara, Martine Adda-Decker & Juan A. Morales-Cordovilla 2014. Pronunciation variation in read and conversational Austrian German. *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Singapore, 1453–1457.

Snoeren, Natalie, Pierre Hallé & Juan Segui 2006. A voice for the voiceless: Production and perception of assimilated stops in French. *Journal of Phonetics* 34. 241–268.

Stolcke, Andreas, Harry Bratt, John Butzberger, Horacio Franco, Venkata R. Rao Gadde, Madelaine Plauché, Colleen Rickey, Elizabeth Shriberg, Kemal Sönmez, Fuliang Weng & Jing Zheng 2000. The SRI March 2000 hub-5 conversational speech transcription system. *Proceedings NIST Speech Transcription Workshop*, College Park.

Strik, Helmer, Diana Binnenpoorte & Catia Cucchiarini 2005. Multiword expressions in spontaneous speech: Do we really speak like that? In *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Lisbon, pp. 1161–1164.

Strik, Helmer, Anna Elffers, Dusan Bavcar & Catia Cucchiarini 2006. Half a word is enough for listeners, but problematic for ASR. In *Proceedings of International Speech Communication Association (ISCA) workshop on Speech Recognition and Intrinsic Variation*, Toulouse, 101–106.

Strik, Helmer & Catia Cucchiarini 1999. Modelling pronunciation variation for ASR: A survey of the literature. *Speech Communication* 29. 225–246.

Torreira, Francisco, Martine Adda-Decker & Mirjam Ernestus 2010. The Nijmegen corpus of casual French, Speech Communication 10 (3). 201–212.

Torreira, Francisco & Mirjam Ernestus 2012. Weakening of intervocalic /s/ in the Nijmegen corpus of casual Spanish. *Phonetica* 69. 124–148.

Tseng, Shu-Chuan 2005. Contracted syllables in Mandarin: Evidence from spontaneous conversation. *Language and Linguistics* 6 (1) 153–180.

Van Son, Robert J.J.H. & Louis C.W. Pols 2003. An acoustic model of communicative efficiency in consonants and vowels taking into account context distinctiveness. In *Proceedings of the 15th ICPhS*, Barcelona, 2141–2143.

Vasilescu, Ioana, Martine Adda-Decker, Lori Lamel & Pierre Hallé 2009. A perceptual investigation of speech transcription errors involving frequent near-homophones in French and American English. *Proceedings International Speech Communication Association (ISCA) Interspeech*, Brighton, 144–147.

Vasilescu, Ioana, Dahlia Yahia, Natalie Snoeren, Lori Lamel & Martine Adda-Decker 2011. Cross-lingual study of ASR errors: on the role of the context in human perception of near homophones. *Proceedings of International Speech Communication Association (ISCA) Interspeech*, Florence, 1949–1952.

Very-Large-Scale Phonetics Research (VLSP 2011). January 28–31, 2011 Philadelphia, PA, USA. http://www.ling.upenn.edu/phonetics/workshop/

Whalen, Douglas H. 1991. Infrequent words are longer in duration than frequent words. *Journal of the Acoustical Society of America* 90 (4). 2311.

Woehrling, Cécile (2009). Accents régionaux en français : perception, analyse et modélisation à partir de grands corpus. Université Paris Sud, Paris XI.

Mirjam Ernestus and Rachel Smith

5 Qualitative and quantitative aspects of phonetic variation in Dutch *eigenlijk*

Abstract: This chapter presents a detailed analysis of 159 tokens of the Dutch discourse marker *eigenlijk*, uttered in casual conversations. We provide quantitative analyses of the pronunciation variants characterized as segmental sequences and qualitative analyses of the more detailed phonetic characteristics of these variants. Our data demonstrate a wide range of variation in the production of the word, ranging from trisyllabic tokens closely resembling the word's citation form (13% of the tokens) to phonetically minimal monosyllabic tokens consisting merely of a vowel followed by a single obstruent consonant (36%). The full form is thus not the most frequent form. The reduced tokens occur both in prosodically weak and strong positions, contrary to what is typically reported for reduced words. The pronunciation variation displayed by *eigenlijk* is conditioned, among other factors, by the rhythm of the phrase, and shows large differences between speakers. Importantly, a form may be reduced in one aspect, but not in another. For instance, a bisyllabic form may be as long as a trisyllabic form and, whereas some forms still contain acoustic cues for the /l/ but not for the fricative, this is the other way around for other forms. Generally, we found that every pronunciation variant of *eigenlijk* includes two landmarks that may be considered to be the main characteristics of the word (the full vowel and a velar/uvular consonant). These findings raise new questions about reduction, including the status of full but infrequent forms, the extent to which reduction is an automatic process, and the role of landmarks in speech processing. We conclude that many aspects of the phenomenon of speech reduction are not yet well understood and call for more detailed qualitative and quantitative analyses of many tokens of individual words produced in casual speech. Like our study, these studies will substantially extend our knowledge about speech reduction and about speech production and perception.

Keywords: conversational speech, rhythm, landmarks, corpus, accent, coarticulation, Dutch

Mirjam Ernestus, Radboud University and Max Planck Institute for Psycholinguistics
Rachel Smith, Glasgow University Laboratory of Phonetics, University of Glasgow

https://doi.org/10.1515/9783110524178-005

5.1 Introduction

In informal conversations in many languages, many word tokens are pronounced with fewer segments or with segments that are articulated more weakly than in careful speech (for an introduction to the phenomenon, see Ernestus and Warner 2011). For instance, the word *particular* may be pronounced like [pʰtʰɪkʰə] and *hilarious* like [hlɛrɛs] (Johnson 2004). These short word pronunciation variants are generally referred to as reduced forms, and we adopt this terminology here. Reduced forms typically occur in weak prosodic positions, especially in unaccented positions in the middle of sentences (e.g., Pluymaekers, Ernestus, and Baayen 2005a). This chapter contributes to our knowledge of the characteristics of reduced forms by studying in detail one word type in Dutch (i.e., *eigenlijk* 'actually') that is known to show wide variation in its realization (e.g., Ernestus 2000: 141). The results shed light on the variation that a word may show and on how speakers from the same sociolinguistic group may differ in how they reduce words. In addition, the results raise questions about the nature of reduced forms, the mental lexicon, and psycholinguistic models of speech production and comprehension.

Nearly all previous research on reduced forms focused on the presence versus absence of segments and on the duration of these segments or (parts of) the words as measures of reduction. These studies have shown that many different factors affect the probability that a given word appears in a reduced form. These factors include speech rate (e.g., Kohler 1990; Raymond, Dautricourt, and Hume 2006), the word's phonological neighborhood density (Gahl, Yao, and Johnson 2012), its prosodic position (e.g., Bell et al. 2003), its a priori probability (e.g., Pluymaekers, Ernestus, and Baayen 2005a; Gahl 2008), its probability in context (e.g., Bell et al. 2003; Bell et al. 2009; Pluymaekers, Ernestus, and Baayen 2005b), and the presence versus absence of a following hesitation (e.g., Bell et al. 2003). The influences of these factors suggest that reduction may result from time pressure: when time pressure is high, for instance because speech rate is high or because the word or the following word was easy to plan and is ready to be articulated, reduction is more likely to occur (e.g., Bell et al. 2009; Gahl et al. 2012).

In addition, several studies focusing on duration and on the presence versus absence of segments suggest that degree of reduction is under the speaker's direct control. For instance, speakers may choose not to reduce at high speech rates (e.g., van Son and Pols 1990, 1992), and degree of reduction correlates with speaker characteristics, including gender (e.g., Guy 1991; Phillips 1994), age (e.g., Guy 1991; Strik, van Doremalen, and Cucchiarini 2008), and socioeconomic status (e.g., Labov 2001). Furthermore, speakers of different regiolects of a language may differ in degree of reduction for some words (e.g., Keune et al. 2005).

Reduction is therefore not a fully automatic process, but is speaker dependent and probably at least partly under the speaker's control.

Only a few studies so far have investigated more detailed phonetic characteristics of reduced forms. Such studies have shown that information about a word's identity is often preserved despite reduction or reorganization of the acoustic features or articulatory gestures that would be found in a canonical form. For example, reduced tokens of *support* may lack any evidence of a vowel portion between /s/ and the closure of /p/, yet may maintain aspiration of /p/, which is consistent with a singleton syllable-initial stop, rather than an /sp/ cluster. Thus, the laryngeal specification of the stop preserves information that prevents reduced *support* from becoming ambiguous with *sport* (Manuel 1991; Manuel et al. 1992; see also Davidson 2006; see Aalders and Ernestus, in preparation, for evidence that this also holds in casual speech). Similarly, some reduced forms of French *c'était* 'it was' can lack a voiced vowel between /s/ and /t/, yet retain traces of the vowel in the form of a lowered spectral center of gravity in the latter part of the /s/ (Torreira and Ernestus 2011). In English *the*, /ð/ can assimilate in manner of articulation to a preceding nasal or lateral (in phrases like *in the*, *all the*), losing any evidence of frication; yet residues of /ð/ tend to be retained in the form of dentality (as cued by F2 at the nasal or lateral boundaries) and duration (Manuel 1995).

In extreme cases of reduction it may be impossible to linearly segment the speech signal, yet sufficient phonetic residue of a word's form may remain as to make it fully identifiable. Kohler (1999) described such residues as "articulatory prosodies," which "persist as nonlinear, suprasegmental features of syllables, reflecting, for example, nasality or labiality that is no longer tied to specific segmental units" and may be quite extended in time (p. 89). For example, the German discourse marker *eigentlich* 'actually' is canonically produced as [aɪgŋtliç], but can be reduced to [aɪŋi] or [aĩĩ], with palatality, nasality, and duration serving to convey the word's "phonetic essence" (Niebuhr and Kohler 2011). Perception tests indicate that listeners may be sensitive to such articulatory prosodies, even in the absence of contextual clues (Niebuhr and Kohler 2011), just as they are to other aspects of phonetic detail in reduced speech (Manuel 1991, 1995). Thus, reduction may involve significant departures from a word's canonical form, while preserving phonological contrast quite well (Warner and Tucker 2011).

The degree of reduction may be affected by the function that a word performs. Plug (2005) investigated reduction of the Dutch discourse marker *eigenlijk* 'actually' as a function of its pragmatic function. Plug analyzed 49 tokens of *eigenlijk* performing two of the word's subfunctions, one being correction or clarification of a statement or assumption in a speaker's own utterance (self-repair), and the other being correction or clarification of something said

or assumed by the interlocutor (other-repair). He observed that tokens with the function of self-repair tended to be produced fast and to be highly reduced in terms of their number of syllables and segments. Tokens whose function was other-repair tended to occur at the edges of prosodic phrases and to be produced slower and with more phonetic elaboration.

Like Plug (2005), the present study focuses on the Dutch word *eigenlijk*. As is the case for nearly all words in every language, we have very little *detailed* knowledge about the possible pronunciation variants of the word, and about how frequently these variants occur and under which conditions. The present study examines over 150 tokens of this word, produced by 18 speakers in informal conversations, examining their detailed phonetic characteristics and when particular clusters of characteristics are most likely to occur. Detailed data on the pronunciation variation of this word will form a good testing ground for common assumptions about reduction, including the assumption that reduced forms only occur in prosodically weak positions, and the related assumption that speakers mainly reduce to cope with time pressure.

Our main reason for studying *eigenlijk* is that it is known to show a wide variation in pronunciation, ranging from trisyllablic /'ɛɪxələk/ (see Figure 5.1 for an example) to monosyllabic variants like /'ɛɪxk/ and /'ɛɪk/ (see, e.g., Ernestus 2000;

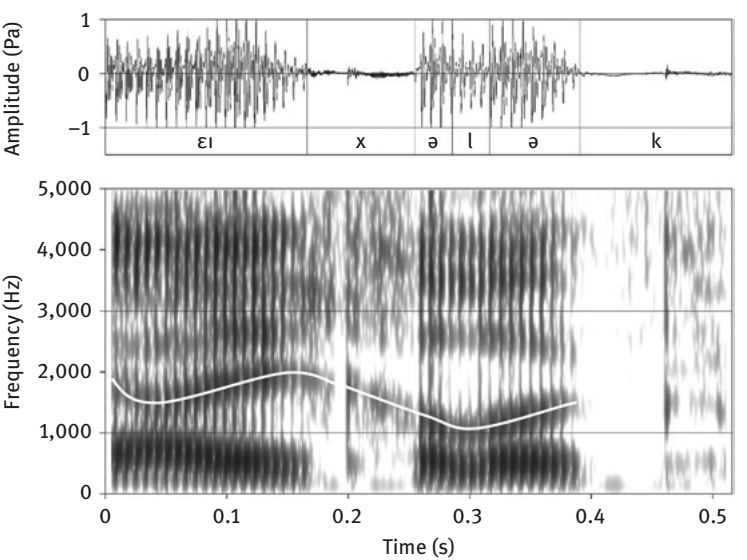

Figure 5.1: Waveform and spectrogram of an unreduced token of *eigenlijk*, produced as [ɛɪxələk] by speaker S in the Ernestus Corpus of Spontaneous Dutch (Ernestus 2000). The white line indicates the F2 trajectory.

Plug 2005; Pluymaekers, Ernestus, and Baayen 2005b). The word shares this variability with many other words also ending in the suffix /ləк/ *-lijk* (e.g., Pluymaekers, Ernestus, and Baayen 2005b). The word occurs relatively frequently in informal conversations (e.g., 1,922 tokens per million word tokens in the Spoken Dutch Corpus, Oostdijk 2000), which allows us to study both intra- and interspeaker variations on the basis of tokens produced in a relatively short period of time.

As mentioned above, the word *eigenlijk* is a discourse marker that in general signals a contrast between what the speaker is saying and what (s)he or the interlocutor just said or implied, or is assumed to believe (e.g., Plug 2005; van Bergen et al. 2011); see, for instance, sentences (1, 2, 3) from our data set.

(1) *één van de, of eigenlijk de oudste, acht geveild zou worden.*
one of the, or actually the oldest, eight [a type of rowing boat] would be auctioned.

(2) *Nee, tenten hoef ik eigenlijk niet*
No, I actually do not need tents.
(In response to the interlocutor's request whether he would like to buy any tents.)

(3) *Van tandartsen word ik altijd eigenlijk helemaal nooit goed.*
I always actually completely never feel good around dentists
(Following the speaker's remark that he has a dentist appointment next Monday)

Within this broad function, several subfunctions can be distinguished. As described above, Plug (2005) investigated phonetic characteristics of 49 tokens representing two of these subfunctions. In the present paper, we will not distinguish between these subfunctions because they are often difficult to distinguish and because a focus on one or several of the subfunctions would severely reduce the number of word tokens that can be analyzed (as in Plug's study).

The Dutch word *eigenlijk* can occur in different positions in the sentence as illustrated in examples (1, 2, 3, 4, 5). Moreover, it can follow and precede different word types, as illustrated in these same examples. Noteworthy is example (3), in which *eigenlijk* is surrounded by other adverbs, as is frequently the case in spontaneous conversations.

(4) *Eigenlijk is dat actief.*
Actually that is active.

(5) *Het gaf heel veel informatie eigenlijk.*
It gave a lot of information actually.

The first part of this study provides an overview of the types of variation that we attest in our data set, and of the frequencies with which specific phonetic characteristics occur. The second part investigates which factors predict certain phonetic properties, including the number of syllables, presence of creaky voice,

and the presence versus absence of /l/. Our analyses will include predictors that have been reported before to correlate with reduction degree (e.g., speech rate, the predictability of the preceding and following word, and the presence of sentential accent).

We also investigate the influence of a new predictor, the rhythm of the sentence. Because the word *eigenlijk* can be preceded and followed by a wide variety of words, the numbers of directly preceding and following unstressed syllables can vary as well. Speakers of Germanic languages prefer alternating patterns of stressed and unstressed syllables (e.g., Kelly and Bock 1988 and references therein). It is therefore possible that speakers of Dutch opt for a variant of *eigenlijk* with a number of unstressed syllables that optimizes the rhythm of the phrase. For instance, they may prefer a stressed monosyllabic variant if the word is followed by several unstressed syllables, and a di- or trisyllabic variant, ending in one or two unstressed syllables, respectively, when the word is followed by a word with initial stress.

The following sections describe the corpus and our selection of the tokens (Section 5.2), the annotation system (Section 5.3), and provide a qualitative description of the attested variation (Section 5.4). We then present the results of our statistical modeling of some of the tokens' characteristics (Section 5.5). We conclude the chapter with a general discussion of these results (Sections 5.6 and 5.7).

5.2 Materials

We extracted the tokens from the Ernestus Corpus of Spontaneous Dutch (Ernestus 2000). This corpus, recorded in the 1990s, contains high-quality recordings of ten conversations, each 90 minutes long, between pairs of friends or direct colleagues. A Digital Audio Tape (DAT) recorder recorded the speech by each interlocutor on a different track of a tape, by means of unidirectional microphones placed on a table between the speakers. The speakers are all male, highly educated, and lived their whole lives in the western part of the Netherlands. They speak a "western" variant of Standard Dutch, which implies, among other things, that they do not distinguish between the voiced and voiceless velar fricative.

The conversation during the first part of each recording was elicited by a third person, who knew at least one of the speakers well. The speakers discussed topics as diverse as television quizzes, how they chose their professions, their experiences with dentists, and their opinions of euthanasia. In the second part of the recording, the speakers participated in a role-play in which one speaker sold tents, sleeping bags, and backpacks to the other speaker, who pretended to be a shop owner. This

Table 5.1: The number of tokens produced by each speaker, the number incorporated in our analyses, and the number of studied tokens that were monosyllabic (see the section on Individual speaker differences).

Speaker ID	Total tokens	Tokens studied	Monosyllabic tokens (percentage of the tokens studied)
A	6	5	5 (100%)
B	10	9	2 (22%)
E	12	8	1 (13%)
F	26	10	0 (0%)
G	8	–	–
H	20	8	4 (50%)
I	15	9	1 (11%)
J	10	9	1 (11%)
K	28	–	–
L	45	11	0 (0%)
M	20	9	7 (78%)
N	13	7	2 (29%)
O	9	9	8 (89%)
P	12	11	2 (20%)
Q	26	10	1 (10%)
R	15	7	5 (71%)
S	17	10	3 (30%)
T	15	9	3 (67%)
U	22	9	7 (78%)
V	10	9	5 (56%)

second part also contained conversations covering a wide range of other topics since the speakers were encouraged to converse freely before and after the negotiations. The speech in the corpus sounds natural and casual, as is evidenced, among other things, by ratings of six other native speakers, the high frequency of discourse markers (including *eigenlijk*), and the amount of gossip in the corpus.

The 20 speakers in the corpus produced in total 339 *eigenlijk* tokens. The number of tokens per speaker ranges from 6 to 45 (see Table 5.1). We randomly selected 159 tokens, taking into account the following constraints and preferences. First, we only incorporated tokens that were produced fluently, without laughing and without much background noise (including the interlocutor's speech), so that detailed phonetic analysis is possible. Second, we wished to have minimally five tokens per speaker, so that we could study intraspeaker variation, and maximally 11 tokens, so that the data set would not be dominated by just a few speakers. Third, we preferred tokens produced in the free conversations over tokens produced in the role-play and we discarded tokens that were produced in the first 10 minutes of a recording because the speaker might not yet have been

speaking very naturally. Many tokens did not meet all these requirements and preferences and, as a consequence, we lost Speaker G. We also decided not to incorporate Speaker K because this speaker often stumbled over his words. Table 5.1 shows the resulting number of tokens per speaker.

5.3 Labeling procedure

A phonemic transcription of the entire intonational phrase containing *eigenlijk* was made by the first author, and checked by the second author. For the token of *eigenlijk*, an allophonic transcription was also made, specifying voicing of /k/ and /x/, but no other detail. The number of syllables in *eigenlijk* was identified. Then, labeling of prosody and of segmental detail was carried out by the two authors independently. All cases where the transcribers used different labels were resolved by joint listening, as were all cases where they placed labels more than 20 ms apart, which was not often the case. There was no obvious bias toward either transcriber's labeling.

For the prosodic labeling, the boundaries of the intonational phrase containing *eigenlijk* were annotated, and all syllables within this phrase were labeled as primary accented, secondary accented, stressed, or unstressed, that is, we distinguished four levels of prosodic strength, defined as follows.

Primary accented: the most prominent pitch-accented syllable in the phrase. All phrases included minimally one primary-accented syllable; only six included two primary-accented syllables, and most of these were produced by the same speaker and contained equally prominent accents on *eigenlijk* and another word.

Secondary accented: lexically stressed syllables that were produced with a pitch movement or, rarely, a substantial increase in loudness in the absence of a pitch accent.

Stressed syllables: lexically stressed syllables that were produced without a pitch movement. Unaccented function words were labeled as stressed if they had a full vowel and no evidence of segmental reduction, or as unstressed otherwise.

Unstressed: syllables lacking lexical stress. Filled pauses were always labeled as unstressed.

For the labeling of segmental detail, we defined a set of articulatory events in the larynx and supraglottal tract which are present in an unreduced token of *eigenlijk*, including the onsets and offsets of periodicity and the onsets and offsets of creaky voice. These events are described in detail in the following paragraph and illustrated in Figure 5.2. If an event was identifiable in the waveform

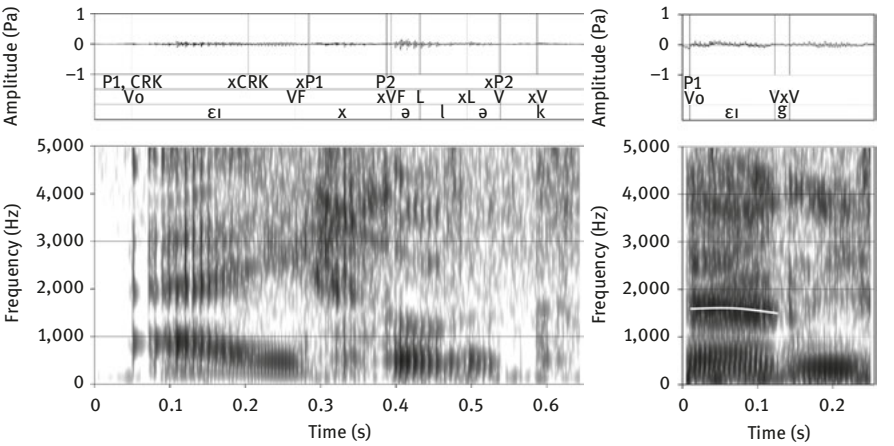

Figure 5.2: Labeling of segmental detail of two tokens of *eigenlijk*. Left: an unreduced token produced as [ɛɪxələk] by speaker H. Right: a reduced token produced as [ɛɪg] by speaker M (the following word, *niet* 'not,' is also shown). In each case, the top tier indicates laryngeal events, the second tier indicates events involving the upper articulators, and the third tier shows the allophonic transcription (see text for details). The white line indicates the F2 trajectory.

and spectrogram, it was labeled. If it was absent or unidentifiable because of reduction, then that label was omitted. By labeling events rather than segments, we aimed to achieve maximum comparability across tokens that had different degrees of reduction, while avoiding parsing the signal exhaustively into phoneme-sized segments, which can be very challenging for reduced speech. We made one exception: if a velar stop was present, we marked its offset in the signal, even if the stop was unreleased, as long as it was directly followed by another stop. In these cases, we placed the boundary for the stop in the middle of the long closure formed by the two stops. We thus maximized the number of velar stops whose durations we could analyze (see below). However, utterance-final unreleased stops were excluded from this analysis, as we had no principled way to estimate their durations.

As Figure 5.2 shows, on the *larynx* tier we labeled the onsets and offsets of periodicity (numbered as P1 and xP1 for the first portion of periodicity, P2 and xP2 for the next, etc.). We also labeled the onset (CRK) and offset (xCRK) of creaky voice, if present during /ɛɪ/. Our criterion for identifying creaky voice was irregularity of periods. On the *upper articulators* tier we labeled the following acoustic events: the onset of the stressed /ɛɪ/ vowel (Vo); and the onset and offset of velar frication (VF and xVF, respectively), of lateral quality (L, xL), and of velar closure (V, xV). These labels allowed us to calculate the duration of the whole

word token (defined as extending from the onset of the stressed vowel to the last labeled event) and durations of individual segments, including the unstressed vowels, if present.

5.4 General description of variation within the word

The Appendix lists the variation that we discuss in this and the following section for which we can provide frequency data.

We first focus on the variation in the number of syllables. Figure 5.3 shows the number of mono-, di-, and trisyllabic tokens for the four prosodic statuses that we distinguished. We see that the majority of tokens are disyllabic, and are thus one syllable shorter than the full form. Moreover, we find that many disyllabic and some monosyllabic tokens are accented (8 monosyllabic tokens are primary accented and 16 secondary accented). The word *eigenlijk* thus does not follow the well-known pattern that accented word tokens show little reduction (e.g., Bell et al. 2003). The variation in the number of syllables does not just result from the prosodic status of the word.

With regard to duration, tokens were on average longer when they had a greater number of syllables (3 syllables: 386 ms; 2 syllables: 310 ms; and 1 syllable: 197 ms), but Figure 5.4 shows that the durational ranges for tri-, di-, and monosyllabic tokens overlapped considerably, and we observed trisyllabic tokens as short as 257 ms.

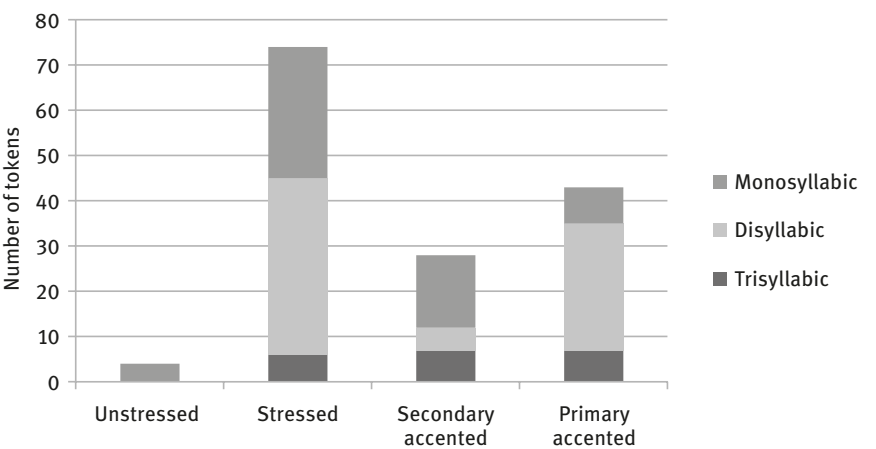

Figure 5.3: Number of trisyllabic, disyllabic, and monosyllabic tokens that are unstressed, stressed, secondary accented, or primary accented.

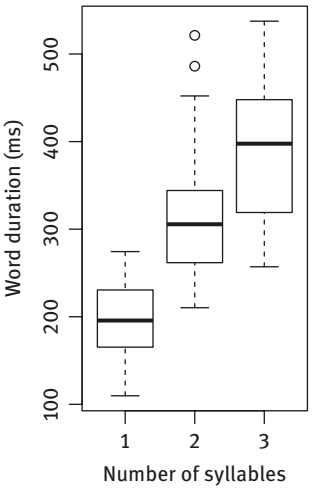

Figure 5.4: Boxplot of the duration of *eigenlijk* (ms), according to its number of syllables. The bottom and top of each box indicate the first and third quartiles; the band inside the box represents the second quartile (the median), while the whiskers extend to the minimum and maximum values that are maximally 1.5 interquartiles from the box. The two small circles are outliers.

Table 5.2 lists the phonemic transcriptions of the tokens, and allophonic transcriptions specified for the voicing of /k/ and /x/ but showing no other detail, along with the frequency of each transcription. The most frequent phonemic form is disyllabic /ɛɪxləkˌ/ (57 tokens). The monosyllabic form /ɛɪkˌ/ (22 tokens) comes in second position, and the full form /ɛɪxələkˌ/ (20 tokens) in third. Also common are monosyllabic forms /ɛɪxkˌ/ (16 tokens) and /ɛɪxˌ/ (13 tokens) and the disyllabic /ɛɪxəkˌ/ (14 tokens).

Importantly, we did not see a clear correlation between the structure of a token and how clearly its segments were articulated. Tokens that are highly reduced in terms of their number of syllables and segments, could nonetheless exhibit clearly articulated and tightly coordinated segments, and vice versa, as discussed below.

Table 5.2 makes clear that all forms in our data set contain, minimally, a full front vowel, and at least one obstruent articulated at the back of the mouth. These appear to be essential phonetic components of *eigenlijk*. The vowel is typically a closing diphthong, but varies in degree of diphthongization, and can look and sound quite monophthongal (e.g., Figure 5.2, right panel; Figure 5.9). As regards the obstruent(s), 114 tokens (72%) were transcribed as containing both a fricative and a stop, 19 (12%) only a fricative, and 26 (16%) only a stop. The obstruents are typically voiceless, but are voiced in 18% of cases. The fricative's place of articulation is normally velar or uvular. Eighty-seven tokens (55%) were transcribed as containing /l/, while several more contain a residual trace of /l/, as discussed further below.

Table 5.2: Tokens of *eigenlijk* found in the corpus. Phonemic and allophonic transcriptions are shown. The allophonic transcriptions specify the voicing of /k/ and /x/, but no other detail. Note that [ɛɪxl] was once perceived as disyllabic and once as monosyllabic.

Token structure	N	Transcription	
		Phonemic	**Allophonic**
Trisyllabic	**20**		
Vowel + fricative + schwa + lateral + schwa + stop	20	/ɛɪxələk/ 20	[ɛɪxələk] 17, [ɛɪxələg] 2, [ɛɪɣələk] 1
Disyllabic	**82**		
Vowel + fricative + lateral + schwa + stop	60	/ɛɪxlək/ 57	[ɛɪxlək] 42, [ɛɪxləg] 7, [ɛɪɣlək] 8, [ɛɪɣləg] 1
		/ɛɪxləŋ/ 1	[ɛɪxləŋ] 1
		/ɛxlək/ 2	[ɛxlək] 1, [ɛxləg] 1
Vowel + fricative + schwa + stop	15	/ɛɪxək/ 14	[ɛɪxək] 6, [ɛɪxəg] 2, [ɛɪɣək] 5, [ɛɪɣəg] 1
		/ɛxək/ 1	[ɛxəg] 1
Vowel + fricative + lateral	1	/ɛɪxl/ 1	[ɛɪxl] 1
Vowel + fricative + schwa + lateral	1	/ɛɪxəl/ 1	[ɛɪxəl] 1
Vowel + lateral + schwa + stop	4	/ɛɪləl/ 4	[ɛɪləl] 3, [ɛɪləg] 1
Monosyllabic	**57**		
Vowel + fricative	16	/ɛɪx/ 13	[ɛɪx] 11, [ɛɪɣ] 2
		/ɛx/ 3	[ɛx] 2, [ɛɣ] 1
Vowel + stop	22	/ɛɪk/ 22	[ɛɪk] 13, [ɛɪg] 9
Vowel + fricative + stop	18	/ɛɪxk/ 16	[ɛɪxk] 15, [ɛɪɣk] 1
		/ɛxk/ 2	[ɛxk] 2
Vowel + fricative + lateral	1	/ɛɪxl/ 1	[ɛɪxl] 1
Total	**159**		

Trisyllabic tokens were produced as the canonical form /ɛɪxələk/. Figures 5.1 and 5.2 (left panel) show typical trisyllabic tokens. Note the formant dynamics, in particular how F2 rises through the first diphthong to reach a maximum at the start of /x/, then falls to reach its minimum during /l/, before rising again into /k/. The first schwa is typically shorter than the second (mean durations 19 versus 46 ms, respectively).

Despite containing all or almost all of the segments expected in the canonical form, some of the trisyllabic tokens did display elements of reduction. Some were quiet and/or breathy, particularly if phrase-final. Others were rapidly articulated: Figure 5.5 shows a trisyllabic token that is very short indeed (268 ms) and whose formant dynamics follow a less extreme version of the pattern described above. Among the types of reduction found in trisyllabic tokens were also monophthongization of the vowel, incomplete closure of the stop, and devoicing of the first schwa.

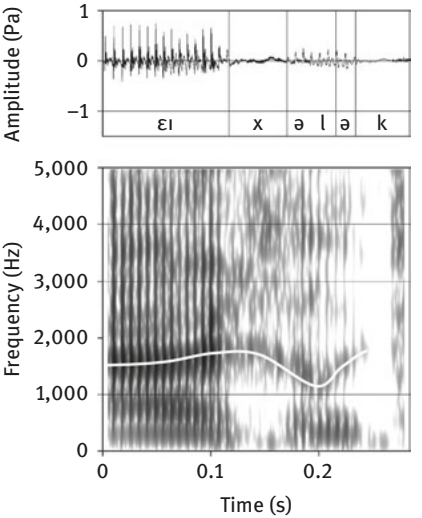

Figure 5.5: Trisyllabic token of *eigenlijk* produced as [ɛɪxələk] by speaker E. Note the short duration and the less extreme excursions of F2 (indicated by the white line) compared to the trisyllabic tokens in Figures 5.1 and 5.2. The white line indicates the F2 trajectory.

Disyllabic tokens took a wide range of forms. The most common disyllabic form was /ɛɪxlək/, very similar to trisyllabic tokens, but with no discernible schwa before /l/. Such schwa loss in unstressed syllables before /l/ and /r/ is very common also in English (e.g., Cruttenden 1994).

Disyllabic tokens evinced a range of interesting phonetic behavior at the juncture between the stem and suffix. First, there was variation in the extent to which the expected laryngeal events were produced, and in how they were aligned with respect to events involving the upper articulators. For example, we encountered numerous cases of progressive voice assimilation of /l/ to the velar fricative, that is, cases where /l/ and on occasion the word's entire second syllable were devoiced (e.g., Figure 5.6). We also found cases of regressive voice assimilation that is, where /x/ was voiced by assimilating to /l/, sometimes resulting in a token that was voiced in its entirety (e.g., Figure 5.7). Voice assimilation involving /l/ has not been described for Dutch so far. When voicing of the fricative occurred, it was sometimes accompanied by weakening of the degree of stricture, and/or loss of place cues, such that the frication sounded glottal rather than velar or uvular. In four disyllabic tokens, phonemically /ɛɪlək/, we found no evidence of velar frication at all, but simply a lateral approximant at the syllable boundary.

Second, in disyllabic tokens we found variation in the alignment of events involving the upper articulators. In several tokens, most of them spoken by speaker F, the /x/ and /l/ appear strongly coarticulated. This speaker seems to produce lateral frication (e.g., Figure 5.8, left panel) as a solution to the problem of producing two very different articulations in swift succession. A different

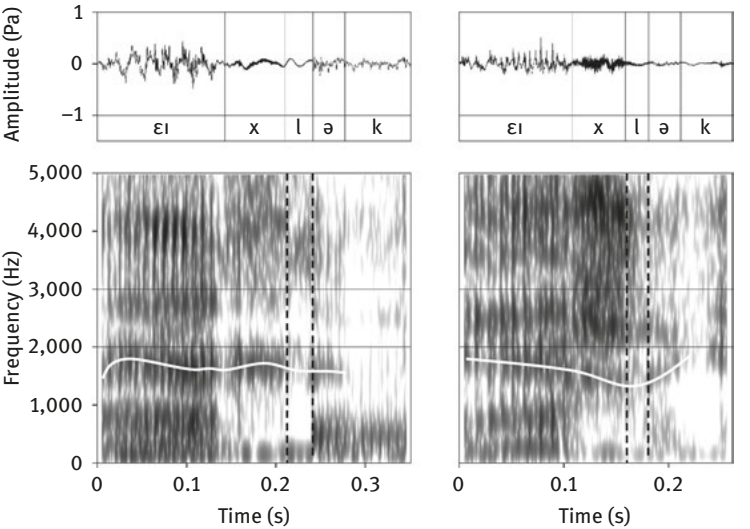

Figure 5.6: Disyllabic tokens of *eigenlijk* produced as [ɛɪxlək], illustrating devoicing in the second syllable. Left: Token produced by speaker N, with a devoiced [l]. Right: Token produced by speaker V, where the second syllable is devoiced, but preserves the formant dynamics consistent with a [lə] sequence. Dashed lines on spectrograms indicate the boundaries of [l]. White lines indicate the F2 trajectory.

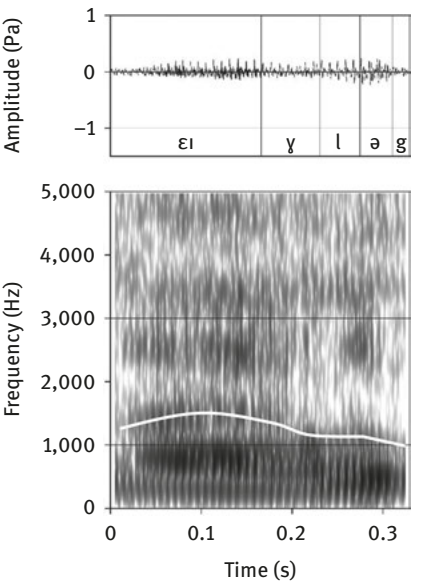

Figure 5.7: Fully voiced disyllabic token of *eigenlijk* produced as [ɛɪɣləg] by speaker T. The white line indicates the F2 trajectory.

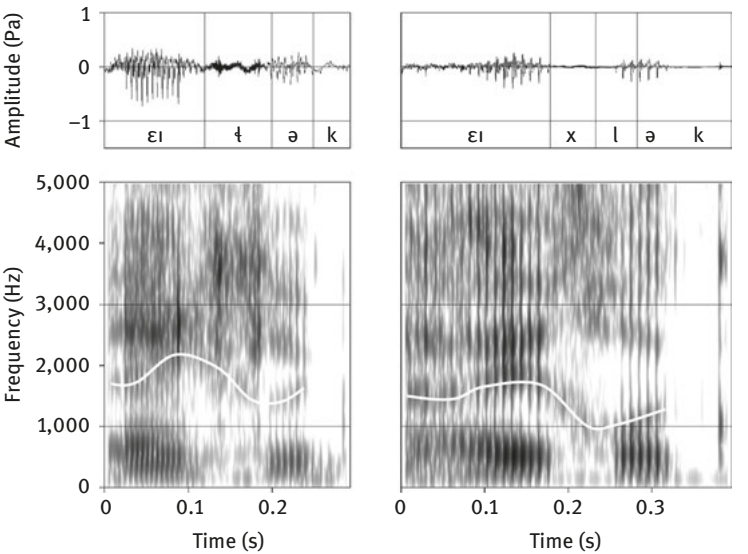

Figure 5.8: Disyllabic tokens of *eigenlijk* in which strong coarticulation between velar frication and laterality is audible. White lines indicate F2 trajectories. Left: token produced as [ɛɪɬək] by speaker F, that is, with an apparent lateral fricative. Right: token produced as [ɛɪxlək] by speaker S, in which F2 (indicated by the white line) falls rather steeply from the start of the frication.

strategy is adopted by speaker S (e.g., Figure 5.8, right panel), who appears to start backing the tongue body in preparation for the /l/ already from the start of the frication, such that F2 reaches its minimum in the middle of the fricative portion. Finally, among the tokens transcribed phonemically as /ɛɪxək/, we also observed cases where an /l/ was not unambiguously present, but nevertheless left some residual trace in the signal, in the form of an F2 dip (e.g., Figure 5.9).

Monosyllabic tokens also took a number of forms. Some ended in a sequence of fricative followed by stop (phonemically /ɛɪxk/). The obstruent cluster /xk/ is not a legitimate syllable coda in Dutch. It was often produced with rather long duration relative to the vowel (e.g., Figure 5.10). A small number of tokens, with particularly long obstruent clusters, were difficult to classify in terms of their number of syllables: despite having only one syllable peak, they sounded almost disyllabic (see Aoyagi 2015 for a possible theoretical account of this finding).

Other monosyllabic tokens contain a single voiceless obstruent, either /x/ or /k/. The formant dynamics of the vowel are quite variable in these tokens. Some tokens show a flat or falling F2 in the vowel (e.g., Figure 5.11). Others have a typically diphthongal vowel ending in a clear velar pinch (convergence of F2 and F3) at the transition into the obstruent (e.g., Figures 5.12 and 5.13). We are not sure to

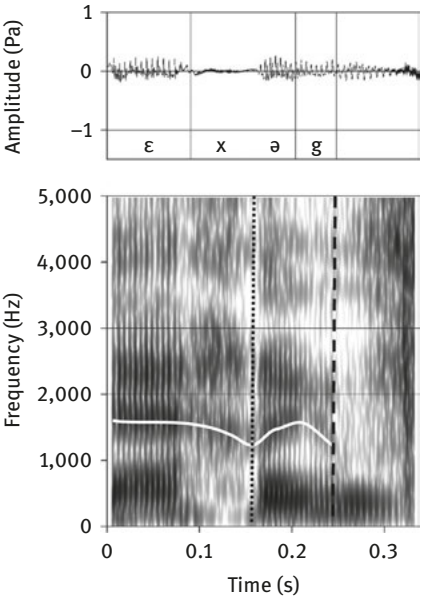

Figure 5.9: Disyllabic token of *eigenlijk*, produced as [ɛxəg] by speaker I in the context *dat doe ik eigenlijk nooit* 'I actually never do that.' This token was not heard as containing a definite [l], but the spectrogram indicates a residual trace of [l], manifest as an F2 dip around 0.15 seconds (the white line indicates F2, and the black dotted line the F2 minimum). Note also the assimilation of the final stop to the following [n] in terms of voicing and nasality. The black dashed line indicates the start of *nooit* (produced with laughter).

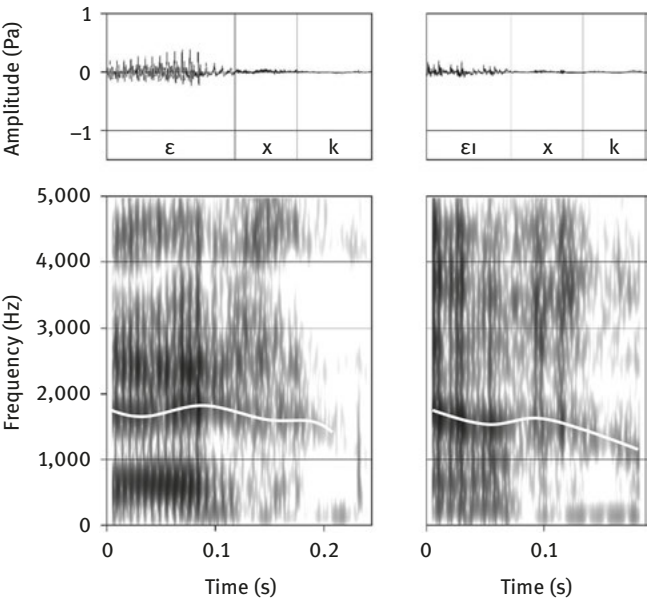

Figure 5.10: Monosyllabic tokens of *eigenlijk*, produced as [ɛxk] by speaker I (left panel) and as [ɛɪxk] by speaker O (right panel). Note the long duration of the obstruent portion compared to the vowel in both tokens. White lines indicate F2 trajectories.

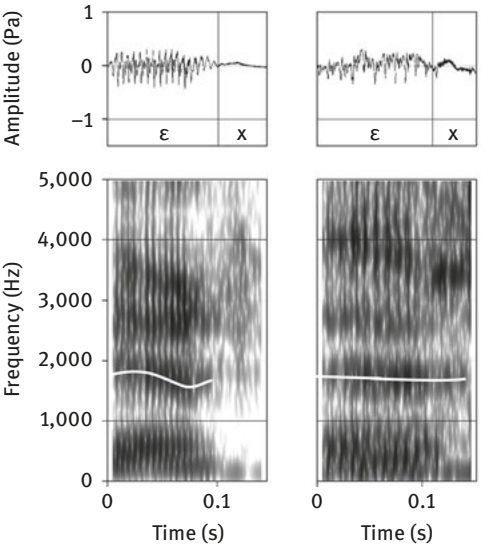

Figure 5.11: Monosyllabic tokens of *eigenlijk*, produced as [ɛx] by speakers P (left panel) and N (right panel). In each case F2 is indicated by a white line; note the flat or falling F2 at the transition into the obstruent. White lines indicate F2 trajectories.

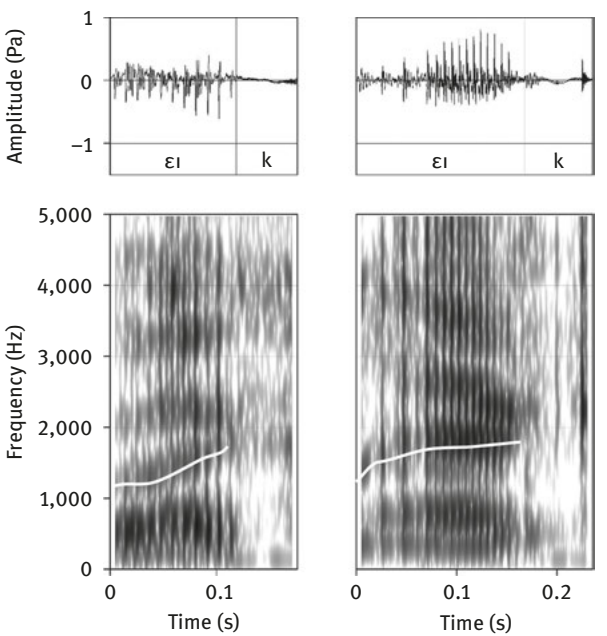

Figure 5.12: Monosyllabic tokens of *eigenlijk*, produced as [ɛɪk] by speakers S (left panel) and R (right panel). F2 is indicated in each case by a white line; note its rise and the convergence of F2 and F3 (velar pinch) at the transition into the obstruent. White lines indicate F2 trajectories.

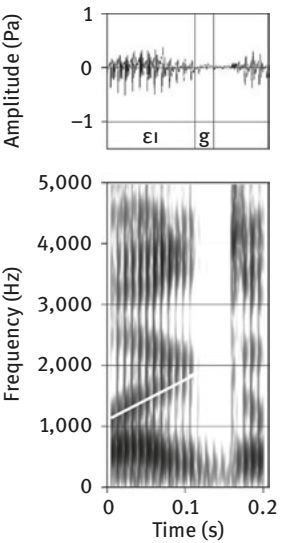

Figure 5.13: Monosyllabic token of *eigenlijk*, produced by speaker S as [ɛɪg] in the context *eigenlijk door* 'actually by.' Note the velar pinch at the transition into the stop. The white line indicates the F2 trajectory. The final stop is also assimilated in voice to the following [d] and is unreleased.

what to attribute the difference between these two patterns, but it may relate to the place of articulation of the obstruent: the diphthongal vowels may precede consonants with a velar place, and the vowels with flat or falling F2 may precede post-velar or uvular consonants (for comments on how a velar versus uvular place distinction can affect vowel formants, see Gordon, Barthmaier and Sands 2002).

Finally, creaky voice was common in the initial full vowel /ɛɪ/. This vowel often carries stress or accent (see, e.g., Figure 5.3), and according to van Jongenburger and van Heuven (1991), the word may therefore be expected to be preceded by a glottal stop or similar phonetic events (i.e., glottalization/creaky voice), especially after a vowel. Indeed, half of our tokens showed creaky voice at the start of the vowel (83 cases, or 52% of the data set), and less frequently other types of non-modal voice quality, such as harshness or breathiness.

5.4.2 Voice assimilation

The *eigenlijk* tokens, whether tri-, di-, or monosyllabic, show unexpected patterns of voice assimilation. Dutch is typically assumed to have only two processes of voice assimilation affecting sequences of obstruents. The first process voices obstruents preceding /b/ and /d/, while the second devoices /v/ and /z/ after voiceless obstruents (e.g., Booij 1995: 58, 59). We observed examples of these processes: see for example Figure 5.13. Yet, we also found cases of assimilation not described in the literature. Thirteen tokens showed voicing of the word-final

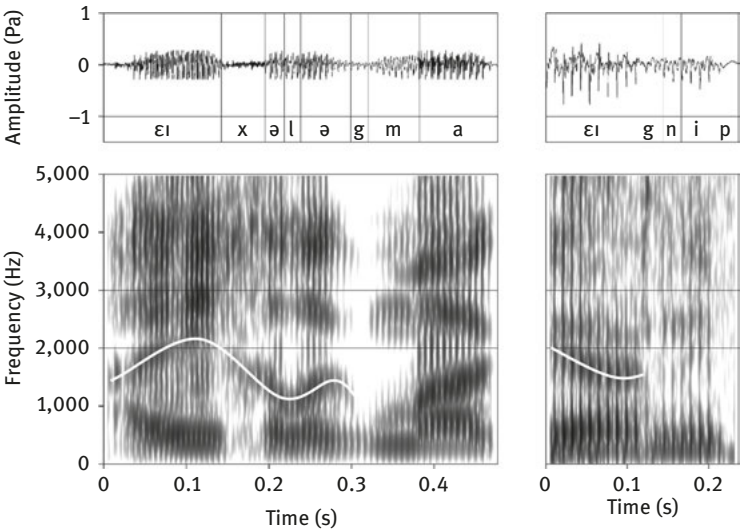

Figure 5.14: Tokens of *eigenlijk* where the final obstruent assimilates in voice to a following nasal segment. Left panel: [ɛɪxələg], in the context *eigenlijk maak ik*… 'actually I make…,' by speaker I. Right panel: [ɛɪg], in the context *ik weet het eigenlijk niet precies* 'I actually do not know exactly,' by speaker U. White lines indicate F2 trajectories.

velar obstruent before nasal-initial words such as *niet* 'not,' *nooit* 'never,' and *nog* 'yet' (e.g., Figure 5.2, right panel, which illustrates an assimilated final stop in a monosyllabic token of *eigenlijk*, in the set phrase *ik weet het eigenlijk niet* 'I actually don't know'; Figure 5.9; Figure 5.14) and we also observed voicing before /ʊ/ in one case. Typically for tokens of this kind, the stop is weak and very short, as little as 20 ms, relative to a vowel lasting 110–160 ms.

Furthermore, both /x/ and /k/ were sometimes voiced when followed by a vowel, whereas intervocalic voice assimilation at prosodic word boundaries is assumed to be restricted to fricatives (Booij 1995: 147). Finally, a handful of tokens of *eigenlijk* were followed by devoiced nasal stops, probably resulting from progressive voice assimilation induced by /k/ (e.g., Figure 5.15). All in all, the tokens of *eigenlijk* in our data set show more voice assimilation than would be expected on the basis of the existing literature.

5.4.3 Individual speaker differences

Although the speakers form a homogeneous group (they are all adult speakers coming from the same region and from the same socioeconomic class), they clearly show individual differences. Speakers vary in their propensity to produce

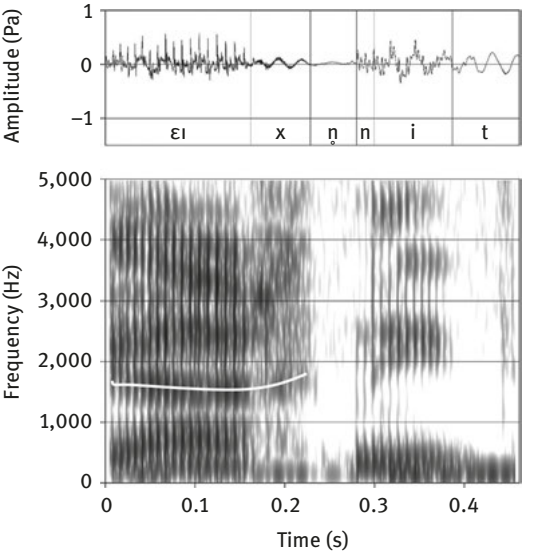

Figure 5.15: Token of *eigenlijk* produced as [ɛɪx] by speaker O in the context *eigenlijk niet* 'actually not.' The token is followed by a partially devoiced [n] which has assimilated in voice to the final segment of *eigenlijk*. The white line indicates the F2 trajectory.

tokens with different numbers of syllables. This is illustrated in Table 5.1, which lists for each speaker the percentage of monosyllabic tokens. Speaker A produced only monosyllabic tokens, speakers F and L produced no monosyllabic tokens, and all other speakers produced a mixture of monosyllabic and di-/trisyllabic tokens. There is also clear variation in the frequency of trisyllabic tokens: half of these were produced by only three speakers (E, J, and S). Part of this variation may be accounted for by individual differences in speech rate (see the next section).

In addition, there may be a role for the position of the word in the prosodic phrase. The speakers varied in terms of where in the prosodic phrase their tokens of *eigenlijk* tended to occur. Speakers E, F, and J produced over half of their tokens at phrase edges, whereas all others produced the majority of their tokens phrase-medially. This variation in prosodic position probably partly explains the interspeaker variation in reduction degree because word tokens at prosodic boundaries tend to be less reduced (e.g., Bell et al. 2003).

At the level of phonetic detail, while most of the patterns observed were common to more than one speaker, there were certain production strategies that appeared to be specific to one or just a few speakers. For instance, speakers F and L were the only ones in the data set who produced overlapping velar frication and /l/-quality. They contrast with speaker N, for instance, who often devoiced the /l/. Furthermore, impressionistically, some speakers were very variable in their realizations of *eigenlijk* whereas others were more consistent. Further research has to investigate to what extent physiological differences between the speakers may explain these individual reduction patterns.

5.5 Analysis of the conditions under which some properties appear

5.5.1 Predictors

We tested whether several properties of the tokens may be conditioned by the following five types of predictors. First, we investigated the influence of the predictability of the preceding and following word, since both have been shown to correlate with the duration and the number of segments of *eigenlijk* (Pluymaekers, Ernestus, and Baayen 2005b). We defined these predictabilities as the logarithms of the numbers of occurrences of these words plus one in the Spoken Dutch Corpus (ranges for both words: 0–12.16).

Second, we studied the influence of a temporal measure. As mentioned in the Introduction, many studies have shown that a higher speech rate generally favors a higher reduction degree. Pluymaekers, Ernestus, and Baayen (2005b) showed that this also holds for *eigenlijk*. We defined speech rate as the logarithm of the number of syllables per second in the citation forms of the words in the labeled phrase (mean: 7.7 syllables/second; range: 1.3–20.5 syllables/second). We used the number of syllables of the citation forms because it is an important predictor of perceived rate (Koreman 2006) though the realized number of syllables also plays a role. We applied a logarithmic transformation because an increase of one syllable per second is likely to have a bigger impact if the rate is one syllable per second than if it is seven syllables per second.

Third, we included prosodic measures: the level of accent on *eigenlijk* and its position in the phrase. We started with regression models in which both predictors had four levels (primary accented, secondary accented, stressed, and unstressed; isolation, phrase-initial position, phrase-final position, phrase-medial position), but models with these predictors did not converge. We therefore simplified these predictors to two-level predictors (accented or not; in phrase-medial position or not).

Further, we included as prosodic measures the number of unstressed syllables preceding *eigenlijk* and the number of unstressed syllables following *eigenlijk* (both ranges: 0–3). They provide information about the rhythm of the phrase. We compared these two continuous measures with two predictors that merely indicate whether the preceding/following syllable is unstressed. The tables with the statistical results show the models with these categorical variables. If these models differ significantly from the models with the continuous variables, this is mentioned in the text.

Fourthly, we investigated the roles of the types of preceding and following segments, since through coarticulation they may directly affect the realization of

the neighboring segments. These variables, however, never played statistically significant roles.

Finally, we also tested whether the number of syllables could predict the other properties of the tokens. Because we feel uncomfortable modeling one dependent variable with another one, we only report the results of these analyses in the text (i.e., without details in tables). For the same reasons, we did not incorporate in the main analysis vowel duration as a predictor for creaky voice.

5.5.2 Statistical analyses

We analyzed the continuous dependent variables with linear mixed effects modeling, and the Boolean dependent variables with generalized linear mixed effects modeling with the logit link function, as implemented in the statistical package R (R Core team, 2014). We tested for random intercepts for speaker, preceding word, and following word. We did not include random slopes because preliminary testing suggested that models including them did not converge or seemed to overfit the data.

Our final models, reported below, only include statistically significant predictors. Fixed predictors were considered significant if their absolute *t*-values or *z*-values were greater than 1.96 (which approximates an alpha level of 0.05). Random intercepts were considered significant if the model with the random intercept outperformed the model without that random intercept as indicated by likelihood ratio tests (again we adopted an alpha level of 0.05). For continuous dependent variables, the final models are only based on those data points that differed less than 2.5 standard deviations from the values predicted by the model.

5.5.3 Results

We first studied which variables predict the number of syllables of *eigenlijk*, by analyzing two dependent Boolean variables: whether the word contains one syllable or whether the word contains three syllables. Trisyllabic tokens are relatively rare in the data set (only 20 tokens) and it is therefore not surprising that there are fewer predictors of the occurrence of trisyllabic than of monosyllabic tokens (see Table 5.3).

We found that monosyllabic tokens are more likely at a higher speech rate and when the preceding word is of a higher frequency of occurrence. In addition, polysyllabic tokens are followed by stressed syllables in 71% of cases, whereas monosyllabic tokens are approximately equally often followed by unstressed and

Table 5.3: Statistical results for the number of syllables. A positive coefficient implies that the predictor increases the probability of one/three syllable(s).

Fixed effects	**Exactly one syllable**		**Exactly three syllables**	
	Coefficient	***z*-value**	**Coefficient**	***z*-value**
Intercept	−11.963	−3.48	3.597	1.81
Speech rate	4.192	2.81	−2.524	−2.53
Preceding word frequency	0.358	2.37	–	–
No following unstressed syllables	−1.996	−2.00	–	–
Phrase-medial position	–	–	−2.311	−3.49
Random effect	**Variance**	**SD**	**Variance**	**SD**
Speaker	6.965	2.64	1.712	1.31
Following word	4.646	2.16	–	–

stressed syllables (51% versus 49%). This suggests a relatively strong tendency for *eigenlijk* to be monosyllabic when followed by an unstressed syllable. The categorical variable, which only provides information about whether the next syllable is stressed, results in a model that is as good as the model with the continuous variable indicating the exact number of following unstressed syllables (the two variables result in models with similar Akaike information criterion values: 152 versus 151, Akaike 1974).

Trisyllabic tokens mostly occurred at lower speech rates and at phrase boundaries. Out of the 20 trisyllabic tokens, only five occurred in phrase-medial position (20%), whereas no fewer than 105 out of the 139 mono- and disyllabic tokens (76%) occurred phrase-medially.

As expected, given our previous observations on differences between speakers (see Section 4.3), speaker is an important random effect. In addition, the probability of a monosyllabic token was conditioned by the identity of the following word. We also found random effects of speaker and following word in the analyses presented below, and we will no longer mention them separately.

We then analyzed the duration of the whole word token, and of the most frequently occurring parts. Table 5.4 shows the statistical results for the duration of the token as a whole and that of its first, full, vowel. As expected, both units are shorter at a higher speech rate and if non-accented (mean duration of non-accented tokens: 267 ms; mean duration of accented tokens: 294 ms). The complete token is also shorter when in phrase-medial position (mean duration: 255 ms) than at phrase boundaries (mean duration: 337 ms).

The duration of the whole token is in addition affected by the rhythm of the phrase: tokens tend to be shorter (mean of 259 ms versus a mean of 293 ms) if they

Table 5.4: Statistical results for word and full vowel durations.

Predictor	Word duration		Vowel duration	
	Coefficient	*t*-value	Coefficient	*t*-value
Intercept	525.115	12.68	229.666	10.47
Speech rate	−112.823	−5.66	−41.141	−3.89
Phrase-medial position	−35.455	−2.86	–	–
Non-accented	−24.350	−2.67	−18.650	−3.79
No following unstressed syllable	24.603	2.37	–	–
Random effect	**Variance**	**SD**	**Variance**	**SD**
Speaker	512.9	22.65	109.7	10.48
Following word	1,606.0	40.08	–	–

are followed by (any number of) unstressed syllables. This result is in line with the results for the probability of a monosyllabic variant, presented above. This effect of the rhythm of the phrase only surfaces in the analysis of token duration if we test the categorical variable. A variable that exactly indicates how many unstressed syllables are following is not predictive of the duration of the token, and results in a model with a slightly higher Akaike information criterion value (1777 versus 1775).

The number of syllables is a good predictor for the durations of both the complete token and the full vowel. The statistical results for vowel duration hardly change if the number of syllables is incorporated as an additional predictor. This is different for the token duration. The effect of the rhythm of the phrase on the duration of the token is no longer significant, which is not surprising because the number of syllables of *eigenlijk* and the rhythm of the phrase are correlated (see above). In addition, the duration of the token is no longer predicted by whether the token is accented or not. This may be more surprising because accentedness does not predict number of syllables in our analyses presented above. Possibly, these analyses were not sufficiently sensitive since we investigated two Boolean dependent variables (monosyllabic or not; trisyllabic or not).

Table 5.5 shows the statistical results for the duration of the velar fricative, if it was present (133 tokens). The fricative was shorter at higher speech rates and when in phrase-medial position (mean duration: 61 ms) rather than at a phrase boundary (mean duration: 68 ms). These variables also predicted token or vowel durations, and in the same directions (see above). The token's number of syllables does not predict the duration of the fricative.

Table 5.5: Statistical results for the durations of the velar fricative and the suffix /lək/.

Predictor	Fricative duration		Suffix duration	
	Coefficient	*t*-value	Coefficient	*t*-value
Intercept	100.916	8.18	192.672	7.98
Speech rate	−17.372	−2.86	−33.653	−2.83
Phrase-medial position	−7.867	−2.34	−39.728	−5.18
Random effect	**Variance**	**SD**	**Variance**	**SD**
Speaker	18.01	4.24	159.2	12.62
Following word	–	–	474.6	21.79

If we test the effect of the rhythm of the phrase on the duration of the fricative with the categorical variables, we found no effects (as indicated in Table 5.5). We find an effect, in contrast, if we test the continuous variable indicating the exact number of following unstressed syllables (coefficient: 6.705; *t*-value: 2.58). Surprisingly, the fricative tends to be longer if *eigenlijk* is followed by a higher number of following unstressed syllables. The absence of an effect of the categorical variable and the unexpected direction of the effect of the continuous variable (which is also opposite to what we found for the probability of a monosyllabic form and for the duration of the complete token) raises the question whether this effect on the fricative duration may be a Type 1 error or just arises because the number of following unstressed syllables happens to be correlated with some other relevant predictor.

Each of the other segments (i.e., the segments in the suffix /lək/) was too often absent to allow for an analysis of its duration. We therefore analyzed the duration of the suffix as a whole, in the 140 tokens in which at least one of its segments was present and in which the final segment was not an unreleased stop followed by silence (see Section 5.3). The results are presented in the last two columns of Table 5.5. The suffix is shorter at higher speech rates and if in phrase-medial position rather than at phrase boundaries (mean durations: 83 ms versus 135 ms). The effect of speech rate disappears if the number of syllables is taken into account. As expected, the suffix is longer if the token contains more syllables and the suffix thus does not only consist of a velar consonant.

From all segments, only /l/ was both present in many tokens (88) and absent in many tokens (71). This was therefore the only segment for which we could analyze which variables predicted its presence. The results are presented in Table 5.6. The segment /l/ was less often present at higher speech rates and when located in phrase-medial position (47% of tokens) rather than at a phrase

Table 5.6: Statistical results for the presence of /l/ and of creaky voice.

Predictor	/l/		Creaky voice	
	Coefficient	z-value	Coefficient	z-value
Intercept	6.1744	−2.76	3.098	1.87
Speech rate	−2.6348	−2.49	−1.715	−2.16
Phrase-medial position	−1.8897	−3.29	−0.920	−2.10
No following unstressed syllable	1.1174	2.22		
Following word frequency	–	–	0.128	2.23
Random effect	**Variance**	**SD**	**Variance**	**SD**
Speaker	2.746	1.66	0.968	0.98

boundary (84%). Moreover, /l/ was more often realized if the *eigenlijk* token was not followed by unstressed syllables (63%) than if one or more unstressed syllables followed (39%). This effect of rhythm only surfaces if we test the categorical rather than the continuous variable indicating the exact number of following unstressed syllables.

All effects on the presence of /l/ disappear if the number of syllables in the spoken *eigenlijk* token is taken into account. The strong effect of the number of syllables is unsurprising: /l/ is seldom present in monosyllabic tokens, quite often present in disyllabic tokens, and always present in trisyllabic tokens (see Table 5.2). The number of syllables is predicted by the same variables as the presence of /l/ (see Table 5.3).

Finally, we analyzed creaky voice, a variable that has not been analyzed so far as a measure of degree of reduction. Creaky voice was present in 83 tokens. We modeled the probability that creaky voice was present as well as the duration of creaky voice if present. The results for the presence of creaky voice are presented in the last two columns of Table 5.6. Creaky voice was less often present at higher speech rates and when the token was in phrase-medial position (in 52% of phrase-medial tokens) rather than in phrase-initial or phrase-final position (in 63% of these tokens). Furthermore, there was a correlation with the frequency of the following word: a more predictable following word appears to increase the likelihood of creaky voice.

The presence of creaky voice cannot be predicted by the number of syllables of the token of *eigenlijk*. In contrast, an additional analysis established that the presence of creaky voice can be predicted by the duration of the vowel. (Recall that we did not include vowel duration in the main analysis for creaky voice because we did not want to model one dependent variable with another one; see Section 5.1.) If vowel duration is incorporated as a predictor in the model for the presence of creaky voice, the effect of the following word frequency is still

statistically significant, while those of speech rate and of phrase-medial position are only marginally significant (ps < 0.07).

The duration of creaky voice, if present, could not be predicted by any of our independent variables. We could only find a high correlation with vowel duration (coefficient: 0.228; $t = 2.02$), showing that the longer the vowel, the longer the part with creaky voice tends to be. In addition, we found a random effect of speaker (variance: 281; S.D. = 16.76).

5.6 General Discussion

This chapter has presented a detailed analysis of the properties of 159 tokens of the Dutch discourse marker *eigenlijk*, which occur in the Ernestus Corpus of Spontaneous Dutch (Ernestus 2000). Our aim was to document the wide variation in the pronunciation of the word and to analyze which properties of the context predict this variation. Previous research suggests that the exact meaning of a token has an influence on its detailed phonetic characteristics (Plug 2005). We did not distinguish between the different meanings because this would have resulted in too small a data set for statistical analyses. Moreover, the meaning of the word only seems to favor some pronunciation variants rather than completely excluding others. As such, it would have only been one of our many predictors. It would be worthwhile investigating the role of the word's exact meaning in a larger data set.

The qualitative analysis of the tokens in our data set supported previous studies (e.g., Ernestus 2000) in demonstrating a wide range of variation in the production of the word, ranging from trisyllabic tokens closely resembling the word's citation form, through to phonetically minimal monosyllabic tokens consisting merely of a vowel followed by a single obstruent consonant. The disyllabic forms occur most frequently (52% of tokens), followed by the monosyllabic variants (36%), while trisyllabic forms are relatively rare (13%).

Reduction of *eigenlijk* may be manifest in a number of ways, in the number of syllables that a token contains, or in its duration, or the clarity of its articulation. We expected these properties to correlate with one another. In fact, we found surprisingly little clear correlation between the different indices of reduction. Although the number of syllables in a token did increase along with duration, the durational ranges found in mono-, di-, and trisyllabic tokens clearly overlapped. Moreover, some tokens that had only one syllable nonetheless had clearly articulated and tightly coordinated segments, while some di- and trisyllabic tokens appeared to be articulated rather laxly. These observations underscore that reduction is not a simple or automatic consequence of speaking under time pressure.

We found some support for the concept of "articulatory prosodies" in the sense of nonlinear features of syllables that are no longer tied to specific segmental units (Niebuhr and Kohler 2011). In particular, the /l/ of *eigenlijk* is not always present, but when a definite lateral articulation is absent, acoustic and auditory traces of /l/ often remain. Also, the weak syllables of the word may be absent yet leave an acoustic residue in the signal, for instance, as extra duration of the consonants in monosyllabic tokens produced as /ɛɪxk/. Following Niebuhr and Kohler's (2011) reasoning further, can we specify a "phonetic essence" of the word *eigenlijk*? Apparently, the only essential components are a front, usually diphthongal vowel, and at least one back (velar or uvular) obstruent. Even these essential parts allow for variation: the vowel can lose its diphthongal quality and can become somewhat more backed; the obstruent can be either stop or fricative, and given an appropriate conditioning context, it can lose its voicelessness.

We observed unexpected patterns of voice assimilation within the word and at word boundaries that are not described in the literature. We found both regressive and progressive voice assimilation of /l/ within the word, and of the final voiceless obstruents of *eigenlijk* to following nasal consonants. Local (2003) proposes that different patterns of assimilation occur for function words (e.g., *I'm*) compared to content words (e.g., *lime*). Our data suggest that the same may be true for discourse markers, and perhaps for frequent sequences such as *eigenlijk niet* 'actually not.' Further work is needed to show whether the observed assimilation patterns are indeed specific for *eigenlijk*.

Also, unexpectedly, the data set showed no clear relationship between a token's prosodic status and its reduction in terms of either duration or number of syllables. We observed a surprisingly large number of accented tokens that were produced with only one syllable (eight primary-accented and 16 secondary-accented monosyllabic tokens). This clearly shows that reduction of *eigenlijk* is not a phenomenon restricted to prosodically weak positions. A token may be heavily reduced in duration or number of syllables, yet may still constitute the most prominent word in its local context. Future studies have to show which other (types of) words can also be drastically reduced in prosodically strong positions.

The data set contains tokens from 18 speakers from a rather homogeneous group (all highly educated men raised in the western part of the Netherlands). Nevertheless, there are substantial differences between speakers in their reduction degree. Some speakers are clearly more likely to produce monosyllabic forms of *eigenlijk* than others. Speakers also differ in how they solve the problem of quickly producing a velar/uvular fricative followed by a lateral, for instance by complete coarticulation of the two sounds, resulting in a lateral fricative, or by weakening of the degree of stricture for the fricative. Finally, in our analyses

investigating which variables predict the duration and presence versus absence of segments, the speaker was always a significant random effect.

These individual patterns confirm and extend the results reported by Hanique, Ernestus, and Boves (2015). By means of computational modeling of automatically generated segmental transcriptions of the speech in the entire corpus, these researchers showed that the speakers in our data set can be better distinguished from each other if not only the word types and combinations of word types that they produced are taken into account but also how these speakers reduced phones and combinations of phones. Our study contributes to extending the results obtained by Hanique and colleagues by documenting interspeaker variation in the pronunciation at the segmental level of one entire word (instead of a single phone or a sequence of three phones), which forms part of the information the computer modeling of Hanqiue and colleagues was based on. On top of this, our results show that the speakers differ at the subsegmental level, which was not taken into account by Hanique and colleagues.

In the second part of the chapter, we investigated which variables may predict the number of syllables of a token, the durations of a token and its parts, and the presence of /l/ and creaky voice. In the Introduction to this chapter, we hypothesized that the rhythm of the sentence may have an effect on reduction degree of *eigenlijk*. Speakers of Germanic languages prefer sentences with approximately equally long intervals between stressed syllables. In order to minimize the sequence of unstressed syllables, they may realize tokens of *eigenlijk* followed by unstressed syllables as (stressed) monosyllabic variants. Conversely, in order to avoid stress clashes, they may prefer a di- or trisyllabic variant, ending in one or two unstressed syllables, respectively, when the word token is followed by a word with initial stress. Our data set provides support for this hypothesis. Tokens of *eigenlijk* appear to be more often monosyllabic, to be shorter in duration, and to be less likely to contain /l/s when followed by unstressed syllables. To our knowledge, effects of rhythm on speech reduction have not been documented before.

Interestingly, the categorical variable that just indicates whether the token of *eigenlijk* was followed by either a stressed or an unstressed syllable outperformed a continuous variable indicating the exact number of following unstressed syllables for token duration and for presence of /l/. (For the probability of a monosyllabic form, the categorical and the continuous variables are equally predictive.) This suggests that only the prosodic status of the immediately following syllable is relevant. Possibly, when producing *eigenlijk*, speakers have only taken decisions about the prosodic status of the next syllable. Another possible explanation is that speakers cannot or are not inclined to reduce *eigenlijk* even more when it is followed by more than one unstressed syllable.

Unexpectedly, it is the exact number of following unstressed syllables rather than the presence of a following unstressed syllable that predicts the duration of the fricative. Moreover, this effect goes in the non-hypothesized direction: this fricative was longer if it was followed by more unstressed syllables. We do not know how to explain this unexpected result.

Of the fixed predictors that have been shown before to correlate with reduction degree, two emerged as statistically significant in (nearly) all analyses: tokens were more reduced at higher speech rates and in phrase-medial position. These results replicate previous findings (e.g., Bell et al. 2003; Kohler 1990; Raymond, Dautricourt, and Hume 2006). We found these effects also for the probability of creaky voice, that is, creaky voice was less likely to occur at higher speech rates and phrase-medially.

Two analyses showed effects of the predictability of a neighboring word. The word *eigenlijk* was more likely to be monosyllabic if it was preceded by a more frequent word. This effect is in line with earlier findings by Pluymaekers, Ernestus, and Baayen (2005b) for this same word. In addition, we found an effect of the frequency of occurrence of the following word on the likelihood of creaky voice in the full vowel: creaky voice was more often present when the word token was followed by words of higher frequencies.

These results with respect to creaky voice are interesting. We see that creaky voice is more often absent under the same conditions where segments tend to be absent and shorter (i.e., at higher speech rates and in phrase-medial position). This suggests that the absence of creaky voice results from the same mechanisms that also reduce segments. This hypothesis, however, does not fit with our observation that creaky voice tends to be more often present if the following word is of a higher frequency: creaky voice behaves differently in this respect from segments, which tend to be more, rather than less, reduced before high-frequency words. We propose that these conflicting results are the consequence of the ambiguous character of creaky voice. On the one hand, creaky voice at the start of a vowel-initial word may function as a phonetic cue to prosodic strength (cf. Jongenburger and van Heuven 1991), which tends to be reduced in the same conditions where segments are reduced. On the other hand, separate from this function in marking vowel-initial word onsets, the presence of creaky voice may in some cases result from reduced articulatory effort in voicing (cf. Gobl and Ní Chasaide 2003). In that case, we expect the presence of creaky voice in those contexts where segments tend to be reduced, including before a high-frequency word. Further research is needed to test this hypothesis.

Unlike previous studies (e.g., Bell et al. 2003, 2009; Pluymaekers, Ernestus, and Baayen 2005b), we did not find an effect of the frequency of the following word on the presence versus absence of segments or on phone, affix, or token durations. A possible explanation is that the previous studies did not incorporate following

word as a random variable. This hypothesis is supported by an analysis showing that the likelihood of a monosyllabic token is predicted by the frequency of the following word, if the following word is not incorporated as a random effect as well. The random effect of the following word was significant in half of our analyses.

These observations may give some clues as to what may be stored in the mental lexicon. On the one hand, the high frequency of highly reduced forms and the existence of consistent phonetic patterns in their production encourage the conclusion that multiple variants are stored as separate pronunciation targets. On the other hand, such variants need to retain some link to the word's canonical form in order to account for the presence of articulatory residues in the phonetic detail of the forms produced. Put differently, a reduced token of *eigenlijk*, even if broadly transcribable as [ɛɪk], probably often differs from the same phoneme string produced in the content word *eik* 'oak,' as other authors have demonstrated for word pairs like (reduced) *support* versus *sport* (e.g., Manuel 1991; Manuel et al. 1992). To substantiate these suggestions requires further acoustic analysis – in particular of spectral properties which we could not address in this paper.

Furthermore, our data show that models of speech production have to take the rhythm in the phrase into account when explaining the reduction degree of *eigenlijk*. This strongly suggests that the assumption that the degree of reduction is only determined by how much time a speaker needs to plan and produce the word or the following word is too simplistic: the speaker also appears to be concerned with at least the rhythm of the phrase.

From a perceptual point of view, the data emphasize that the speech comprehension system has to deal with a great deal of variation when it comes to recognizing *eigenlijk*. One aspect that may help the listener is that several core properties of *eigenlijk* are high in acoustic salience: the formant dynamics characteristic of the diphthong /ɛɪ/, the compact mid-frequency spectral prominences that characterize velar and uvular obstruents, and the strident nature of uvular fricatives. These landmarks may guide the listener, enabling the detection of finer details that cue the word's identity, such as traces of /l/. Perceptual experimentation is needed to test the role of the various gross and subtle acoustic characteristics that we have identified.

5.7 Rethinking reduction

The research reported on in this chapter has generated data that may cast doubt on some of the common assumptions about the phenomenon of speech reduction. We showed that the variation in the pronunciation of discourse markers may

be substantial. The unreduced variant may be less common than reduced variants (in our case only 13% of the tokens are unreduced). This raises the question which form of such a word should be considered as canonical: the full form, which is represented in orthography, or the most frequently occurring reduced form.

Furthermore, our data show that, in contrast to what is generally assumed, highly reduced discourse markers can occur in prosodically strong positions. Reduction is therefore not for all words restricted to unaccented positions. The occurrence of highly reduced forms in accented positions underlines that reduced forms of at least some words are not special and can occur without restrictions.

The pronunciation variation displayed by *eigenlijk* is conditioned, among other factors, by the rhythm of the phrase, and shows large differences between speakers. Moreover, a form may be reduced in one aspect, but not in another. This strongly suggests that reduction is not a fully automatic process that arises when speakers are under time pressure. Speakers clearly have a choice whether to reduce and how to reduce, and they make this choice, among others, on the basis of the rhythm of the phrase, while adhering to their own speech habits.

Finally, we found that every pronunciation variant of *eigenlijk* appears to include two landmarks that may be considered to be the main characteristics of the word (the full vowel and a velar/uvular consonant). These landmarks raise questions about speech processing. Are these landmarks indicated in the mental lexicon? How do listeners use these landmarks during word recognition?

We conclude that this corpus study has shown that many aspects of the phenomenon of speech reduction are not yet well understood. We call for more detailed qualitative and quantitative analyses of many tokens of individual words produced in casual speech because these studies substantially extend our knowledge about speech reduction and about speech production and perception.

Acknowledgments: We would like to thank Harald Baayen for facilitating the second author's stay at the Max Planck Institute for Psycholinguistics, and Brechtje Post for her advice on the prosodic analysis of our phrases. Furthermore, we would like to thank Sarah Hawkins, Ellen Aalders, and three anonymous reviewers for their helpful comments on an earlier version of the paper. This research was partly funded by a vici grant from the Netherlands Organization for Scientific Research.

References

Aalders, Ellen & Mirjam Ernestus In preparation. Is there a link between schwa reduction and the realization of following /p, t/ in casual speech?

Akaike, Htrotugu 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19. 716–723.

Aoyagi, Maki 2015. Japanese vowel devoicing and voicing control in the C/D model. *Journal of the Phonetic Society of Japan*. 22–32.

Bell, Alan, Jason. M. Brenier, Michelle Gregory, Cynthia Girand & Dan Jurafsky 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60. 92–111.

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory & Daniel Gidea 2003. Effects of disfluencies, predictability and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113 (2). 1001–1024.

Bergen, Geertje van, Rik van Gijn, Lotte Hogeweg & Sander Lestrade 2011. Discourse marking and the subtle art of mind-reading: The case of Dutch eigenlijk. *Journal of Pragmatics* 43. 3877–3892.

Booij, Geert 1995. *The phonology of Dutch*. Oxford: Clarendon Press.

Cruttenden, Alan 1994. *Gimson's pronunciation of English* (5th edition). London: Edward Arnold.

Davidson, Lisa. 2006. Schwa elision in fast speech: segmental deletion or gestural overlap? *Phonetica* 63. 79–112.

Ernestus, Mirjam 2000. *Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface*. Utrecht: Netherlands National Graduate School of Linguistics.

Ernestus, Mirjam & Natasha Warner 2011. An introduction to reduced pronunciation variants. *Journal of Phonetics* 39. 253–260.

Gahl, Susanne. 2008. "Thyme" and "Time" are not homophones. Word durations in spontaneous speech. *Language* 84. 474–496.

Gobl, Christer & Ailbhe Ní Chasaide 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40. 189–212.

Gahl, Susanne, Yao Yao & Keith Johnson 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66. 789–806.

Gordon, Matthew, Paul Barthmaier & Kathy Sands 2002. A cross-linguistic acoustic study of fricatives. *Journal of the International Phonetic Association* 32. 141–174.

Guy, Gregory R. 1991. Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3. 1–32.

Hanique, Iris, Mirjam Ernestus & Lou Boves 2015. Choice and pronunciation of words: Individual differences within a homogeneous group of speakers. *Corpus Linguistics and Linguistic Theory* 11. 161–185.

Johnson, Keith 2004. Massive reduction in conversational American English. In Kiyoko Yoneyama & Kikuo Maekawa (eds.), Spontaneous speech: Data and analysis. *Proceedings of the 1st Session of the 10th International Symposium*, 29–54. Tokyo: The National International Institute for Japanese Language.

Jongenburger, Willy & Vincent van Heuven 1991. The distribution of (word-initial) glottal stop in Dutch. In Ans van Kemenade & Frank Drijkoningen (eds.), *Linguistics in the Netherlands*, 101–110. Amsterdam: John Benjamins Publishing Company.

Kelly, Michael H. & J. Kathryn Bock 1988. Stress in time. *Journal of Experimental Psychology: Human Perception and Performance* 14. 389–403.

Keune, Karen, Mirjam Ernestus, Roeland van Hout & R. Harald Baayen 2005. Social, geographical, and register variation in Dutch: From written mogelijk to spoken mok. *Corpus Linguistics and Linguistic Theory* 1. 183–223.

Kohler, Klaus J. 1990. Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In William J. Hardcastle & Alain Marchal (eds.), *Speech production and speech modelling*, 21–33. Dordrecht: Kluwer Academic Publishers.

Kohler, Klaus J. 1999. Articulatory prosodies in German reduced speech. *Proceedings of the XIVth International Congress of Phonetic Sciences* 1. 89–92.

Koreman, Jacques 2006. Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *Journal of the Acoustical Society of America* 119. 582–596.

Labov, William L. 2001. *Principles of linguistic change: Social factors*. Oxford: Blackwell Publishers.

Local, John 2003. Variable domains and variable relevance: Interpreting phonetic exponents. *Journal of Phonetics* 31 (3). 321–339.

Manuel, Sharon Y. 1991. Recovery of deleted schwa. *Perilus: Papers from the Symposium on Current Phonetic Research Paradigms for Speech Motor Control*, University of Stockholm Institute of Linguistics. 115–118.

Manuel, Sharon Y. 1995. Speakers nasalize /ð/ after /n/, but listeners still hear /ð/. *Journal of Phonetics* 23 (4). 453–476.

Manuel, Sharon Y., Stefanie Shattuck-Hufnagel, Marie K. Huffman, Kenneth N. Stevens, Rolf Carlson & Sheri Hunnicutt 1992. Studies of vowel and consonant reduction. *Proceedings of the Second International Conference on Spoken Language Processing*. 943–946.

Niebuhr, Oliver & Klaus J. Kohler 2011. Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics* 39 (3). 319–329.

Oostdijk, Nelleke 2000. The Spoken Dutch Corpus project. The ELRA newsletter 5 (2). 4–8.

Phillips, Betty S. (1994). Southern English glide deletion revisited. *American Speech* 69. 15–127.

Plug, Leendert 2005. From words to actions: the phonetics of eigenlijk in two communicative contexts. *Phonetica* 62. 131–145.

Pluymaekers, Mark, Mirjam Ernestus & R. Harald Baayen 2005a. Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America* 118. 2561–2569.

Pluymaekers, Mark, Mirjam Ernestus & R. Harald Baayen 2005b. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62. 146–159.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raymond, William D., Robin Dautricourt & Elizabeth Hume 2006. Word-internal /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change* 18. 55–97.

Son, Rob J. J. H. van & Louis C. W. Pols 1990. Formant frequencies of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America* 88. 1683–1693.

Son, Rob J. J. H. van & Louis C. W. Pols 1992. Formant movements of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America* 92. 121–127.

Strik, Helme, Joost van Doremalen & Catia Cucchiarini 2008. Pronunciation reduction: how it relates to speech style, gender, and age. *Interspeech* 9. 1477–1480.

Torreira, Francisco & Mirjam Ernestus 2011. Vowel elision in casual French: the case of vowel /e/ in the word c'était. *Journal of Phonetics* 39. 50–58.
Warner, Natasha & Benjamin V. Tucker 2011. Phonetic variability of stops and flaps in spontaneous and careful speech. *The Journal of the Acoustic Society of America* 130 (3). 1606–1617.

APPENDIX

Variation discussed in Sections 5.4 and 5.5 for which we can provide the number of tokens. The first two columns provide information about segmental variation, and the second two columns about subsegmental variation.

Segmental variation	**Number of tokens (out of 159)**	**Subsegmental variation**	**Number of tokens (out of 159)**
Initial vowel is monophthongal	6	Creaky voice for first vowel	83
First schwa is absent	139	Voicing of velar stop	25
Velar fricative is absent	22	At least partly devoiced /l/	21
Second schwa is absent	58	At least partly devoiced schwa	7
Velar stop is absent	19		

Jennifer Cole and Stefanie Shattuck-Hufnagel

6 Quantifying phonetic variation: Landmark labelling of imitated utterances

Abstract: Speech is known to be highly variable across speakers and situations, and listeners pay attention to some of this phonetic detail for the rich contextual information it carries. In this chapter we introduce a method for investigating phonetic variation from the dual perspectives of perception and production. We analyse serial imitations of a heard utterance, where the linguistic object to be produced is fixed syntactically, lexically and prosodically, and employ a novel method for quantifying phonetic variation using acoustic landmarks (LMs) (Stevens 2002) as correlates of phonologically contrastive manner features. Imitated utterances produced by ten native speakers of American English resulted in 3,500+ consonant and vowel LMs, which were labelled and compared both to the lexically specified LMs, and to the LMs as produced in the stimulus. We report five main observations from this exploratory study: (1) Phonetic reduction due to variation in LM realization occurs even in the highly constrained imitation task; (2) variation is asymmetric across classes of LMs: Vowel LMs seldom vary, while glide LMs are most vulnerable; (3) certain patterns of LM deletion were very frequent in our data, but no pattern of phonetic variation prevailed over all imitated instances across or within speakers; (4) phonetically reduced forms in the stimulus, identified in terms of LMs, are not reliably imitated; (5) about 20% of lexically predicted LMs are produced with variable outcomes, both within speakers (across repetitions) and across speakers. These findings demonstrate and quantify systematicity in phonetic reduction as measured in terms of LMs. They also reveal that speakers exercise choice in phonetic implementation, deviating both from lexical targets and from the phonetic detail of the heard stimulus. These results hold promise for the use of imitated speech in the study of phonetic variation, and for the use of LMs (and by extension other feature cues) as a phonologically grounded measure of variation in speech production.

Keywords: phonetic reduction, phonetic variation, landmarks, imitation.

Jennifer Cole, University of Illinois at Urbana-Champaign and Northwestern University
Stefanie Shattuck-Hufnagel, Massachusetts Institute of Technology

https://doi.org/10.1515/9783110524178-006

6.1 Introduction

Research over the past few decades provides mounting evidence of systematic and contextually governed phonetic variation in continuous speech. This variation, which is non-contrastive in the lexical sense, arises due to coarticulation with adjacent segments (Cole et al. 2010; Farnetani and Recasens 1997) or to prosodic structure (e.g. Cho 2005; Choi et al. 2005; Cole et al. 2007; Turk and Shattuck-Hufnagel 2007), and there is also variation in the cues to prosodic features themselves (Dilley, Shattuck-Hufnagel, and Ostendorf 1996; Mo 2011; see also Cole 2015). Phonetic variation can take the form of reduction, strengthening or other kinds of pronunciation change, and has long been seen as a driving force in sound change.

Recent studies have shown the ability of language users to hear, learn and make use of these systematic context-driven and speaker-specific patterns. Evidence for this is seen in phenomena such as the facilitation effect of a familiar talker's voice on word recognition (e.g. Goldinger 1998), phonetic convergence between interlocutors (e.g. Pardo 2006; Pardo et al. 2012), and the perceptual "retuning" of phoneme category boundaries based on auditory exposure to acoustically ambiguous stimuli (Norris, McQueen, and Cutler 2003). Studies of perceptual learning further demonstrate that listeners can learn to associate specific patterns of segmental phonetic variation, synthetically created, to the voice of an individual talker (e.g. Allen and Miller 2004; Eisner and McQueen 2005; Kraljic and Samuels 2005, 2007). Kraljic et al. (2008) argue that perceptual learning of this sort may be critical in enabling listeners to differentially accommodate variation that is idiosyncratic to an individual talker and also the more systematic patterns that characterize dialectal variation; Cutler (2008, 2010) argues that such accommodation is accomplished by reference to abstract phonemic segments in the lexicon, and contributes to the efficiency of lexical access.

In light of these findings, a comprehensive model of phonetic variation must not only provide an inventory of the types of variation that can occur and the contexts in which each type is licensed, but it must also take into account how individual speakers attend to, store and make use of variable phonetic patterns. In this study we explore a new methodology that addresses two challenges for developing such a model, the first concerning data collection and the second concerning measurement of phonetic variation.

6.1.1 Data collection through elicited imitation

Contextual factors play a significant role in conditioning phonetic variation, so in order to discover systematic patterns in variation it is desirable to have multiple

instances of the same lexical items (defining the production targets) in the same contexts, and produced by numerous speakers. For this we use an elicitation task of repeated imitation that offers substantial control over the linguistic object produced by the speaker. Prior studies using imitation and the related task of speech shadowing show that speakers converge to the sub-phonemic detail of heard speech, (Goldinger 1998; Pardo 2013), with measurable effects on, e.g. VOT (Voice Onset Time, measured from stop release burst to onset of voicing for the following vowel) and vowel formants (Babel 2012; Neilson 2011; Shockley, Sabadini, and Fowler 2004). There is some evidence to suggest that imitation is limited to phonetic detail that cues phonological contrast (Mitterer and Ernestus 2008). The findings from these studies predict that imitators will reproduce phonetically reduced forms that they hear, especially when the reduction affects cues to phonologically contrastive features.

Alternative methods to elicited imitation are not as suitable for investigating phonetic reduction. In studies that rely on the production of written stimulus materials, it is not easy to control the prosodic structure (which is known to influence surface phonetics); in studies that rely on corpora of spontaneous speech, it is not easy to control the lexical and syntactic content to make direct comparisons across speakers possible. In contrast, the imitation task severely constrains the syntactic, lexical and phonological (i.e. segmental and prosodic) shape of the utterance, and this means that the effects of these factors are relatively consistent across speakers and across imitations by a single speaker (for evidence of consistency in prosodic imitation see Section 6.2.2). Thus, any consistent patterns of phonetic reduction/variation produced in this task can be understood as patterns that are favoured by the language in those contexts. Although a complete inventory of such processes is well beyond the scope of this exploratory study, if results are promising, it will motivate future expansion of the method, to provide a comprehensive inventory of the nature and scope of surface phonetic variation.

6.1.2 Measuring phonetic variation with landmarks

To model systematic patterns of phonetic variation, including reduction, we need a measure that captures variation in the mapping between discrete lexically contrastive units (e.g. phonemes) and continuous-valued acoustic parameters. As observed by Pardo (2013) in her study of phonetic entrainment, it's not an easy task to identify raw acoustic measures that capture both what is heard by listeners and the phonetic adjustments controlled by speakers. We think the solution lies not in raw acoustic measures, but in a measure that is more directly related to units involved in speech processing, for example, phonological units. Here, we explore the use of landmarks (LMs), as proposed by Stevens (2002), as a

quantifiable, acoustic-phonetic metric that captures variation in the realization of cues to phonological manner features at a finer level of detail than the symbolic allophone, however narrowly defined, permits. LMs provide a way to discretize information from acoustic measures as cues to phonologically contrastive manner features.

In the rest of this section we introduce LMs and expand on the reasons why we expect LMs to be appropriate measures of phonetic variation. Definitions of the LMs used in this study, with examples from our speech data, and the methods for LM labelling are presented in Section 6.2.

We use LMs as the units for measuring pronunciation variation, rather than the symbolic allophone, based on the proposal of Stevens (2002) that individual acoustic cues to contrastive features, rather than symbolic allophones, are significant units of representation in human speech processing. Stevens proposed that the first step in the processing of a perceived utterance by a human listener is the detection and identification of LMs, i.e. the abrupt acoustic discontinuities associated with consonant closures and releases, as well as intensity minima and maxima in glides and vowels, respectively. LMs are a particular class of feature cues that signal information about one class of contrastive phonological features (i.e. the articulator-free features, after Halle 1992, which roughly correspond to the manner features). In this framework, LMs (like other feature cues) are not raw acoustic measures, but are derived from acoustic measures; they are acoustic edges or inflection points, i.e. events which require comparison across multiple measurement values. The LMs used in this study, adopted without modification from Stevens' proposal, mark the acoustic expression of the closure and release of consonantal constrictions for plosives, affricates, fricatives and nasals (e.g. stop-closure, stop-release), the energy valley for glides, and the energy peak for vowels. These LMs are further described, with illustrative examples, in Section 6.2.3.

As an illustration, consider the LM representation for the word peak (from our database) in its unreduced (full) form. The lexical specification of this word identifies the phoneme sequence of the unreduced form as /pi:k/. The initial and final consonants are plosives, which have two LMs, one marking the abrupt intensity drop across a range of frequencies corresponding to the onset of the closure interval and the other marking the abrupt intensity spike marking the onset of stop release noise. The vowel has a single LM marking an intensity maximum. Thus, the LM sequence for this word consists of five LMs: stop-closure, stop-release, V, stop-closure, stop-release, which specifies an unreduced CVC (consonant-vowel-consonant) structure. LMs are particularly informative acoustic events for listeners, since they not only signal the identity of or changes in manner (providing an initial estimate of the CV (consonant-vowel) structure of an utterance), but also identify regions that are rich in cues to the voicing and place features, such as formant transitions and release-burst spectra.

Stevens' proposal was originally concerned with individual feature cues in perceptual processing; here we begin to explore the possibility of extending it to the task of speech transcription (and by implication to speech production). By annotating imitations of the target utterances in terms of LMs, we lay the groundwork for testing the hypothesis that individual feature cues are an appropriate vocabulary for capturing patterns of context-driven surface phonetic variation. That is, LMs may constitute a level of description that links the abstract symbolic specification of lexical items (i.e. in terms of features that define phonemic manner categories) to the continuous-valued variation in the speech signal (i.e. in terms of quantitative parameter values for the cues to manner features). We emphasize here that LMs by themselves will not capture all information about phonetic variation, nor are they the only acoustic cues to inform speech processing. Acoustic cues to voicing and place features, and other spectral information not captured by LMs will also be informative, as will the specific parameter values for the cues, but here we restrict our focus to the presence vs. absence of LMs as cues to manner features.

Labelling LMs (and, eventually, acoustic cues to other kinds of phonological features) offers several advantages over positional allophones for capturing the type of systematic context-driven phonetic variation that has become evident in detailed acoustic-phonetic studies of large corpora, and that experimental studies have indicated are under speaker control and attended to by listeners. For example, individual cues to a given feature are sometimes omitted or added independently, leaving other predicted cues to the features of a target sound segment intact, as when a sequence of two stop consonants is produced without the release burst for C1 and without a closure LM for C2, or when a final /t/ is produced with both glottalization and a release burst. Similarly, a speaker may omit the LM cues to a stop coda, but retain the duration cues to its voicing in the duration of the preceding vowel. Segmental transcription requires a binary decision as to whether a segment was included in the surface form of the utterance or not; cue-based labelling permits a more fine-grained annotation which can capture the fact that some cues may remain to the features of an apparently "deleted" segment. Niebuhr and Kohler (2011) have described such phenomena as the "phonetic residue" of apparent segment deletion processes. LMs (and other feature cues) can also capture detailed (and potentially significant) differences among tokens within an allophonic category. For example, the allophonic category "flap" is applied in American English to a wide range of tokens, from a very short-closure /t/ with clear acoustic evidence for a closure and release burst, to a small glide-like dip in amplitude in a voiced region, with or without a small release burst (due to some build-up of pressure behind the incomplete constriction). If we want to determine whether these variations within an

allophonic category are perceptible, learnable and reproducible by language users, it is useful to have a labelling system which captures them. Individual feature cue labelling also permits the capture of temporal asynchronies among feature cues in the signal, as when frication noise for a voiceless fricative begins before the voicing for a preceding vowel ends, or when the velum opens to create a nasal formant for a coda nasal, somewhere in the preceding vowel. Transcription using sequences of symbols, no matter how detailed and narrowly defined, require the annotator to determine where in the signal the acoustic implementation of one symbol ends and the implementation of the following symbol begins; as practitioners of phonetic labelling are only too well aware, this requirement is often impracticable. That is, in many cases the various cues to a feature (or to the segment that the features define) are spread in time, so that they overlap with cues to adjacent segments (as when the duration of a vowel correlates with the [voice] feature of a following coda consonant) or they are limited in time, so that they do not extend throughout the region that a labeller must designate as corresponding to the relevant phonetic symbol (as when vocal fold vibration is limited to just a few pulses at the beginning of the frication noise associated with a voiced fricative). By labelling individual cues, such asynchronies can be captured and studied for their systematicity, with potentially profound implications for the types of acoustic-phonetic information that are represented and controlled by language users.

LMs and other feature cues are also more amenable to fine-grained quantification than allophonic categories are. For example, it may be difficult to use allophonic symbols to specify the sense in which two speakers' voice onset times become more similar during a conversation, since these changes are typically sub-phonemic (Neilson 2011). But in a transcription system based on individual feature cues, quantitative specification of cue values will be natural and precise; to the degree that such transcriptional analyses reveal systematic control by speakers, it will open the door to the development and testing of speech processing models that incorporate representations of individual cues to contrastive features. Finally, LM transcription provides a simple way of quantifying certain aspects of variation: counting the number and type of LMs that are modified from the lexically predicted pattern (or in the case of imitation studies, from the heard stimulus) allows a straightforward comparison between utterances. In this study, we restrict ourselves to labelling LM cues, because this class of feature cues is particularly robust. But if LM-based transcription emerges as a useful tool for capturing some of the systematic phonetic variation produced by speakers in an imitation task, it will serve as the basis for developing an approach to speech analysis that is more robustly and extensively based on individual acoustic cues to phonological features, and their parameter values.

6.1.3 Research questions

In this study we pose a number of specific questions about the LM behaviour of the speakers in our small sample:

Q1: *Variability in LM outcomes*: Does phonetic variation, as measured by LM modification, occur even in the highly constrained imitation task? If so, what type of modification is most common (e.g. deletion, substitution or insertion)?

Q2: *Variability by LM class*: Are different types of LMs, representing different manner classes, differentially likely to be modified?

Q3: *Between-speaker variation*: Are some patterns of LM modification (e.g. in specific words) consistently produced across speakers? If so, what are the phonological environments that most frequently condition variable outcomes?

Q4: *Accuracy in imitation vs. realization of lexical target*: Do speakers differ in the accuracy with which they realize lexically specified LMs? Do they differ in the accuracy of imitation? What happens when the target of imitation differs from the lexically specified target?

Q5: *Within-speaker variation*: Are speakers internally consistent in the way they realize a LM in a given phonological context, producing an individual "phonetic signature" in terms of preferred patterns of phonetic reduction?

We emphasize that this is an initial exploration, undertaken to evaluate the viability of combining imitated elicitation and LM analysis as a measure of certain aspects of variability and reduction. The domain of the study is restricted to 60 utterances by 10 speakers, i.e. 3 imitations per speaker of 2 target utterances, but, as shown below, the resulting 3,502 LM annotations provide a window into the contexts in which phonetic variation occurs, the nature of that variation and the insights that LM annotation can provide into the processes that underlie it. Thus, the results serve as an initial demonstration of the usefulness of LM labelling as a tool for the quantitative comparison of the phonetic similarities and differences between utterances of the same phonemically specified sentences.

A final comment on terminology is in order here. While we are broadly interested in patterns of phonetic variation as measured by LMs, the findings presented below reveal that the most common patterns of variation involve the loss of a lexically predicted LM, i.e. LM reduction. Other patterns show substitutions of lexically predicted LMs, which, like the examples of LM loss, often result in the partial or complete loss of information about the manner class of phonemes specified in the unreduced lexical form of a word. In what follows we use the terms

"reduction" and "variation" interchangeably in referring to variable outcomes in the imitation data. Distinguishing between these terms will necessitate further work measuring the degree to which lexically specified phonological information is recoverable for the listener.

6.2 Methods

6.2.1 Imitation experiment

Stimuli: Target utterances for the imitation task were drawn from the American English Map Task (AEMT) Corpus of task-driven spontaneous speech (Shattuck-Hufnagel and Veilleux 2007). This corpus was collected using the Map Task elicitation method, described in Anderson et al. (1991). In this speech elicitation task, two speakers (one the instruction giver, the other the instruction receiver) are each furnished with a map; the instruction giver's map shows a path through the items pictured on the map, and the instruction giver is asked to guide the instruction receiver through the task of reproducing the path on the receivers map which shows no path. The two maps differ slightly in the geographical items shown, but this fact is initially unknown to the participants, since neither participant can see the other's map; this manipulation introduces just enough complexity into the task so that the two speakers soon become absorbed in solving the problem and begin speaking in a very natural manner. The resulting speech exhibits the kinds of surface phonetic modification of word forms that occurs widely in natural speaking situations, but is otherwise more difficult to elicit in controlled conditions of laboratory recording which afford the opportunity for the highest-quality acoustic recording and pre-specification of target lexical items.

Thirty-two utterances from 4 of the 16 dialogues in the AEMT were selected for the imitation task; all 4 of these dialogues concerned the same pair of maps. Eight utterances from the instruction giver were selected from the middle portion of each dialogue. The extracted utterances were 7–15 words long (average length 11.5 words), and were chosen to minimize disfluent intervals and laughter. Data from imitations of two of the target utterances are presented here:

Utterance 1
Um Kate d'you see the Canadian Paradise?

Utterance 2
Um you're gonna be standing at the peak of the mountain on the Canadian Paradise.

These two were selected to represent a short and long utterance, and had a minimum of lexical substitutions and disfluent imitations relative to some of the other utterances in the full data set. Both utterances begin with *um* ending in a mid-level pitch plateau that marks a fluent continuation into the following phrase. These *um*s were included in the stimulus utterances for the analysis of prosodic imitation, a part of the larger project for which these data were elicited, but which is not reported here. Note that the orthographic rendering of these utterances reflects three contracted elements: *d'you, you're*, and *gonna*. LMs for these items are discussed below (Section 6.2.3.1). These two utterances, unlike most of the others in the data set, have a common word sequence as well, *Canadian Paradise*, which allows us a small opportunity to look at variation for lexical items across sentence contexts.

Participants: The imitated speech analysed here was recorded from 10 female speakers (18–25 years old) recruited from the student body at the University of Illinois, and paid $10 for participating in this study. The restriction to young female participants was intentional, since the speech to be imitated was taken from dialogues between young female speakers of similar age range. All participants were speakers of the Midland dialect area of American English, and reported no history of speech or hearing deficits.

Procedure: Participants were seated in a quiet room where they received brief instructions from the experimenter and provided written consent prior to the start of the experiment. Participants were equipped with a head-mounted cardioid microphone (AKG C520) and headphones. Target utterances were presented to participants in auditory form through the headphones, with no accompanying text presentation. Participants were told they would be reproducing utterances recorded from a dialogue, and the nature of the Map Task was briefly described to provide context for the dialogue excerpts they would be imitating. The experimenter instructed participants to reproduce each utterance by "repeating the words and the way the utterance was said". Participants listened first to an example utterance to get familiarized with the speech materials, and then proceeded to the imitation task. The auditory stimulus was presented three times in succession with a 2-s pause between presentations. Participants were instructed to reproduce the utterance three times in succession immediately following the three auditory presentations, for a total of 96 imitated productions per subject (32 utterances × 3 repetitions).[1] The timing of the repetitions and the speech rate

1 The intention of the instructions was that participants would reproduce not only the lexical and syntactic content of the stimuli, but also the prosody and other pronunciation qualities representative of the speech style, such as speech rate. The word "imitation" was not used in the

were produced by the participant without instruction. Experiment sessions lasted about 30 minutes. Imitated productions were recorded through a head-mounted microphone (AKG-C520) onto a Marantz solid-state digital recorder, and later transferred to computer for processing and analysis.

6.2.2 Prosodic annotation

Impressionistically, the imitated utterances succeeded in reproducing the spontaneous speech style of the stimuli, and were in fact very hard to distinguish from the set of original productions of the Map Task speakers. To evaluate the extent to which the prosody of the imitated utterances was a match to the prosody of the stimulus utterances, an agreement analysis was conducted on prosodic labels assigned to both stimulus and imitated utterances. The stimuli were prosodically labelled for pitch accents and prosodic boundaries using the full ToBI transcription system (Silverman, et al. 1990). Imitated utterances were prosodically labelled for the location of pitch accents and prosodic boundaries (using the labels "A" and "B"), but without annotation of tonal melodies, and treating intonational and intermediate phrase boundaries as alike. A comparison of prosodic labels between the stimulus utterance and the third imitated production was performed. This comparison using the third imitation rather than the first or second was considered to be a more conservative test of prosodic imitation, on the grounds that the auditory record of the stimulus utterance would be more remote in short-term auditory memory, or not present at all, so a match in prosodic features with the imitated utterance should reflect the cognitive representation of those features in the mind of the imitator.

Cohen's kappa scores for pitch accent and boundary were calculated as the agreement metric for a subset of six imitators. This statistic measures observed agreement against expected agreement, taking into account the frequency of each label. Kappa scores range from 0 (no agreement) to 1 (perfect agreement), and the scores for stimulus-imitation agreement are in the range of 0.61–0.71 for the location of pitch accent, and between 0.6 and 0.7 for the location of prosodic phrase boundaries. These represent substantial agreement according to the common interpretation of this statistic. The kappa scores are in the same range as has been reported for trained transcribers doing a ToBI-style "A" and "B" annotation of a

instruction, to avoid the suggestion that participants should attempt to reproduce pitch range or other aspects of the stimulus speaker's voice that reflect physical characteristics of the speaker rather than linguistic or communicative features of the speech.

similar genre of American English spontaneous speech, with kappa scores of 0.75 for accent and ~0.65 for boundaries (Yoon et al. 2004). Further details of the prosodic annotation, agreement analysis and phonetic measures of prosodic similarity are reported in Cole and Shattuck-Hufnagel (2011) and Mixdorff et al. (2012).

6.2.3 Landmark labelling

The acoustic-phonetic labelling scheme employed in this study was designed to capture the ways in which the predicted LMs, as well as the LMs produced by the speakers of the stimulus utterances, were implemented in the productions of the imitators. We define the predicted LMs to be those that derive from the lexically specified phonemes, i.e. the contrastive segmental units of the full, unreduced pronunciation of the word. For the data analysed here, the lexically predicted LMs were identified by the authors (native speakers of American English) based on their understanding of English phonology and familiarity with the words in this sample.

A further comment is in order here regarding the status of unreduced pronunciations. In using the unreduced form as the reference form against which variable, reduced pronunciations are measured, we do not claim that full, unreduced forms are the *only* kinds of representation encoded by language users, or even that they are the forms that are the most likely to be produced in a given context. Frequent patterns of reduction may be encoded, for example the intervocalic flapped /t/ in *butter*, or deletion of the medial unstressed vowel deletion in *fam(i)ly*. But to the extent that the unreduced pronunciation is possible, perhaps associated with certain conditions of speech style or rate (e.g. extremely clear speech), we hold that it has a privileged status as the form which links all potential productions of a word, including both reduced and strengthened forms. Exemplars on their own do not capture the systematic relationships between surface forms, nor do they capture relationships between exemplars that generalize across lexemes. We maintain that the unreduced form must be available and identifiable as such, even in theories that propose a lexicon defined over clusters of phonetically detailed exemplars. As discussed below, our findings lend some support to this view, as a reduced word in the stimulus is sometimes restored to its full, unreduced form in imitation.

To carry out LM labelling, we used criteria that have been developed in a LM labelling project at the Speech Communication Group at MIT (Shattuck-Hufnagel and Veilleux 2007), based on the ideas of Stevens (2002). In this approach, LMs are initially defined in terms of the acoustic characteristics of a segment (consonant or vowel) in its canonical context. In this sense, "canonical" is defined as the form that a LM takes when it occurs in its most definitive context. For consonants,

the canonical context is between two full (in English, stressed) vowels, as for the closure and release LMs for /b/ in /aba/ or for /m/ in /imi/. (We use the terms "closure" and "release" to designate the acoustic outcome of the articulatory events which cause them; this close mapping between acoustic events and their articulatory causes is an important aspect of Stevens' (2002) proposal.) Results of experiments perturbing the acoustic-perceptual consequences of a speaker's articulatory configurations (Villacorta, Perkell, and Guenther 2007) support the view that, despite this close mapping, the targets of speech production are acoustic in nature. Consonantal stops, fricatives and nasals are predicted to have two LMs, i.e. one created at the moment of formation of the oral constriction and one at the oral release, while affricates have three, i.e. one generated at the moment of constriction, one at the partial release of closure into a configuration that produces frication noise and one at the final release of that constriction. Examples are illustrated in Figure 6.1, panels a–c. In contrast, canonical vowel segments are produced with just one LM, which represents the acoustic consequences of the maximum opening of the vocal tract, i.e. when the vocal tract cross-sectional area is greatest (example in Figure 6.1, panel d). Canonical intervocalic glides are produced with a single minimum opening occurring when the vocal tract is the most constricted, i.e. has the smallest cross-sectional area, and the glide LM marks the valley of the corresponding dip in acoustic energy (example in Figure 6.1, panel e). The string of predicted LMs for an utterance is derived from the string of phonemes that define each word in the lexicon.

6.2.3.1 Predicted and observed LMs for the stimulus utterances

Tables A1 and A2 in the appendix display the phonemes, the predicted LMs for the full, unreduced form for each word in the two target utterances, and the LMs and prosodic features that were realized in the utterances as they were produced by the Map Task speaker and labelled by the authors. As already noted, the orthographic rendering of these utterances reflects three contracted elements: *d'you*, *you're* and *gonna*. These contractions exist in the language as reductions from full forms (*do you*, *you are* and *going to*), but we allow the possibility that the reduced forms are the lexical targets for contractions such as these that have a conventionalized spelling. Thus, we establish the lexically predicted LMs for these items based on the contracted forms, not the corresponding full forms.

There are 40 predicted LMs in Utterance 1, and 82 in Utterance 2, making a total of 122 predicted LMs (see Tables A1 and A2 in appendix). These LMs comprise the lexically specified targets for the imitation task, and are predicted to occur in any clearly produced instance of the words in these utterances. Of the 122 target LMs, six LMs in Utterance 2 were excluded from the analysis of imitated

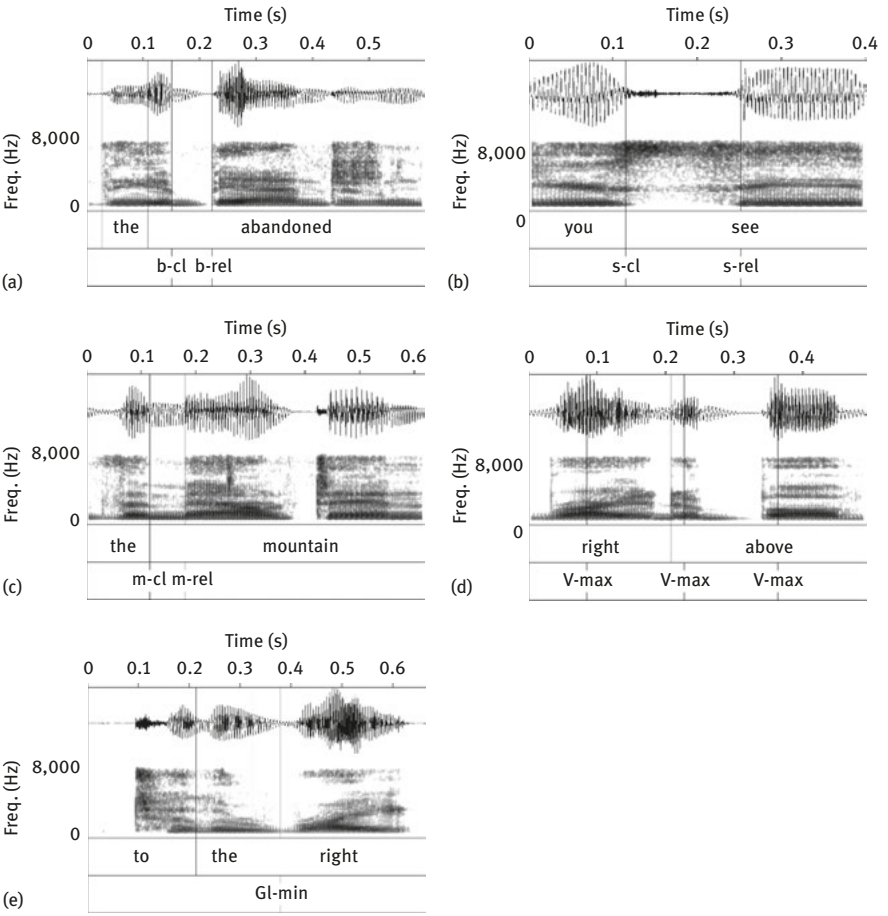

Figure 6.1: Examples of landmarks (LMs) as labeled in stimulus utterances. (a) Stop closure and release LMs for /b/ (spkr 1 utt 4). (b) Fricative closure and release LMs for /s/ (spkr 1 utt 1). (c) Nasal closure and release LMs for /m/ (spkr 1 utt 2). (d) Vowel LMs at amplitude maxima for three vowels in *right above* (spkr 1, utt 3). (e) Glide LM at amplitude minimum for /r/ (spkr 1 utt 4).

productions reported below. The excluded LMs are from the prepositions *at* and *on*, which were frequently subject to lexical substitution in the imitated productions. Thus, where the stimulus contained "... *at the peak ... on the Canadian...*," imitators frequently swapped the prepositions or used the same preposition twice, e.g. "*... on the peak ... at the Canadian ...*". These lexical errors were frequent and variable across speakers, but occurred in otherwise fluent imitated productions, suggesting that the lexical target for the imitator may have been different than the word produced by the Map Task speaker. LMs for these variably produced prepositions were removed from all imitated utterances and from the stimulus prior to

Table 6.1: Number of target LMs in each manner class for Utterances and 1 and 2, combined.

Plosive		Fricative		Nasal		Glide	Vowel	Total
Closure	Release	Closure	Release	Closure	Release			
18	18	9	9	12	12	5	33	116

measuring agreement in LM production. A breakdown of the remaining 116 target LMs by manner class is shown in Table 6.1.

6.2.3.2 Categorizing LM outcomes as intact or deviant

The way each predicted LM was implemented in each utterance was labelled by hand, as follows: the LM was either (a) implemented in its canonical form (termed "no change"), (b) merged with the following LM (see below for further discussion of LM merges), (c) modified to a different type of LM, or (d) deleted. In addition, occasionally an unpredicted LM was produced, labelled as (e) inserted. Labelling was done on the basis of visual inspection of the speech waveform and spectrogram, in conjunction with listening. The two authors labelled about 10% of the data from both utterances together to achieve consistency in labelling, and then labelled the remaining data independently, with regular discussion to resolve ambiguous cases.

Figure 6.1 provides illustrative examples of the 8 canonical LM types for American English: stop closure, stop release, fricative closure, fricative release, nasal closure, nasal release, glide and V. LM locations are labelled in the textgrid for each panel. (Affricates, which combine stop closure with fricative closure and release LMs, did not occur in our data sample.) The examples shown here are drawn from the larger corpus of stimulus utterances, including utterances whose analysis is not included in this study, chosen to provide the clearest illustrations of canonical LM realization.

As noted above, four different codes were used to annotate the outcome of each predicted LM.

- **No Change:** When the acoustic characteristics of a predicted LM matched those of the canonical definition described above. No Change is also described as a Match to the prediction.[2]

2 Note that the label No Change refers only to the acoustic properties that define the LM, and does not imply that other acoustic properties predicted by lexically specified features, or other acoustic properties present in the imitation stimulus, are realized intact.

- **Merge:** When two target consonants occurred in sequence with the release of the first C occurring simultaneously with the closure for the next C. For example, in an /st/ cluster, the LM associated with the release or end of the frication noise for the /s/ often coincides with the LM at the closure for the /t/.[3] In this case a single abrupt spectral change is simultaneously signalling the release of one constriction and the formation of another.
- **Substitution:** When the predicted LM was replaced by a different LM, i.e. when the cues in the signal matched those predicted for a different manner category.
- **Deletion:** When the predicted LM was missing altogether, and no substituted LM occurred between the preceding and following predicted LMs.
- **Insertion:** When a non-predicted LM was produced.

No Change and Merged LM outcomes are considered to be *intact* – the LM is produced as expected, given the lexical specification of the unreduced form and taking into account the adjacent context (for Merge). In Merge contexts, such as sequences of stops consonants and/or fricatives, merged LMs are expected to occur even in clear speech. Substitutions, deletions and insertions are considered as *deviant* LM outcomes, where the expected LMs are not realized. Perceptually salient reduction that relates to manner features, or C/V structure more generally, is expected to occur in contexts with deviant LM outcomes, though there may also be deviant outcomes that are transcribed based on evidence from the acoustic signal but which are not perceived.

Examples of Merges, Substitutions and Deletions – the three most common outcomes other than No Change – are illustrated in Figure 6.2. In the **Merge** example of Figure 6.2a, notice the abrupt end of frication noise for /s/ that is simultaneous with the abrupt beginning of silence for /t/ closure. In the **Substitution** example in Figure 6.2b, the abrupt spectral changes of the predicted LMs marking /d/ closure and release are not present and instead there is a gradual valley in intensity resembling a glide, with voicing continuing throughout. But note that not all alveolar stops that would be transcribed as flapped show this pattern of LM substitution, as shown in Figure 6.2c, where closure and

3 The spectral characteristics of an acoustic LM associated with a change in manner can differ substantially, depending on the manner feature of the adjacent phoneme. For example, the spectral characteristics of the release LM for /s/ are quite different if /s/ is followed by a target stop vs. by a target nasal vs. by a target vowel. In fact, it would be difficult to imagine the same LM outcomes for /s/-release across these contexts. In these cases, the existence of a robust acoustic edge can serve as a perceptual cue to both the occurrence and the nature of the change in manner features.

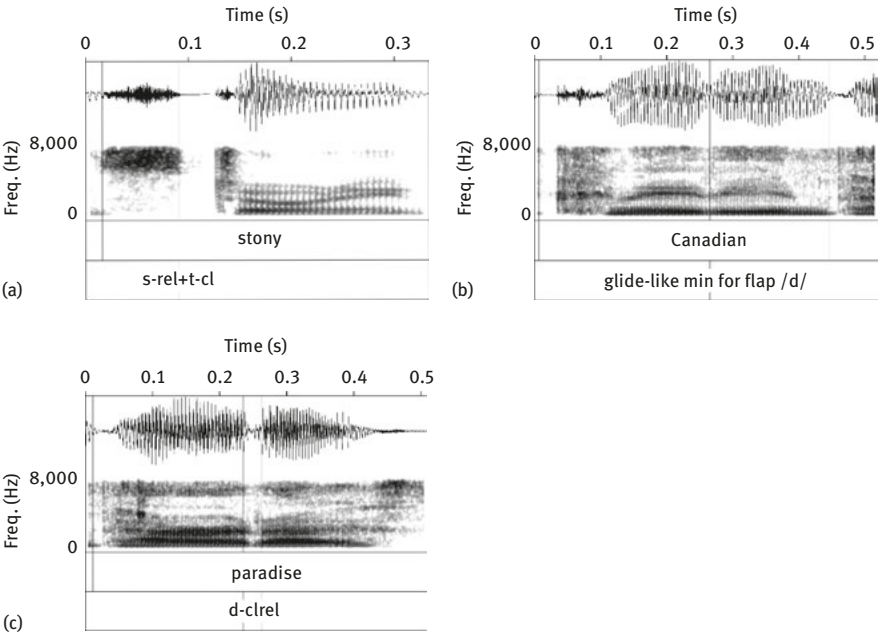

Figure 6.2: Examples of landmarks (LMs) as labeled in stimulus utterances illustrating variable LM outcomes. (a) **Merged** realization of fricative release LM for /s/ and the stop closure LM for /t/ in *stony* (spkr 2 utt 2). (b) A reduced glide-like /d/ in *Canadian* showing **substitution** of the predicted LMs marking stop closure and release with a glide LM marking an intensity valley with voicing continuing throughout (spkr 1 utt 1). (c) A reduced flap-like /d/ in *paradise* showing intact realization of the stop closure and release LMs, with abrupt spectral transitions and closure duration of 22 ms. (spkr 1 utt 1). **Deletion** of expected LMs occurs for /n/ in *stony* (panel a), for the schwa of the first syllable of *Canadian* (panel b), and for the /r/ of *paradise* (panel c), as described in text.

release LMs are observed in a flapped /d/. All LM substitutions in our sample involve realizations of the phoneme /d/ as lenited (i.e. an approximant realization) or flapped. Substitution occurs when the realization involves a change in manner – the stop closure or release are not realized. When the /d/ is fully lenited and manifests as a glide the substitution results in the loss of a LM: this is labelled as substitution of a Glide-Min LM for the expected d-cl LM and deletion of the d-rel LM. In other cases, the /d/ may be partially lenited, manifesting with a glide-like transition at the left or right edge, but with an intact stop closure or release at the opposite edge. Such tokens would be labelled as having one substitution and one unchanged LM, and represent hybrid realizations – part stop and part flap or glide, which would be difficult to capture with a segmental transcription.

Examples of **Deletion** can be seen in these figures as well. In Figure 6.2a the predicted abrupt spectral transitions marking the closure and release LMs for /n/ in *stony* are absent; there are no LMs between the /o/ and /i/ vowels. In Figure 6.2b, the word-initial /k/ of *Canadian* is released directly into the nasalized voiced region for the /n/, with deletion of the predicted V LM in the initial syllable. Finally, in Figure 6.2c, we expect a glide LM for /ɹ/ in *paradise*, but there is no amplitude minimum in the vocalic interval spanning the first two syllables. Instead, the /ɹ/ is realized in rhoticization that extends over a long portion of the vocalic interval.

6.3 Results

6.3.1 LM outcomes: Comparing stimulus to lexically predicted LMs

The first objective of this study is to measure variability in the realization of the target LMs that are predicted from the lexical specification of a word in its unreduced, full form. We begin by evaluating the LM outcomes of the stimulus utterances. There is evidence of phonetic reduction in the stimulus utterances as they were originally produced by the Map Task speaker (see charts of LMs for Utterances 1 and 2 in appendix Tables A1 and A2). In Utterance 1, the speaker produces *Kate* as [kejtˀ] (in familiar allophonic terms, i.e. with an unreleased /t/), which is represented as deletion of the t-closure LM. She also produces *Canadian* as [kʰnejɹɪən], with no vowel LM for the unstressed schwa in the first syllable and with an approximant realization of /d/, which is labelled as deletion of the vowel LM and with a glide LM that substitutes for the predicted d-closure and d-release LMs.

Utterance 2 displays many more variable LM outcomes. One surprising feature of this utterance is the relatively oral-sounding production of the medial /n/ in gonna, transcribed as [gʊdə], which appears on the spectrogram as an oral [d]. This word is produced in the rapid, phrase-initial, unaccented sequence you're gonna be ..... We have no way of knowing if the oralization of /n/ reflects a speech error from an intrusive /d/ target, or if the intended target was a nasal that was ineffectively implemented. Nasal and oral stops share the same LM specification, with closure and release LMs, so the oral realization of /n/ in this utterance is considered to have intact LMs. Looking further into Utterance 2 we observe reduction of the medial /d/ and final /ŋ/ of standing, produced as [stænɪ̃]. The initial /ð/ of *the* is produced after an interval of irregular pitch periods (ipp), which effectively masks the cue to ð-closure LM, although the ð-rel is intact. The /v- ð/ sequence in *of the* exhibits merger of the v-release and the ð-closure, which is an expected

realization for a sequence of two fricatives (or stops). More reduction follows, with a deleted vowel LM for *the*, frication of the beginning of the /k/ in *Canadian* that is marked by substitution of the k-cl LM with x-cl, and subsequent deletions that yield the reduced form [x̂kẽĭrɪm]. The assimilation of the final /n/ of *Canadian* to the labial place of the following /p/ in *Paradise* is not an LM effect, but the expected merger of the n-release and the p-closure is noted. *Paradise* exhibits one more reduction, with a lenited realization of the medial /d/ that it has an intact d-closure but deletion of the d-release.

We turn next to consider the patterns of reduction in imitated productions of these two utterances, where we are especially interested to see if the specific reductions that are present in the stimulus utterances are imitated in the same way.

6.3.2 LM outcomes: Comparing imitations to lexically predicted LMs

6.3.2.1 Frequency of intact and deviant LM outcomes

We turn now to examine the realization of lexically predicted LMs in the imitated productions of Utterances 1 and 2. Recall that we examine all three imitations from each participant, for both of the stimulus utterances. The reader should also bear in mind that the target LMs refer to the lexically derived LMs of the unreduced form of the words in the utterance, which are not always realized as intact in the stimulus utterances themselves, as shown in the preceding subsection.

There are a total of 116 target LMs from the combined stimulus Utterances 1 and 2 (Table 6.1). Each LM was produced 30 times in the imitations (10 speakers × 3 repetitions), and, including 22 inserted LMs (not predicted from lexical specification), there was a total of 3,502 LM *outcomes* in our data. As shown in Table 6.2, the large majority (79%) of these target LMs are realized in their predicted form with no change (NC), or in the form that is predicted from the immediately adjacent phones (Merged). We refer to these as *intact* outcomes. There are also instances of LM substitution and deletion, and a few additional cases of inserted LMs associated with segments not included in the lexical specification of a word, such as sporadic appearance of a full glide LM for a [j] inserted between the last two vowels in *Canadian* [kʰənɛɪdijən]. We refer to LM substitutions, deletions and insertions as *deviant* outcomes, which collectively represent 21% of the total number of LM outcomes produced by our participants.

Among the deviant LMs, the most common outcome is deletion, representing 16% of total outcomes and 77% of the deviations. In comparison, insertion and substitution account for 1% and 4% of LM outcomes, respectively. This finding indicates that LMs are capturing some aspects of the patterns of phonetic

Table 6.2: Classification of produced LMs relative to their predicted form: Intact (No Change or Merged) and Deviant (Deletions, Insertions, Substitutions). Each cell reports the number of LMs produced (outcomes), and in parentheses that number as the proportion of outcomes from 3 repetitions of each utterance by 10 speakers.

	Utterance 1	**Utterance 2**	**Total (Utts. 1–2)**
No change	794 (0.66)	1,477 (0.64)	2,271 (0.65)
Merged	188 (0.16)	306 (0.13)	494 (0.14)
Intact (NC + Merg)	**982 (0.81)**	**1,783 (0.78)**	**2,765 (0.79)**
Substitution	56 (0.05)	95 (0.04)	151 (0.04)
Deletion	164 (0.14)	402 (0.18)	566 (0.16)
Insertion	11 (0.01)	11 (0.00)	20 (0.01)
Deviant (S+D+I)	**229 (0.19)**	**508 (0.22)**	**737 (0.21)**
Total (Intact + Dev.)	1,211	2,291	3,502

reduction in the sense of a production that is reduced with reference to its full form, by virtue of providing fewer cues to signal the presence of a phoneme (cues to syntagmatic structure), or by providing fewer cues to signal contrastive manner features (cues to paradigmatic contrast). In this sense, LMs provide a means to measure certain of the missing components from speech. This merits further analysis of the patterns of deviant LMs in our data.

6.3.2.2 LM outcomes by manner class

Having established that LMs index some patterns of phonetic reduction, we turn to the second objective of this study, which is to determine if all LMs are equally susceptible to variation in production outcome, or if deviant outcomes occur more often for some LMs than for others. In this analysis we compare outcomes based on the manner class of each LM, irrespective of its local (left and right) context, for these manner classes: Plosive, Fricative, Nasal stop, Glide (/r, l, j, w/) and Vowel. As described in Section 6.2, there are distinct LMs marking the closure and release of Plosives, Fricatives and Nasal stops.

Figure 6.3 shows the percentage of LM outcomes that are deviant, for each manner class. These figures reveal a number of interesting asymmetries. The most frequent types of deviant LMs that occur in our small data set are the plosive

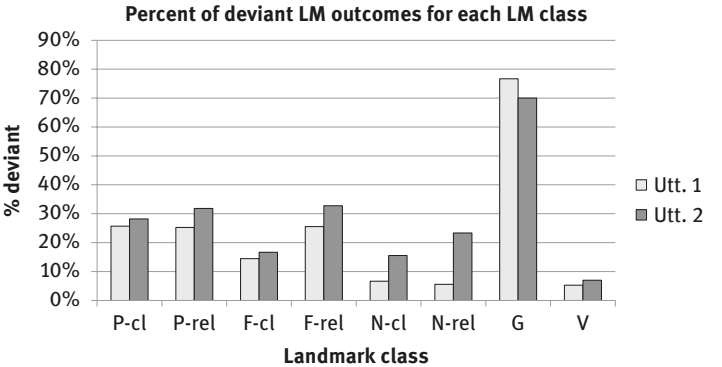

Figure 6.3: Percent of LM outcomes that are deviant relative to the lexically specified target, for LMs grouped by manner class: Plosive, Fricative, Nasal, Glide and Vowel. Closure (-cl) and release (-rel) LMs are coded separately for Plosive, Fricative and Nasal LMs. Utterance 1 LMs shown in light bars, Utterance 2 LMs in dark bars. LMs pooled over all speakers and all repetitions. Percentage values are based on the total number of LM targets for each class that are produced by 10 speakers over 3 repetitions of utterances 1 and 2 (see Table 1).

closure and release LMs, though this is primarily due to the greater number of plosives in this small sample, compared to the other consonantal LMs. When deviant plosive LMs are counted as a proportion of the total number of plosive LMs (Figure 6.3), the deviant outcomes are only slightly more frequent for plosives than for fricatives, with the biggest difference between these two manner classes found for the closure LM. Furthermore, despite occurring with medium frequency relative to plosives and fricatives, nasal LMs are overall less likely to have deviant outcomes, which are especially infrequent in the case of nasal closure LMs. Glide LMs, though less frequent in our sample than other consonant types, are realized as deviant in 70% and 77% of outcomes in Utterances 1 and 2, respectively. The opposite pattern is found for vowel LMs, the most frequent LM type by a large margin, which are almost always realized as intact, with only 5–7% deviant outcomes. A final observation from these findings is that LM outcomes are different for the closure and release components of fricatives and to a lesser degree for nasals and plosives. For these "two-phase" consonants, the release LM is more likely to have a deviant outcome than the closure LM.

6.3.2.3 Contexts for frequent deletion of target LMs

Our LM observations are based on repeated productions of only two utterances, with unequal representation of LMs from the different manner classes, and represent only a fraction of the local contexts in which each LM type may occur,

given the phonotactics of English. We are therefore cautious in drawing generalizations from these data, especially concerning the relative likelihood of deviant LM outcomes for different LM classes. It is useful, though, to examine the specific lexical, prosodic and segmental context for the most common types of deviant LM outcomes in this small corpus as an indication of the contextual factors that condition phonetic reduction. Recall that deletion is the most common type of deviant outcome, with substitutions and insertions occurring much more sporadically. Towards this end, we examined our data to identify the individual consonant LMs in the stimuli that exhibit the most frequent occurrence of deletion outcomes. We qualitatively characterize the contexts with the highest incidence of LM deletion (identified to be 10 or more deleted outcomes out of 30 possible):

a. Consonants in intervocalic position, preceding an unstressed vowel: Here we find frequent deletion of the closure and release LMs for the nasal in /... VŋV .../ in *standing at*, and for the /r/ in /... VrV .../ in *Paradise*. This is also the context for optional flapping of /d/, which occurs often but not always in *Canadian* and *Paradise*. An intervocalic context is also the frequent context for deleted word-initial /g/ LMs in *gonna* following a deleted /r/ LM in *you're* in the phrase *you're gonna*. Here, speakers often produce a very weakly constricted velar approximant with no evident closure or release LM.
b. Consonant clusters: In CC clusters like the /st/ in *standing*, the /nt/ in *mountain*, and the /td/ across a word boundary in *Kate dyou*, there is frequent but not consistent deletion of the release LM of the first consonant with or without deletion of the closure LM of the second consonant. When both LMs are deleted in the same production, the result is a noticeably reduced pronunciation, e.g. when the /st/ cluster is realized as an [s] that releases into a burst characteristic of /t/, but without the prior /t/-closure interval. It's interesting to note that deletion in CC clusters often leaves one LM for each consonant intact, which may support the perceptual identification of both consonants. In our corpus such deletion is most extensive in the word *mountain*, with fully 10 instances out of 30 having deletion of all four LMs for the /nt/ cluster, leaving nasalization of the preceding vowel as the only clear clue to the /n/, and no more than a miniscule burst of irregular glottal pulsing as a cue to the /t/.
c. Schwa vowels in syllables preceding the stressed syllable: This is one of the very few contexts in which a vowel LM is deleted in our sample, and it is the most frequent outcome for the schwa LM in the initial syllable of *Canadian* in both utterances. A similar context is found for the schwa in the final syllable of *mountain*, which is deleted in the 10 (out of 30) productions that have a syllabic /n/ instead.
d. Coda /t/: The final /t/ of *Kate*, which also often occurs before a pausal juncture, is very often entirely deleted in our data, leaving at most a trace of

irregular pitch periods marking the characteristic glottal constriction that accompanies coda /t/.

e. Onset /j/: In 16 of the 30 tokens, the initial glide in *you're* is manifest primarily in the formant transitions into the following vowel, but without evidence of the diminished amplitude that defines the glide LM. In this case, we might consider whether the glide has migrated from the onset to the nuclear position, creating a [ɪɔ] diphthong. Similarly, the glide in the onset cluster /dj/ of *d'you* is present in only one token, with other tokens showing a fricated release of /d/ and a heavily fronted /u/ vowel. The frequency of the deviant LM pattern in this word suggests a stored specification of the reduced variant, though we do note the occurrence of one production that preserves a clear glide LM.

6.3.2.4 Between-speaker variability in LM realization

It is clear even from this limited data set that variation in the production of target LMs is fairly systematic. Across speakers, the frequency of deviant LMs is relatively low, with deletion as the favoured deviant outcome. In addition, deviant LM outcomes are more likely for glide and plosive LMs than for vowel or nasal LMs, and the most common deleted variants tend to occur in specific phonological contexts. Considering these systematicities, we ask if they hold uniformly across speakers.

The distribution of LM outcomes across the Intact and Deviant classes shown in Table 6.2 is representative of the distributions for each speaker, as shown in Figure 6.4, which plots the distribution of LM outcomes over the total number of LMs produced by each speaker. Overall, there are very consistent patterns of LM realization across speakers, in terms of the frequency of intact LMs and the relative frequency of deletion compared to substitutions or insertions in deviant LM outcomes.

The fact that different speakers produce deviant LMs with similar frequency raises the question of whether different speakers are producing deviant LM outcomes for the *same* target LMs in the stimulus. This would be the case if deviant LM outcomes are systematically produced in certain segmental or prosodic contexts. Identifying the contexts where target LMs are systematically produced as deviant is important because it might yield insight into the mechanisms of articulation and speech planning that give rise to phonetic reduction. We examined each target LM in the stimulus for the consistency of outcomes across speakers, counting the number of speakers who produced an intact outcome for that LM in all three repetition of the stimulus (the "high-intact" LMs), as well as the number of speakers who produced no intact outcomes – i.e. for whom every outcome was deviant with reference to the target (the "high-deviant" LMs). These two

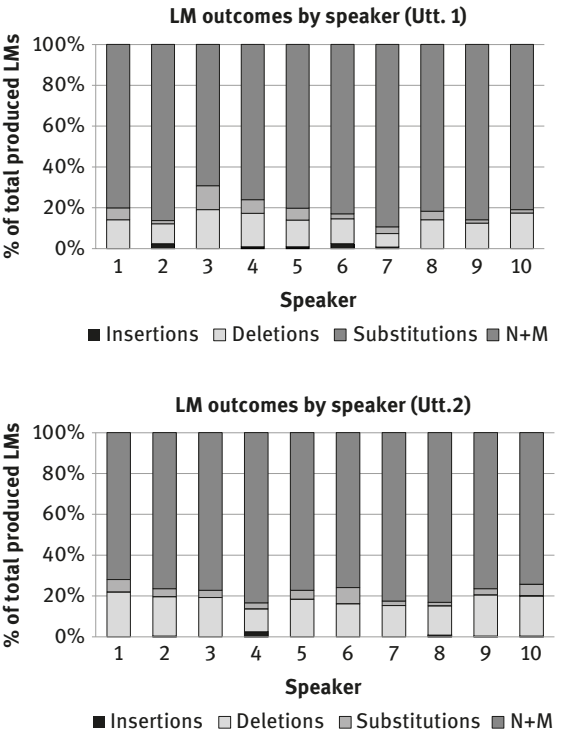

Figure 6.4: Utterance 1 (top) and Utterance 2 (bottom) LMs produced by each speaker (1-10, on the x-axis) and classified by outcome: No Change and Merges (N+M), Substitutions, Deletions, and Insertions (legend below graphs). Bar graph shows each outcome type as a percentage of the total number of LM outcomes produced by each speaker.

categories were taken as the endpoints of a deviance scale with values from 0 to 10, with high-intact LMs assigned a value of zero and high-deviant LMs assigned a value of 10. Intermediate values were assigned to each LM based on the number of speakers (out of 10) who produced one or more deviant outcomes for that LM. If there is consistency among speakers in producing deviant LM outcomes for certain targets, e.g. as determined by LM type and the phonological context of the LM, then we expect to find a lot of LMs in both the high-intact and high-deviant groups. On the other hand, if there are strong individual differences or token-by-token differences in LM realization, we expect to find more variable outcomes, and a higher number of LMs with intermediate values on the deviance scale. This prediction is only partially confirmed.

Figure 6.5 shows the distribution of the LMs (Utterances 1 and 2 pooled) along the deviance scale. While we observe a concentration of LMs in the high-intact category (42 out of 116, or 36%), there are very few LMs in the high-deviant category (4 out of 116, or 3%), and relatively few at intermediate values. This finding indicates

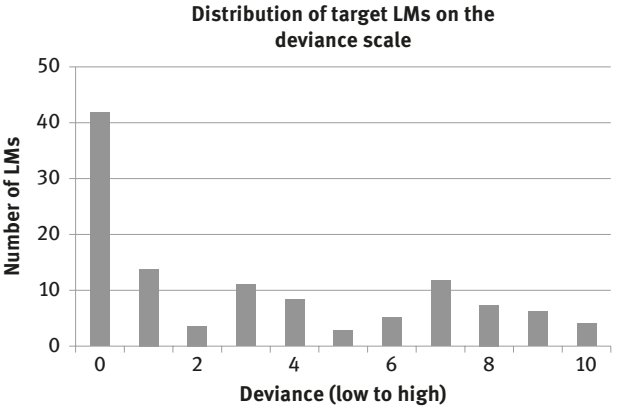

Figure 6.5: The distribution of target LMs from the stimulus utterances 1 and 2 (combined) along the deviance continuum. The high-intact LMs (with low Deviance values, in the leftmost column) are produced as intact in all three repetitions by all 10 speakers. The high-deviant LMs (rightmost column) are produced as deviant in all three repetitions by all 10 speakers. Intermediary values as described in text.

that while many LMs (36%) are consistently produced as intact across speakers, there is less consistency across speakers in the production of deviant LMs.

Extending the "high-deviant" category to include the target LMs with values from 8 to 10 on the deviance scale (i.e. those for which 8 out of 10 speakers produce one or more deviant outcomes), there are 17 out of 116 LMs, or 15%. The deviant outcomes of these LMs represent several well-known types of phonetic reduction in American English: schwa deletion (*C*a*nadian*), lenition of nasal stop closure in NC clusters following a nasalized vowel (*stan*d*ing, mou*n*tain*), deletion of /r/ following an r-coloured vowel (*you'*r*e, pa*r*adise*), lenition of stop or fricative closure intervocalically (*see* t*he, you(re) g*o*nna, go*nn*a*), and loss of oral closure for post-vocalic coda /t/ in the presence of glottal constriction (*Ka*t*e*). These LMs are also among those listed earlier as having the highest occurrence of deletion outcomes over the entire data set (considering all repetitions from all speakers).

6.3.3 LM outcomes: Comparing imitations to stimulus

In the preceding paragraphs we examined variability in the consistency with which a target LM was produced as intact in all three outcomes by the speakers in our study, and we looked into the identities of the LMs with consistently deviant outcomes for all or most speakers. A question arises here as to whether deviant outcomes of the target LMs are in fact produced as a faithful imitation of the stimulus. After all, the original speakers of these utterances from the Map Task corpus themselves produce deviant LMs for some of the LM targets in their speech.

Figure 6.6 illustrates, for each speaker in our study (the imitators), the number of LM outcomes that are deviant with reference to the target LM, and those that are deviant with reference to the stimulus (i.e. where the imitation fails to match the LM outcome in the stimulus). In Utterance 1, LMs differ from the target at about the same frequency as they differ from the stimulus, but in Utterance 2 there are somewhat more LM outcomes that differ from the stimulus compared to those that differ from the target. That means that in Utterance 2, which is the longer utterance, speakers are relatively more reliable in producing LMs as projected from the dictionary specification of each word than they are in accurately imitating the phonetic realization of LMs in the stimulus.

In considering the behaviour of individual speakers across the two utterances, we ask if some speakers are overall more accurate in producing intact outcomes for target LMs, or conversely, if some speakers are more accurate in imitating

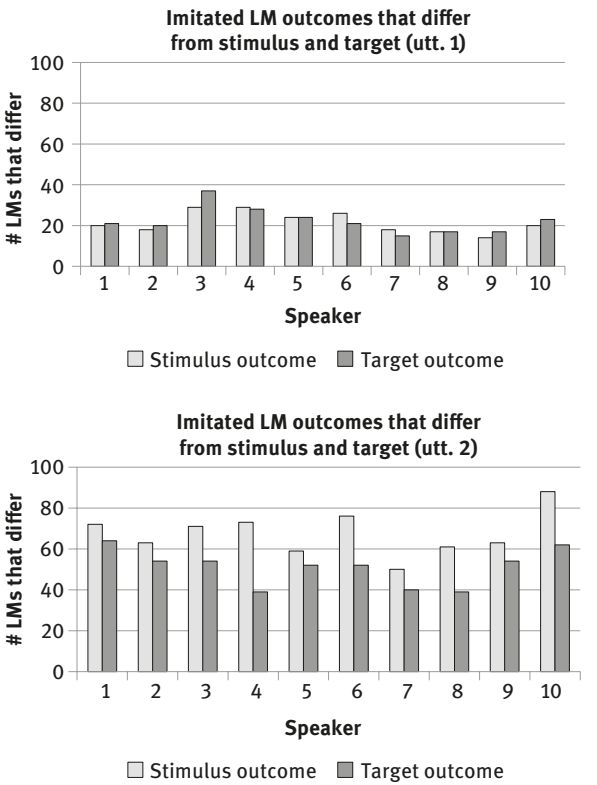

Figure 6.6: The count of LM outcomes in Utterance 1 (top) and Utterance 2 (bottom) that differ from the target LM (dark bars) and the number that differ from the stimulus, grouped by speaker. Data for each speaker include 120 LM outcomes from Utterance 1 (40 target LMs x 3 repetitions) and 228 LM outcomes from Utterance 2 (76 target LMs x 3 repetitions).

LM-related phonetic detail from the stimulus. Comparing the data shown in Figure 6.6, we do not see a relationship between the accuracy in either dimension for these ten speakers. For instance, among all the speakers it is Speaker 3 who produces the greatest number of inaccurate (mismatched) outcomes of target LMs in Utterance 1, but this speaker's productions are not very different from the other speakers for Utterance 2. To be clear, there are individual differences in accuracy among the speakers, both in comparing outcomes to the target and to the stimulus, but it's not clear if the differences reflect individual speaker differences that would generalize across more utterances. An alternative scenario is that the observed differences in accuracy of LM production (or imitation) reflect a range of variation in speech production, with comparable variation within and between speakers.

To determine if speakers are producing deviant LM outcomes (i.e. deviant with respect to the dictionary-predicted LM) in order to more accurately imitate those LMs that are deviant in the stimulus (i.e. that the stimulus speaker produced as deviant), we examine the LM outcomes for the LMs that were deviant in the stimulus utterances. There are 4 deviant LMs in stimulus Utterance 1 and 11 in Utterance 2, making a total of 15 stimulus-deviant outcomes. Considering only the imitations of those 15 stimulus-deviants, we ask how often our speakers produced identical deviant outcomes. Figure 6.7 displays the number of matching vs. mismatching outcomes for these stimulus-deviant LMs as relative proportions, out of 12 for Utterance 1 (4 stimulus-deviant LMs × 3 repetitions) and 33 for Utterance 2 (11 stimulus-deviant LMs x 3 repetitions). The highest rates of matching to stimulus (Speakers 2 and 3, but only for Utterance 1) still fail to match the precise pattern of phonetic reduction in the stimulus about 20% of the time, which is the overall rate of variability in LM outcomes in our data set. It's possible to claim that for the shorter utterance these two speakers have achieved the maximum imitation precision possible for this task. But even these same speakers do not perform as well for the longer utterance, Utterance 2, and all other speakers show imitation match that is far lower than the overall rate of variation for LM outcomes. The relative proportion of matched (to stimulus) vs. mismatched LM outcomes is quite variable across speakers. It appears from this small data set that speakers do not reliably or consistently imitate deviant LMs in the imitation stimulus. In other words, reduced forms are not imitated in the same way as they occur in the stimulus.

6.3.4 Within-speaker variability in LM realization

Although the overall frequency of deviant outcomes, among all 3,502 outcomes in the data set, is moderate, at 21% (see Table 6.2), when we consider the 116 target LMs individually, we observe that fully 74 (64%) are realized with variable

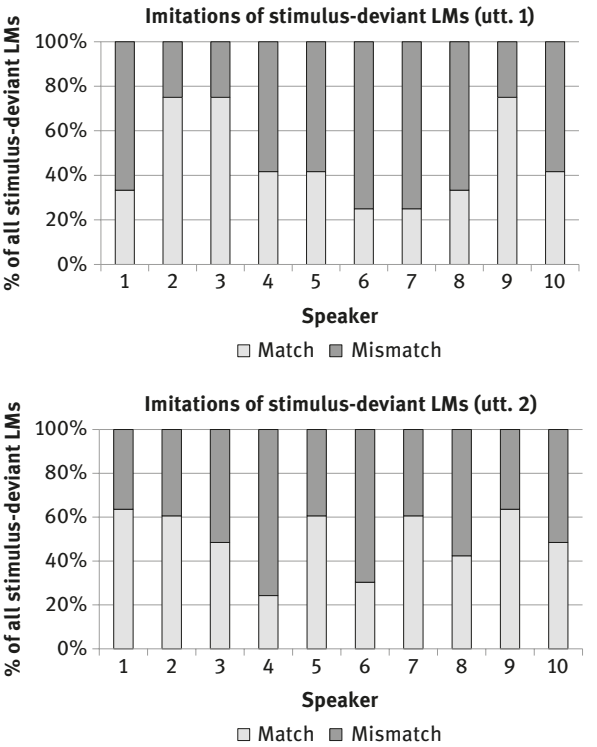

Figure 6.7: Bar graph showing how LMs that have deviant outcomes in the stimulus are imitated by each speaker. Imitations that match stimulus-deviant LMs are shown with light shading, and imitations that do not match stimulus-deviant LMs are shown with dark shading. Breakdown of matching and mismatching LMs are calculated for each speaker as the percentage out of 12 LM outcomes for Utterance 1 in top panel (3 repetitions of each of the 4 stimulus-deviant LMs) and 33 LM outcomes for Utterance 2 in bottom panel (3 repetitions x 11 stimulus-deviant LMs).

outcomes by one or more speakers (these are the LMs with values greater than zero on the deviance continuum, see Figure 6.5). Together these facts point to within-speaker variability in the production of LMs. We assess within-speaker variability to determine the extent to which an individual speaker is consistent in her implementation of LMs. Consistency could result from the faithful realization of lexically predicted LMs, or from the careful imitation of the stimulus, or even from an individual speaker's idiosyncratic patterns of deviant LM production – a kind of phonetic "signature".

Figure 6.8 shows, for each speaker, the relative proportion of target LMs that are realized with the same outcome over all three repetitions, and the number that are realized with two outcomes, out of the 116 target LMs in Utterances 1 and

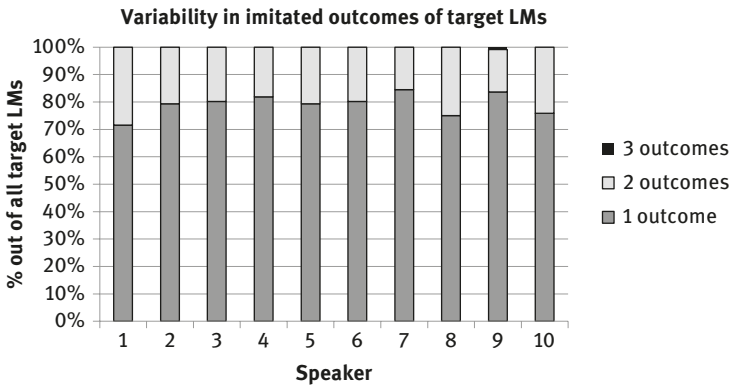

Figure 6.8: Variable outcomes for target LMs as a percent of the total number of target LMs (116), for each speaker's LM outcomes over three repetitions. LMs produced with the same outcome in dark gray bar segment, LMs produced with two outcomes in light gray. LMs pooled from utterances 1 and 2.

2 combined. It is notable that over the entire data set there is only one occurrence of a target LM that is produced in three distinct outcomes by a single speaker, and that was the /p/-closure LM for the /p/ in *paradise*, which Speaker 9 produced once as merged with the preceding word-final nasal of *Canadian*, once as a voiced /b/, and once "deleted" (with no interval of voiceless closure). All other target LMs were produced either with the same outcome, or with two distinct outcomes over three repetitions by a given speaker. Across speakers, the vast majority of target LMs (79%) were produced with a single outcome in all three repetitions by the same speaker. For the other 21% of target LMs, there was within-speaker variation in LM outcomes over three repetitions. Note that this level of within-speaker variation is similar to the overall level of variation in LM production, where we find 21% of the total LM outcomes to differ from the lexically predicted LM (see Table 6.2).

Figures 6.9–6.11 provide examples of how imitated tokens can either reproduce the LMs of the stimulus or modify them, even within a single speaker. Figure 6.9 shows the stimulus word *mountain*, where all of the LMs predicted for the *-ntain* sequence were produced. Figure 6.10 shows three consistent imitations from a single speaker, which are produced with a set of LMs that are different from those in the stimulus. Specifically, the /n/ and /t/ in the medial cluster lack closure and release LMs, and the V LM for the vowel in the second syllable is also absent, with an /n/ closure and release signalling the syllable nucleus. Note that irregular pitch periods are present in each imitation, providing perceptual cues to the obstruent and voiceless features of the LM-less medial /t/, while nasalization of the vowel in the first syllable (not labelled) provides a cue to the medial /n/.

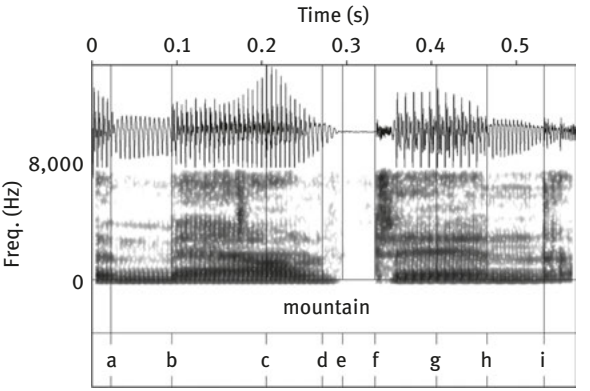

Figure 6.9: The word *mountain* as produced by the MapTask speaker in the stimulus utterance 2. The labels on the second TextGrid row correspond to the following LMs: m-closure (a), m-release (b), V-maximum (c), n-closure (d), n-release merged with t-closure (e), t-release (f), V-maximum (g), n-closure (h), n-release (i).

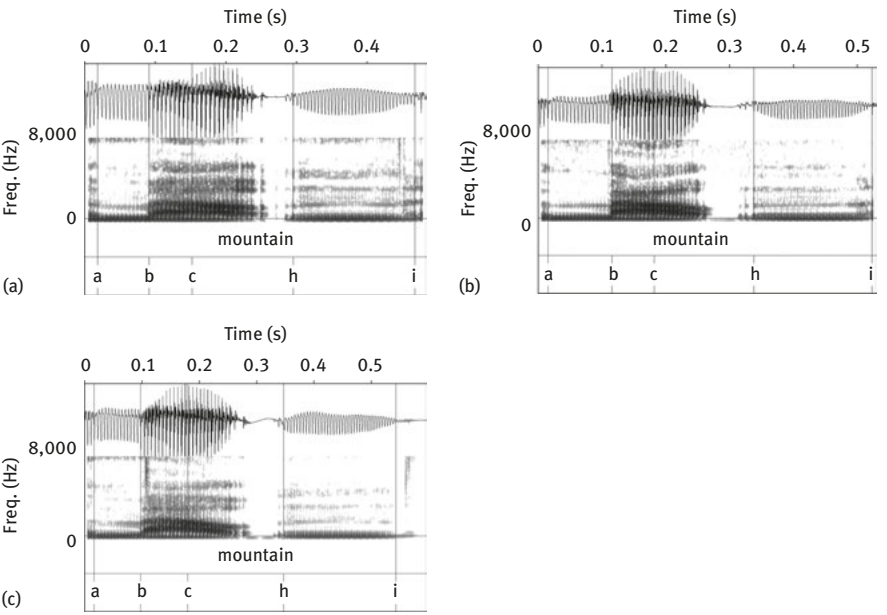

Figure 6.10: The word *mountain* produced as an imitation of the same word in stimulus utterance 2, by speaker 2. Labels on second TextGrid row are the same as corresponding labels in Fig. 9: m-closure (a), m-release (b), V-maximum (c), n-closure (h), n-release (i). LMs corresponding to labels d-g from Fig. 9 are not produced by this speaker. See text for details.

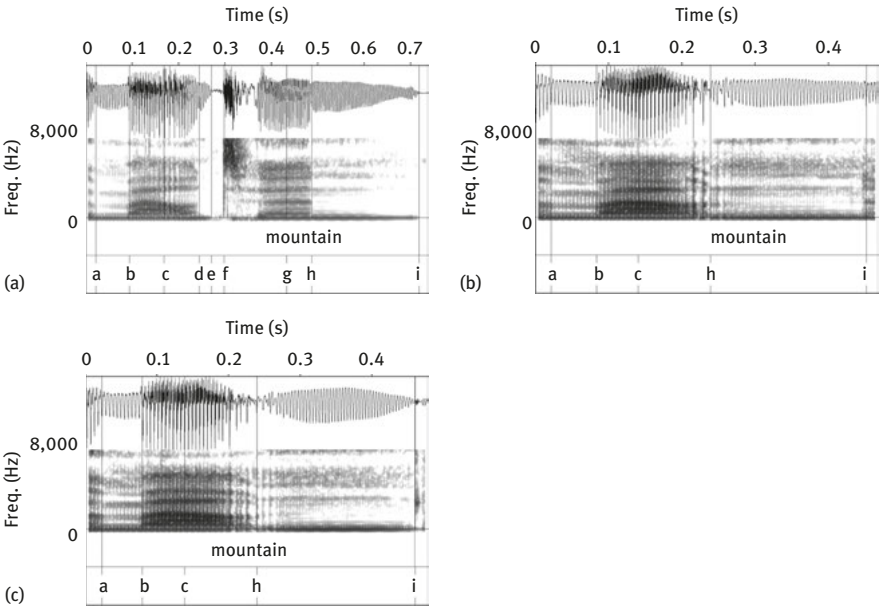

Figure 6.11: The word *mountain* produced as an imitation of the same word in stimulus utterance 2, by speaker 1. Labels on second TextGrid row are the same as corresponding labels in Fig. 9: m-closure (a), m-release (b), V-maximum (c), n-closure (h), n-release (i). LMs corresponding to labels d-g from Fig. 9 are not produced by this speaker. See text for details.

Figure 6.11 illustrates, in contrast, an example of intra-speaker variation. The first of these three imitations produced by a single speaker reproduces the closure and stop cues to the medial /t/ and the vowel LM for the reduced second-syllable vowel that are present in the stimulus shown in Figure 9. But for the second and third imitations, this speaker produces a different set of cues for the medial /t/, including irregular pitch periods, and deletes the V LM for the second-syllable vowel (whose presence may be cued in part by the long duration of the oral closure for the final /n/).

The final observation we report for within-speaker variability is whether there were any effects of repetition order in the pattern of LM outcomes. We compared the distribution of intact and deviant LMs for each repetition. There were no distinct patterns of increase or decrease in deviant LMs across the three repetitions. Indeed, the number of deviant LM outcomes of each type was very similar across the repetitions by a single speaker, for each speaker.

These examples illustrate a significant aspect of surface phonetic variation which emerges even in this highly constrained imitation task: the degree to which speakers can choose among surface forms which differ in their acoustic details but nevertheless provide significant cues to many of the defining features of the

phonemes of their intended words. As Gow (2003) has shown, even highly overlapping articulations (as for the target /t/-/b/ sequence in *right berries*) which result in explicit transcription of a different segment (e.g. a final /p/) often provide enough information in the signal to permit a listener to perceive the target /t/ when tested with a more sensitive online task such as lexical decision. The question of which feature cues are present in a given utterance is thus not always best answered by a transcription in terms of a string of discrete symbolic elements like allophones. The reduced and overlapping cue patterns suggest the need for an approach to transcription that can capture information about individual cues to contrastive features. LM labelling is a first step in that direction, providing a means of capturing the target segments whose presence is robustly signalled in the utterance, vs. those whose presence is less robustly signalled, and opening the way for a more comprehensive labelling system which can capture the additional feature cues that are richly represented in nearby regions of the signal, such as formant transitions that correlate with place features, and vocal fold vibration patterns that correlate with voicing.

6.4 Discussion

We have analysed three serial imitations of two stimulus utterances, from each of ten speakers of American English, to explore patterns of phonetic variation as indexed by LMs – spectral cues to the phonemically contrastive manner features of plosives, fricatives, nasal stops, liquids, glides and vowels. Our broad goal is to identify patterns of phonetic variation, including reduction, that are common to many or all speakers, and to look for effects of phonological context on phonetic variation. We are also interested in establishing the usefulness of imitated speech for investigating phonetic variation, and the adequacy of LMs for indexing and quantifying variable speech outcomes relative to a lexically specified target, or relative to the phonetic details of the heard stimulus. This study reveals interesting findings related to each of these objectives, which are summarized here in relation to the five research questions posed in Section 6.1.

Q1: Variability in LM outcomes: The most general finding is that phonetic variation does occur across imitated productions of a target utterance, as measured by LM modification patterns, even when the lexical, syntactic and phonological contexts for each sound are held constant across imitated productions. This variation could be due to choices an individual speaker makes about other factors influencing variation, such as speech rate or speaking style (careful or casual), to the extent that the imitation does not match the stimulus utterance on these

dimensions. Another possible explanation for variable phonetic outcomes is that to some degree speakers do not control non-contrastive phonetic detail – in other words, some degree of variability is inherent in the speech production process, perhaps reflecting bounds on the precision with which articulator movement is controlled. However, this account is hard to reconcile with other findings in the literature, such as conversational convergence (e.g. Babel 2012; Pardo 2006; Pardo et al. 2012), in which two speakers in a conversation appear to exercise exquisite degrees of control over aspects of acoustic-phonetic variation such as non-contrastive variation in vowel formant trajectories or VOT. The question about sources of variation in imitated speech cannot be fully resolved from analysis of the data presented here, but our findings confirm that imitated speech is useful for the study of phonetic variation. Imitated speech allows for the analysis of the same phonological content, holding constant other features of the linguistic context, so that phonetic outcomes can be quantitatively compared within and across speakers.

Phonetic variation was assessed by labelling LM outcomes as intact (either unchanged or modified in a way that is predicted by the adjacent segments), or as deviant (substituted, deleted or inserted). Out of 3,502 LM outcomes, 79% were intact realizations of target LMs in the stimulus. The most common type of deviant outcome was deletion, accounting for 16% of the total outcomes, and 76% of all deviant outcomes. This finding offers clear evidence that reduction – in the sense of fewer phonetic cues to contrastive phonological units – is the primary source of variation affecting LMs as cues to contrastive manner class features, in this imitation task.

Q2: Variability by LM class: Deviant LM outcomes are not equally probable for LMs from each manner class. Focusing on deleted LM outcomes, we find that glide LMs are the most susceptible to deletion, when deletions are counted in proportion to the total number of glide LMs in the stimulus utterances. Plosive-closure, plosive-release and fricative-release LMs have smaller and roughly equal proportions of deleted outcomes, while fricative-closure, nasal-closure, and nasal-release LMs are even less likely to be deleted. Vowel LMs are almost always realized as intact, with far fewer deviant outcomes than any type of consonant LM. The fact that closure and release LMs differ in their frequency for some manner classes (fricative and nasal stop) suggests an advantage for LMs over segment-sized symbolic features in characterizing patterns of phonetic reduction, and motivate the further pursuit of this comparison in future work.

The very high likelihood of intact outcomes for vowel LMs points to an important difference between consonants and vowels: phonetic variation affecting manner class features, including variation resulting in the reduction of acoustic cues, is much more likely for consonants than for vowels. This C/V asymmetry is further heightened by the fact that glides, which among consonants are the most

similar to vowels by acoustic criteria, show the opposite pattern, with deviant outcomes being the most likely: 74% of glide LMs are deleted in the imitated utterances, while the proportion of deleted vowel LMs is only 6%.

The finding that vowels are much more robust to variation in LM realization than consonants are may be understood in terms of syntagmatic structure and paradigmatic contrast. Consonants from all the manner classes can occupy many of the same positions in syllable and word structure, so the substitution of a consonant LM of one class for that of another class will often result in a phonotactically legal outcome – for example, when the /g/ in *gonna* is realized with a weakened approximant constriction rather than the predicted plosive closure and release, the resulting C/V structure is unchanged. Similarly, in most contexts the deletion of a consonant LM does not create a phonotactic violation, e.g. loss of the coda /t/ in *Kate* results in a legal [CV:] syllable. Turning to vowels, we might expect that a nasal or liquid consonant LM could substitute for a vowel LM, as syllabic consonants, since the substitution is not likely to induce a phonotactic violation. Yet such substitutions are not observed and would be surprising, e.g. if *Kate* were realized as [kn̩t].[4] Clearly, there are factors beyond phonotactic output constraints that must play a role in shaping LM outcomes. On the other hand, phonotactic considerations may help explain the rarity of vowel LM deletions, as there are many contexts in which the loss of a vowel LM would result in a phonotactically illegal consonant cluster. For example, the (hypothetical) loss of the vowel in *Kate* would result in the unsyllabifiable sequence [kt].[5] Another observation that may relate to the robustness of vowels is that there is an apparently parallel finding from studies of elicited speech errors, which have reported that errors on vowels alone, and not in combination with consonants, are very rare relative to errors on consonants (Rusaw and Cole 2011). It is possible that the relative stability of vowels in speech production, compared to consonants, reflects their status as the locus of C/V coordination in speech production

4 Note that the occurrence of a syllabic nasal in a word like mountain (second syllable) is not an example of substitution of a vowel LM for that of a nasal stop, since in this case the nasal is present in the lexical representation, so the reduced pronunciation results from deletion of the vowel LM, leaving the predicted nasal LMs intact.

5 We do observe frequent deletion of an unstressed vowel LM in contexts where the flanking consonants do not form legal onset or coda clusters, e.g. in productions of *Canadian* where the LM for schwa in the first syllable is deleted. The resulting [kn-] sequence is not a legal syllable (or word) onset, but in this case speakers seem to be recruiting the /n/ as a syllable nucleus, filling the position that the deleted V LM would otherwise occupy in syllable structure. In light of examples like this, it may be more appropriate to talk of syllable constraints on LM outcomes, more specifically.

(Browman and Goldstein 1988). Clearly, more research is needed to fully understand the source of vowel stability in speech production.

Q3: Between-speaker variation: This study was designed to elicit many instances of the same LMs from individual speakers, and across speakers, but with the trade-off that we do not have equally representative data from all LM types in all the contexts where they may occur in English. This limits our ability to identify contexts that condition reduction and other variable LM outcomes. Nonetheless, we observe that some LMs are very often realized with deviant outcomes across speakers in this corpus, and without exception these phenomena represent familiar patterns of phonetic variation. Specifically, we observe deletion of consonant LMs in intervocalic position before an unstressed vowel (e.g. *pa<u>r</u>adise*), deletion of the release LM for the initial consonant in a CC cluster (e.g. *mou<u>n</u>tain*), deletion of the vowel LM for schwa (e.g. *C<u>a</u>nadian*), deletion of coda /t/ (e.g. *Ka<u>t</u>e*) and deletion of post-consonantal onset /j/ (e.g. *d'<u>y</u>ou*).

What strikes us as most remarkable about LM deletion in the contexts described above is not that deletion is consistent across speakers, but that it is not universal. Of the 116 target LMs in this study, only one is never realized intact, and that is the vowel LM for the schwa in the first syllable of *C<u>a</u>nadian*. All other target LMs that undergo frequent deletion are realized as intact in one or more imitated productions. A related observation is that the production of a deviant LM outcome is not strongly predicted by the LM pattern in the stimulus that the imitator heard. Even if the stimulus is deviant (with respect to the lexically specified LMs), the imitation does not reliably reproduce the same deviant LM outcome, nor are all the deviant outcomes in the imitation matched to deviant outcomes of lexically specified LMs in the stimulus. The non-uniformity in deviant outcomes tells us that although a particular phonological context may license modification of a lexically specified LM, such a modification is not obligatory, at least not in most cases. Rather, the findings from this study suggest that speakers have a range of options for the phonetic implementation of a lexical item, and the choice among them is not fully determined by the linguistic context.

Q4: Accuracy in imitation vs. realization of lexical target: LM outcomes were compared across speakers to test for differences in the frequency of intact realization of lexically specified (target) LMs, or in the accurate imitation of deviant LMs in the stimulus. We found no evidence for either. In other words, among the ten subjects in this study there are no exceptionally clear speakers, and no superior imitators. Rather, all speakers exhibit very similar overall patterns in the distribution of intact and deviant LM outcomes, along with some variation in the choice of deviant productions.

Q5: Within-speaker variation: Related to the question of whether some speakers are more accurate in producing target LMs, or in imitation, is the question of

whether speakers are internally consistent in speech production, favouring one particular outcome for a given target LM across all repetitions. We reason that internal consistency in the production of phonetic variants would help to establish individual speech patterns that could function to index social identity. The data provide scant evidence that speakers are using LM variation in this manner. Target LMs are produced by the same speaker with a unique outcome and with divergent outcomes in nearly equal proportions.

The finding of moderate within-speaker variation contributes to the overall picture of phonetic variation as an inherent property of speech production, with very similar frequencies of variable outcomes across and within speakers. In the speech sample analysed here, variable LM outcomes occur in about 20% of the outcomes overall (counting all speakers and all repetitions), and in about 20% of an individual speaker's productions (counting all repetitions). The same proportion of target LMs, about 20%, are produced with variable outcomes in one or more imitations. Furthermore, this finding is an exact replication of the finding from Shattuck-Hufnagel and Veilleux (2007), who report 20% deviant outcomes for LM realization on a larger sample of the Map Task corpus, the same corpus from which the stimuli in our study were taken. The remarkable consistency among these measures of variability lends further support to the idea that some degree of variability is inherent to the speech production process, which we can estimate at 20% on the basis of the present speech sample.

6.5 Conclusion

In this exploratory study of phonetic variation in an imitation task that highly constrains the syntactic, prosodic and lexical aspects of the utterance, we have demonstrated that speakers do not always reliably reproduce the target utterance at the level of detail measured by the cues to manner features known as acoustic LMs. The results provide support for the view that imitation is a useful method for eliciting phonetic variation under controlled conditions, and that LM labelling is a useful tool for quantifying certain aspects of the degree and type of phonetic variation. Because the range of types of phonetic variation in a spoken language is very large, affecting many aspects of an utterance beyond the LM cues to manner features, the hand-labelling method we have used in this preliminary study may not be practical for more comprehensive studies, aimed at inventorying the full set of variation patterns and the way these patterns are used in different contexts and by different speakers. However, the results of this study have provided some initial observations of

this type, and have demonstrated the usefulness of this approach to the study of phonetic variation.

Future work can build on the patterns of LM variation reported here through controlled studies with specific phonological targets, or by manipulating prosodic context, speech rate, or other features of the elicitation stimuli. Developments in the direction of automatic detection of LMs and other cues to feature contrasts, applied to existing large corpora of spontaneous speech, will enable the study of the broader principles that govern patterns phonetic variation in spoken language. Future research along these lines would provide a more comprehensive test of the hypothesis that individual feature cues provide a useful vocabulary for annotating phonetic variation. In linking the perception of phonetic reduction with the listener's subsequent speech production behaviour, this line of research may also shed light on the mechanisms of sound change. Towards these goals, it will be instructive to compare the appropriateness of narrow symbolic transcription, raw acoustic measures and listeners' perceptual judgements for quantifying systematic phonetic variation.

References

Allen, J. Sean & Joanne L. Miller 2004. Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America* 115 (6). 3171–3183.

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson & Regina Weinert 1991. The HCRC map task corpus. *Language and Speech* 34 (4). 351–366.

Babel, Molly. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics* 40 (1). 177–189.

Browman, Catherine P. & Louis Goldstein 1988. Some notes on syllable structure in articulatory phonology. *Phonetica* 45.2–4: 140–155.

Choi, J-Y., Hasegawa-Johnson, M. and Cole, J. 2005. Finding intonational boundaries using acoustic cues related to the voice source. *Journal of the Acoustical Society of America*, 118(4): 2579–88.

Choi, Jeung-Yoon & Stefanie Shattuck-Hufnagel 2014. Quantifying surface phonetic variation using acoustic landmarks as feature cues. *The Journal of the Acoustical Society of America* 136 (4). 2174.

Cho, Tae-Hong 2005. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a,i/ in English. *The Journal of the Acoustical Society of America* 117 (6). 3867–3878.

Cole, Jennifer 2015. Prosody in context: A review. *Language, Cognition and Neuroscience* 30 (1–2): 1–31.

Cole, Jennifer., Kim, H., Choi, H. and Mark Hasegawa-Johnson 2007. Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics* 35. 180–209.

Cole, Jennifer, Gary Linebaugh, Cheyenne Munson & Bob McMurray 2010. Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics* 38 (2). 167–184.

Cole, Jennifer & Stefanie Shattuck-Hufnagel 2011. The phonology and phonetics of perceived prosody: What do listeners imitate? *Proceedings of INTERSPEECH-2011*, 969–972.

Cutler, Anne 2008. The abstract representations in speech processing. *Quarterly Journal of Experimental Psychology* 61 (11). 1601–1619. [omit? doi:10.1080/13803390802218542.]

Cutler, Anne 2010. Abstraction-based efficiency in the lexicon. *Laboratory Phonology* 1 (2). 301–318.

Dilley, Laura, Stefanie Shattuck-Hufnagel & Mari Ostendorf 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24 (4). 423–444.

Eisner, Frank, & James M. McQueen 2005. The specificity of perceptual learning in speech processing. *Perception & Psychophysics* 67 (2). 224–238.

Farnetani, Edda, & Daniel Recasens 1997. Coarticulation and connected speech processes. In William J. Hardcastle, John Laver and Fiona E. Gibbon (eds.), *The handbook of phonetic sciences*, 371–404. Chichester: Wiley-Blackwell.

Goldinger, Steven D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105 (2). 251–279.

Gow, David W. 2003. Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics* 65 (4). 575–590.

Halle, Morris. 1992. Phonological features. In W. Bright (Ed.), *International encyclopedia of linguistics*, Vol. 3, 207–212. Oxford: Oxford University Press.

Kraljic, Tanya & Arthur G. Samuel 2005. Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology* 51 (2). 141–178.

Kraljic, Tanya & Arthur G. Samuel 2007. Perceptual adjustments to multiple speakers. *Journal of Memory and Language* 56 (1). 1–15.

Kraljic, Tanya, Arthur G. Samuel & Susan E. Brennan. 2008. First impressions and last resorts how listeners adjust to speaker variability. *Psychological Science* 19 (4). 332–338.

Mitterer, Holger & Miriam Ernestus. 2008. The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition* 109 (1). 168–173.

Mixdorff, Hansjoerg, Jennifer Cole & Stefanie Shattuck-Hufnagel 2012. Prosodic similarity – evidence from an imitation study. *Proceedings of Speech Prosody 6* (Shanghai). 571–574.

Mo, Yoonsook 2011. *Prosody production and perception with conversational speech*. Urbana-Champagne: University of Illinois dissertation.

Niebuhr, Oliver & Klaus J. Kohler 2011. Perception of phonetic detail in the identification of highly reduced words. Journal of Phonetics 39 (3). 319–329

Nielsen, Kuniko 2011. Specificity and abstractness of VOT imitation. *Journal of Phonetics* 39 (2). 132–142.

Norris, Dennis, James M. McQueen & Anne Cutler 2003. Perceptual learning in speech. *Cognitive Psychology* 47 (2). 204–238.

Pardo, Jennifer S. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America* 119 (4). 2382–2393.

Pardo, Jennifer S., Rachel Gibbons, Alexandra Suppes & Robert M. Krauss 2012. Phonetic convergence in college roommates. *Journal of Phonetics* 40 (1). 190–197.

Pardo, Jennifer S., Kelly Jordan, Rolliene Mallari, Caitlin Scanlon & Eva Lewandowski 2013. Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language* 69 (3). 183–195.

Rusaw, Erin & Jennifer Cole 2011. Speech error evidence on the role of the vowel in syllable structure. *Proceedings of the International Congress on Phonetic Sciences*, Hong Kong, 1734–1737.

Shattuck-Hufnagel, Stefanie & Nanette Veilleux 2007. Robustness of acoustic landmarks in spontaneously-spoken American English. *Proceedings of the 16th meeting of the International Congress of Phonetic Sciences, Saarbrueken*, 925–928.

Shockley, Kevin, Laura Sabadini & Carol A. Fowler 2004. Imitation in shadowing words. *Attention, Perception, & Psychophysics* 66 (3). 422–429.

Silverman, Kim E.A. & Janet B. Pierrehumbert 1990. The timing of prenuclear high accents in English. In John Kingston & Mary E. Beckman (eds.), *Papers in laboratory phonology 1: Between the grammar and the physics of speech*, 72–106. Cambridge: Cambridge University Press.

Stevens, Kenneth N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America* 111 (4). 1872–1891.

Turk, Alice E. & Stefanie Shattuck-Hufnagel 2007. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics* 3 (4). 445–472.

Villacorta, Vergilio M., Joseph S. Perkell & Frank H. Guenther 2007. Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America* 122 (4). 2306–2319.

Yoon, Taejin, Sandra Chavarría, Jennifer Cole & Mark Hasegawa-Johnson 2004. Intertranscriber reliability of prosodic labelling on telephone conversation using ToBI. *Proceedings of INTERSPEECH*-2004, Jeju Island, Korea. 2729–2732.

APPENDIX

A1. Utterance 1: *Um, Kate, d'you see the Canadian Paradise?*

Words	**um**			**Kate**					**d'you**				**see**			**the**		
Prosody	H*	H-L%		H*			L-L%		H*				H*					
Phonemes	ʌ	m		k		ej	t		d		j	uw	s		ij	ð		ə
LMs-pred.	V-max	m-cl	m-rel	k-cl	k-rel	V	t-cl	t-rel	d-cl	d-rel	G	V-max	s-cl	s-rel	V-max	ð-cl	ð-rel	V-max
LMs-real.	V-max	m-cl	m-rel	k-cl	k-rel	V	t-cl	DEL	d-cl	d-rel	G	V-max	s-cl	s-rel	V-max	ð-cl	ð-rel	V-max

Words	**Canadian**											
Prosody												
Phonemes	k		ə	n		ej	d		ij	ə	n	
LMs-pred.	k-cl	k-rel	V-max	n-cl	n-rel	V-max	d-cl	d-rel	V-max	V-max	n-cl	n-rel
LMs-real.	k-cl	k-rel	DEL	n-cl	n-rel	V-max	G-min	--	V-max	V-max	n-cl	n-rel

Words	**Paradise**									
Prosody	L+H%									H-
Phonemes	p		æ	ɹ	ə	d		aj	s	
LMs-pred.	p-cl	p-rel	V	G-min	V-max	d-cl	d-rel	V-max	s-cl	s-rel
LMs-real.	p-cl	p-rel	V	G-min	V-max	d-cl	d-rel	V-max	s-cl	s-rel

A2. Utterance 2: *Um you're gonna be standing at the peak of the mountain on the Canadian Paradise*.

Words	**Um**			**you're**			**gonna**						**be**		
Prosody	H*			H-L%											
Phonemes	ʌ	m		j	ow	ɹ	g		ə	n		ə	b		ij
LMs-pred.	V-max	m-cl	m-rel	G-min	V-max	G-min	g-cl	g-rel	V-max	n-cl	n-rel	V-max	b-cl	b-rel	V-max
LMs-real.	V-max	m-cl	m-rel	G-min	V-max	G-min	g-cl	g-rel	V-max	d-cl	d-rel	V-max	b-cl	b-rel	V-max

Words	**Standing**												**at**		
Prosody	!H*										H-				
Phonemes	s		t		æ	n		d		ɪ	ŋ		æ	t	
LMs-pred.	s-cl	s-rel	t-cl	t-rel	V-max	n-cl	n-rel	d-cl	d-rel	V-max	ŋ-cl	ŋ-rel	V-max	t-cl	t-rel
LMs-real.	s-cl	s-rel	t-cl	t-rel	V-max	n-cl	n-rel	DEL	DEL	V-max	DEL	DEL	V-max	t-cl	t-rel

Words	**The**			**peak**					**of**			**the**		
Prosody				L+H*		L-								
Phonemes	ð		ə	p		ij	k		ə	v		ð		ə
LMs-pred.	ð-cl	ð-rel	V-max	p-cl	p-rel	V	k-cl	k-rel	V-max	v-cl	v-rel	ð-cl	ð-rel	V-max
LMs-real.	DEL	ð-rel	V-max	p-cl	p-rel	V	k-cl	k-rel	V-max	v-cl	MERGED		ð-rel	V-max

Words	**of**			**The**			**mountain**										**on**		
Prosody							L*									H-L%			
Phonemes	ə	v		ð		ə	m		aw	n		t		ə	n		a	n	
LMs-pred.	V	v-cl	v	ð-cl	ð-rel	V	m-cl	m-rel	V	n-cl	n	t-cl	t-rel	V	n-cl	n-rel	V-max	n-cl	n-rel
LMs-real.	V	v-cl	MERGED		ð-rel	V	m-cl	m-rel	V	DEL	DEL	t-cl	t-tel	V	n-cl	DEL	V-max	DEL	DEL

Words	**The**			**Canadian**												**P...**
Prosody																
Phonemes	ð		Ax	k		ax	n		ej	d		ij	ax	n		p
LMs-pred.	ð-cl	ð-rel	V-max	k-cl	k-rel	V-max	n-cl	n-rel	V-max	d-cl	d-rel	V-max	V-max	n-cl	n-rel	p-cl
LMs-real.	ð-cl	ð-rel	DEL	SUBS x-cl	k-rel	DEL	DEL	DEL	V-max	G-min	DEL	V-max	DEL	n-cl	MERGED	

Words	**...n**		**Paradise**									
Prosody			L*									H-H%
Phonemes			p		æ	ɹ	ə	d		ay	s	
LMs-pred.	n-rel		p-cl	p-rel	V-max	G-min	V-max	d-cl	d-rel	V-max	s-cl	s-rel
LMs-real.	MERGED			p-rel	V-max	G-min	V-max	d-cl	DEL	V-max	s-cl	s-rel

Tables A1 and A2: Stimulus Utterances 1 (A1) and 2 (A2), used for elicited imitations. The second row displays the prosodic feature (pitch accent and boundary tone) as ToBI-labelled by the authors, and left blank if no accent or boundary label was assigned. The third row displays the phonemes specified for the unreduced form of the word. The fourth row displays the predicted LMs for each phoneme, and the bottom row displays the LMs as realized by the speaker (LM outcomes) and labelled by the authors. Cells for realized LMs (bottom row) that are deleted, substituted or merged relative to the predicted LMs are lightly shaded. Dark shading is used for cells corresponding to the words, at and on, in Utterance 2 that are excluded from analysis are the imitated utterance (see text for further details).

Francesco Cutugno, Antonio Origlia and Valentina Schettino

7 Syllable structure, automatic syllabification and reduction phenomena

Abstract: Our contribution aims at describing the differences between the expected syllabification and the actual realization of syllable chains in connected speech. In particular, starting from a definition of both phonetic and phonological syllables, we will concentrate on the problem of automatically detecting these units in the acoustic sequence. We will describe the relationship between a segmentation obtained by analyzing the physical information provided by the sound and the one obtained through lexically/phonologically predicted syllable analysis. Mismatches between automatically detected and expected syllabic units will be analyzed; syllabic units obtained by applying automatic analysis will be classified in order to list the most relevant (and frequent in our testing corpus) reduction phenomena. We will start by separating syllables that effectively appear as reduced from the nonreduced ones. The reduced ones will be classified on the basis of a "positional" criterion related to syllabic substructure (onset, nucleus, coda). To conclude, we will show how important it is for the engineering and linguistic communities to cooperate on this task. While automatic segmentation fails to capture specific phenomena and needs linguistic support to improve its performance, the expected segmentation should be critically analyzed with respect to the automatic one. We will propose that the two levels, although strongly related, are different in nature. This means that neither can be considered "right," as each one should be considered more appropriate in different situations. The course of action we suggest is to concentrate on the analysis of mismatches, in order to create a rule-based interface between the two levels that in many cases could lead us to predict where and when reduction can fall. This interface level will intrinsically be language dependent as, in this view, reduction is a "normal" property, whose constraints characterize the predictability of each language.

Keywords: Syllable, automatic syllabification, expectation mismatch

Francesco Cutugno, University "Federico II" of Naples
Antonio Origlia, University "Federico II" of Naples
Valentina Schettino, University "L'Orientale" of Naples - Italy

https://doi.org/10.1515/9783110524178-007

7.1 Introduction

Speech communication is an extremely complex process whose success depends on a number of different factors and on their interaction: a message must be decoded independently of different acoustic realizations, diversity of diaphasic and diamesic variables – that is to say, different registers or means of communication – and other noncanonical phenomena. Nevertheless, listeners are generally able to understand the linguistic message. In this work, we investigate the alterations found in the speech signal in relationship with phenomena predicted by using phonological rules. We also examine what kinds of phenomena shape the linguistic message during connected, spontaneous speech. Our interest is in the effect reduction processes have on syllabic structures. We specifically refer to Greenberg (1999), Lindblom (1990) and Savy and Cutugno (1997). As Greenberg (1999:159) stated: "No two speakers utter the same words in precisely the same way, and it is rare for the speech of even the same individual to repeat precisely over the course of a day (or even a lifetime), despite the apparent ease with which the acoustic waveform is linguistically decoded."

It would seem that canonical, phonologically predictable forms are an artificial construction, far from the actual realization of linguistic messages: phonemes are not always produced in an expected way due to contextual and articulatory reasons.

According to Lindblom (1990), even trained phoneticians encounter great difficulties in trying to predict which phones will appear in the signal. Lindblom explains this peculiarity of spontaneous speech with the help of biological constraints, the principles of economy and plasticity: speakers default to a low-cost form of behavior, acting in a purpose-driven, output-oriented way. Linguistic articulation, then, has to be seen as the consequence of hypo- or hyperspeech: the speakers adapt in each situation to the necessary speech style, along with physiological, cognitive and social constraints.

This view is also supported by Savy and Cutugno (1997), who examine and describe different types of vowel reduction in Italian:

> The hypoarticulation phenomenon is a general one: it affects the whole segmental structure of speech. In connected speech production most of the articulatory target positions are not reached, consequently acoustic parameters corresponding to the gestures appear to be significantly different from the ones observed in the formal production. In this view, the term hypoarticulation is used to define a class of reduction phenomena. (2)

In this view, hypoarticulation is closely connected to speech accuracy – that is to say, to the degree to which an articulatory target position is reached – and is distinguished from other phenomena whose predictability can take place in the grammar (see Farnetani and Recasens 1997). In particular, Savy and Cutugno (1997) argue that centralization – defined as the speakers' tendency to realize

unstressed vowels in a more central articulatory position – has a different status compared to hypoarticulation. The former is a systematic reduction linked with the presence/absence of the accent, whereas the latter is linked to unpredictable (i.e., nonsystematic) reduction processes and reduced articulatory effort. However, their effects are combined to different degrees in different speech styles.

As a matter of fact, reduction processes deeply determine the structure of spontaneous speech, so that actual realizations and canonical forms appear to be two different levels which are only loosely connected.

Many authors, for example, Nam et al. (2009), have noted the preponderance of CV (henceforth: C=consonant, V=vowel; here, for example, CV means Consonant-Vowel syllable type) syllables across different types of languages: in fact, despite (or by means of) reduction processes, this structure appears to be the most resistant to reduction as well as the most widespread. This fact seems to hint at and confirm the syllabic basis of spontaneous speech, as Greenberg (1999: 168) set forth:

> The importance of the syllable as an organizational unit of spoken language becomes manifest when considering pronunciation variation. In spontaneous speech the phonetic realization often differs markedly from the canonical, phonological representation [...] Entire phonetic elements are often dropped [...] or transformed into other phonetic segments [...] Such patterns of deletion and substitutions appear rather complex and somewhat arbitrary when analyzed at the level of the phonetic or phonological segment. However, this variation becomes more systematic when placed within the framework of the syllable.

Greenberg (1999), measuring syllabic types' distribution in a corpus of spontaneous American English, observes that the great majority of words appearing in spontaneous speech are monosyllabic and disyllabic: this statistical datum suggests that the decoding of the speech chain is deeply linked with syllabic units and lexical structure, which in this language are almost identical. However, what happens in languages with a wider distribution? Our contribution to this debate will concern Italian and German. Moreover, as we will see in our data, syllable structures of languages traditionally described as "stress-timed" can be much more complicated than those of so-called syllable-timed ones (Pike 1945), because of complex consonantal clusters allowed both in the onset and in the coda. However, spontaneous speech is composed mostly of CV-syllable types in both rhythmical types. This aspect hints at the central importance of the syllabic unit for spontaneous, connected speech. For this reason, in our empirical analysis we will investigate the syllabic types that most often undergo reduction, as well as their degree of resistance to reduction, in order to verify Greenberg's hypothesis, and to be able to give a better interpretation of the role of syllables in spontaneous speech. Therefore, we investigate reduction processes in connection with the syllable structure, comparing the predictions of a phonological analysis with the actual phonetic realization, thus examining the degree of predictability of actual,

realized spontaneous speech. We are also motivated by the positive implications of considering syllables as the basic unit of description in many linguistic theories and models related to Natural Language Processing. Indeed, using syllables to investigate speech production processes – for example, reduction – could make it easier to introduce linguistic units in computer applications, as opposed to technological units.

We argue that, in principle, speech technology should rely on phonological and phonetic theoretical models to handle speech and language. In practice, technology often disregards linguistics by relying more on statistics and practical rules than on linguistic structure. While these approaches are, indeed, inspired by linguistics, a limited comprehension of the relationship between linguistic expectations and phonetic realizations makes it difficult to analyze and model the observable signal, which is the basis of any speech-based technological system.

In the following, we will describe both the theoretical conditions and the empirical tests we performed. In Section 7.2, we discuss the syllable as a phonological/phonetic unit, highlighting its role in the decoding of connected speech and the problems related with its definition, which influences the analysis of reduction processes. In Section 7.3.1, we will describe our automatic syllabification system. In the rest of the work, we will present our data and the criteria used for the annotation: we will talk about our corpus and its contents, depicting the phonological rules and the phonetic constraints used for the respective annotations; we will also explain how we segmented the signal, making special references to extra-lexical reduction phenomena. An in-depth analysis of data will follow, regarding both Italian and German: results will be discussed, and in particular, the relationship between phonological and phonetic syllables and related reductions.

7.2 Syllable in phonology, phonetics and technology

Even if the syllable is beginning to assume a fundamental role for the study of stress, accent, intonation, duration and rhythm, a clear and uncontested definition of the syllable is yet to be found. As Jones et al. (1997) state, syllables are perhaps easier to identify than to define.

The problem with the definition of syllables is that the various features linked with the syllabic unit have not been settled yet. Bell and Hooper (1978: 4) list the great abundance of various aspects of the syllable that have been investigated, and show that no shared conclusions have been reached.

In the field of articulatory phonetics, for example, the syllable has often been connected with jaw movements (De Saussure 1967) or chest outbursts

(Stetson 1951). From an acoustic point of view, energy plays a very important role: Jespersen (1920) was the first author to recognize that syllable nuclei usually correspond to energy maxima, whereas energy minima are normally associated with syllable boundaries.

In attempting to solve the problem of syllable definition, it is important to remember that during speech production, many processes take place (mainly in the form of spontaneous reductions of the phonetic material) leading to deep changes in the syllabic structure.

> [...] word production is a compromise between articulatory economy for the speaker and acoustic distinctivity for the listener. [...] These physically constrained tendencies to reduce effort are in their turn controlled by linguistic structures at all levels, from phonology to syntax and semantics, and therefore have different manifestations and distributions in different languages, although basic types can be generalized. (Kohler 1998: 29)

Speakers tend to change speech sounds' structure in order to reduce articulatory effort; many coarticulation and assimilation processes (see Farnetani and Recasens 1997) take place in this way. While the latter are included in most phonological theories, the former are hardly explicable from a phonological point of view and they are usually confined to the acoustic descriptive domain (surface phenomena) and not included in any grammatical rule set. The difference between the actual spoken chain and the expected structures emphasizes the fact that the syllable must be analyzed and defined taking into account both phonological and phonetic properties and concepts.

The definition of syllables also depends on the observed language, on the phonotactic rules involved, on the morpho-phonological description adopted for that language and on some particular phonetic constraints. A pure phonetic definition of the syllabic unit can also be difficult: "[...] The concept of the syllable as an entity at the phonological level enjoys no more general a consensus than that of the phonetic syllable" (Laver 1994: 114).

Nespor (1993) agrees with this point of view, and additionally declares that it is nearly impossible to find a good definition of the syllabic unit because it is difficult to identify syllable boundaries on purely phonetic grounds. However, since phonological theories are thus far unable to deal with all types of reduction, the attempt to find a phonetic definition may be helpful.

A definition of the phonetic syllable that we consider to be nearly acceptable was proposed by Laver (1994: 114): "[...] a complex unit made up of nuclear and marginal elements. Nuclear elements, as phonological entities, are what we have been calling vowels. Marginal elements are what we have been calling consonants."

This definition has the advantage of being very general; yet the listed properties are not precise enough to explain and describe all the complexity included

within the syllabic structure. Another attempt was made by Roach (2000: 70), where syllables "[...] are usually described as consisting of a centre which has little or no obstruction to airflow and which sounds comparatively loud; before and after that centre [...] there will be greater obstruction to airflow and/or less loud sound."

All attempts at defining the syllable have encountered some difficulties; as a consequence, it is useful to make an empirical comparison between the phonetic realization and the phonological predictions made for a given set of utterances. In this way, we can finally understand in what ways we need to change our phonological, theoretical expectations, in order to adjust them to the actual phonetic realizations.

Given their representative power from a modeling point of view, any technological systems using syllables as basic units, however, have been designed in such a way that the signal correlates of syllables were prioritized with respect to phonological expectations. The definition given in D'Alessandro and Mertens (1995: 267) for the *phonetic syllable* (sometimes called the *pseudosyllable* as in Martin 2010) is: "[...] a continuous voiced segment of speech organized around one local loudness peak, and possibly preceded and/or followed by voiceless segments."

However, while we approve the term *phonetic syllables*, Roach's definition better accounts for voiced consonants and, in our opinion, is particularly interesting for technological approaches as it is entirely based on phenomena observable in the signal. Roach's definition will be the one we will refer to in the rest of the chapter when talking about *phonetic syllables*. This will be discussed further in Section 7.3.

In order to understand the real nature of the syllabic structure in connected speech, we make a comparison between actual phonetic realizations and predicted phonological syllables. We hypothesize that, as a consequence of the explicit reduction and coarticulation processes, the two levels will have a substantial amount of differences, and that the first consequence will be the alteration of the expected syllable segmentation.

7.3 Automatic and linguistic syllabification

As a unit of analysis for signal processing, the syllable has played a role in a number of applications. In this section, we present some examples of how the phonetic syllable has influenced the definition of speech analysis algorithms. As a consequence, segmental structure defined by the occurrence of phonetic

syllables influences the technological systems presented here. Reduction processes, by contributing to the occurrence of such units, play a significant role. We concentrate on automatic segmentation in syllable units and pitch stylization, as these have been used in recent years to perform prosodic analysis in experimental fields such as emotion tracking (Origlia, Galatà and Cutugno 2014, 2015) and personality perception (Mohammadi et al. 2012). Moreover, we analyze how the segmentation of the syllabic chain comes into being from the linguistic perspective. We describe the major set of extra-lexical phenomena that influence syllabification, that is, reduction processes and related events.

7.3.1 Automatic syllabification

The process of automatically identifying syllable boundaries on the basis of acoustic information detected in the speech signal is called syllable segmentation. This process is important in speech processing because it is connected to prosodic factors including rhythm and tempo and also because the hypothesis that syllables can be used as basic units in speech recognition has been investigated for a long time, cf. Wu et al. (1997) and Jones et al. (1997).

In this work, we are interested in which acoustic cues are useful for an automatic syllabic segmentation. In the field of articulatory phonetics and phonology, some authors link syllables with jaw movement (De Saussure 1967), others to chest burst (Stetson 1951); alternatively, they consider syllables as the basic units of speech programming (Kozhevnikov and Chistovich 1965). From the acoustic point of view, energy temporal patterns play a fundamental role: Jespersen (1920) was the first to link syllabification with energy oscillation, observing that syllable nuclei are usually found in correspondence with energy maxima, while syllable boundaries correlate with energy minima. A first attempt at the automatic segmentation of a speech utterance into syllabic portions was presented in Mermelstein (1975). In this work, a loudness function obtained by assigning a weight to each element within a set of spectral bands was used. An algorithm evaluating the shape of the loudness pattern (convex-hull) was then employed to find syllable boundaries.

We report results obtained on a corpus of Italian read numbers by two approaches concentrating on energy fluctuations, presented in Petrillo and Cutugno (2003) and Ludusan (2010). The difference between the two lies in the fact that the former considers energy valleys directly in order to identify syllable boundaries. The latter also considers syllable nuclei, described as voiced energy peaks, as well as the first derivative of the energy profile, to identify syllable markers.

The accuracy of these segmentations is assessed on the basis of the occurrence of three types of errors:

- Deletions
- Insertions
- Substitutions

A deletion error occurs when a syllable is not recognized at all, that is, when two syllables are expected and only one segment is found. Insertion errors, conversely, occur when an expected syllable is split into two segments. The final type of error, substitutions, produces a difference in the temporal positions of the boundary markers between the two types of annotation. This does not alter the number of detected segments with respect to the reference annotation.

Table 7.1 summarizes the results obtained by the two approaches with the employed evaluation system. The presented summary shows that using syllable nuclei as a reference to start looking for syllable boundaries produces an improvement in the system's precision in marker positioning. Correctness and accuracy are reported for both approaches. Correctness is defined as

$$C = \frac{N-D-S}{N} \tag{1}$$

where N is the number of automatically positioned markers, D is the number of deletions and S is the number of substitutions. Accuracy is defined as

$$A = \frac{N-D-S-I}{N} \tag{2}$$

where N is the number of automatically positioned markers, D is the number of deletions, S is the number of substitutions and I is the number of insertions. Not considering insertion in the correctness measure provides data about the ability of the system to detect and position syllable boundaries. Accuracy, on the other hand, gives a more complete overview of the system by also considering false positives, and is usually the main measure used to evaluate syllabification algorithms.

An approach based on syllable nuclei detection (Ludusan) increases the number of insertions due to false positives in the nuclei detection step but to a negligible degree. In fact, both correctness and accuracy are higher in the nuclei detection approach than in the original approach based on energy minima only. Table 7.1 also highlights that the improvement is caused by an increase in correctly positioned markers as the number of substitution drops.

Automatic syllabification must be compared with a manual annotation in order to be evaluated. As the algorithm is based on the objective analysis of

Table 7.1: Results obtained on the SPEECON corpus (in %) by the new algorithm (Ludusan) and by the baseline approach for Italian (Petrillo and Cutugno).

	Sub	Ins	Del	Corr	Acc
Ludusan	2.03	6.19	5.72	91.74	85.14
Petrillo–Cutugno	4.13	5.74	5.61	89.67	83.58

acoustic features, it is possible to observe mismatches in the comparison that are actually hard to judge. An evaluation algorithm always considers the manual annotation as reference and, by doing so, penalizes automatic algorithms even in those cases where the proposal would be acceptable because the situation is uncertain. This points toward a use of automatic syllabification algorithms in semi-supervised applications, where a first proposal is produced by the system and then refined by a human, thus reducing the amount of work for the human annotator.

7.3.1.1 Pitch stylization

Automatic syllabification is based on a perceptual account of syllables. That is, Roach's definition, along with others, relates the occurrence of syllable boundaries to specific acoustic patterns. While no one definition appears to satisfy the researchers, however, they at least have in common the idea of a repeating acoustic pattern that can be exploited to extract other kinds of information from the signal, mainly of perceptual value. We must therefore consider the relationship between subparts of phonetic syllables and the perceived pitch movements.

While fundamental frequency (F0) is the main acoustic correlate of intonation, it does not necessarily represent what it is actually *heard* by the human ear. t'Hart et al. (1990: 25) argue that "No matter how systematically a phenomenon may be found to occur through a visual inspection of F0 curves, if it cannot be heard, it cannot play a part in communication." This led to the definition of *stylization* as an approximation of the F0 curve by means of linear segments. This can be defined as a sequence of segments that "[. . .] should eventually be auditorily indistinguishable from the resynthesized original and it must contain the smallest possible number of straight-line segments with which the desired perceptual equality can be achieved" (t'Hart et al. 1990: 42).

Among attempts to account for intonation, the MOMEL (MOdelling-MELody) algorithm (Hirst and Espesser 1993) has been widely used in the literature. This algorithm does not produce a stylization as per the definition given above, since its goal is to produce a model of the macroprosodic component, which can be used

together with the microprosodic component the algorithm produces to rebuild the original pitch curve. In this sense, a stylization can be seen as a lossy filter for microprosody while the output of the MOMEL algorithm does not discard the microprosodic component. Nevertheless, the macroprosodic profile obtained with MOMEL is usually considered as reference for stylization algorithms.

D'Alessandro and Mertens (1995) use the concept of dynamic tones, or glissandos in order to produce a stylization of the pitch curve. The Prosogram, a perceptually motivated representation of the pitch curve (Mertens 2004), is based on this algorithm. This representation involves a segmentation of the utterance under consideration into syllables, to represent the pitch curve in terms of glissandos and static tones. Wypych (2006) and Ravuri and Ellis (2008) also used the concept of syllables to position the linear segments used in the stylization. In Ghosh and Narayanan (2009), the pitch stylization problem was treated as an optimization problem for the first time by using a Dynamic Programming algorithm designed to optimize the position of a predefined number of segments estimated on the basis of the findings presented in Wang and Narayanan (2005). As a quality measure, this algorithm used the statistical closeness between the stylized curve and the original one.

Statistical closeness, however, is a visual, rather than acoustically perceptual, measure. The goal of the stylization process, as defined by t'Hart et al. (1990), is to obtain a perceptually equivalent representation of pitch. Investigation into the human ability to perceive pitch movements has shown that pitch perception does not rely solely on the F0 curve. Several studies have found that this process depends on the dynamic interaction of F0 with other factors, mainly the intensity shape and the position of relevant movements with respect to the syllabic structure. The influence of intensity on pitch perception has been studied extensively (Feth 1972; Maiwald 1967; Møller 1974; Rossi 1978; Zwicker 1962); however, some results obtained in the 1990s gave clear indications regarding other perceptual phenomena that should be taken into account when performing pitch stylization.

House (1990) introduces the Spectral Constraint Hypothesis, which states that "As the complexity of the signal increases [. . .], pitch sensitivity decreases"; that is, tonal movement perception capability was described as inversely proportional to the amount of change in energy and spectral information independently of the type of change. House (1996) updated the optimal tonal perception model to take into account the syllabic structure, stating that

> The area of maximum new spectral and intensity change occurring typically between syllable onset and syllable nucleus appears to be a crucial point for the timing of tonal movement. Movement through this area will be recoded as tonal levels as indicated in the earlier model. However, movement through the beginning of the syllable coda can be perceived as movement per se and described using movement features. (2051)

By phonetically reinterpreting the framework used by House, we can link tone perception to synchronized energy behavior both in terms of movement type (rising/falling) and in terms of rate of change (slope). These features are basically the same as those that describe the occurrence of a phonetic syllable. It is, then, that the occurrence of phonetic syllables impacts the perception of synchronized pitch movements. Such superimposition was used by Origlia et al. (2013) to develop the Optimal Stylization algorithm, later refined in Origlia and Cutugno (2014) and presented as the Simplified Optimal Stylization (SOps) algorithm. The basic idea behind SOpS is that glissando perception should not be interpreted as a fixed threshold, as in Mertens (2004).

In t'Hart et al. (1990), the first glissando perception threshold was derived by comparing the different results that were obtained in previous works and it was defined as

$$g_{\text{thr}} = 0.16/T^2 \tag{3}$$

where T is the time interval in which the movement is realized. Mertens (2004), however, found that, when comparing an automatic prosodic transcription, the *Prosogram*, with one provided by humans, the best result could be obtained by doubling the constant value at the numerator of Equation (3) and using it to produce a pitch stylization by the method presented in D'Alessandro and Mertens (1995). The way in which the formulas in t'Hart et al. (1990) and D'Alessandro and Mertens (1995) are constructed implicitly assumes that the threshold they define is intended to be absolute. As a matter of fact, in t'Hart et al. (1990: 33) Equation (3) was named "absolute threshold of pitch change". Given the perceptual nature of the problem, however, an improved solution would be to consider a continuous likelihood measure, taking Mertens' threshold as reference for the maximum likelihood value, and to include synchronized pitch movements taking into account the findings of both Rossi (1978) and House (1996). To define such a measure, we describe a point of the S curve as

$$s_x = (v_{s_x}, t_{s_x}) \tag{4}$$

where v_{s_x} is the value in semitones of the point and t_{s_x} is the time instant of the point in ms. Then, we can define the rate of change of the segment $[s_x, s_y]$ as follows:

$$V([s_1, s_2]) = s_x = \frac{|v_{s_y} - v_{s_x}|}{t_{s_y} - t_{s_x}}. \tag{5}$$

We can then define the likelihood of a linear pitch movement to be heard as a glissando as

$$r_g([s_1, s_2]) = \begin{cases} 1 & if\ V([s_1, s_2]) > \dfrac{0.32}{T^2} \\ \left(\dfrac{V([s_1, s_2])T^2}{0.32}\right)^{\gamma}, & \text{otherwise} \end{cases} \quad (6)$$

where γ is a correction factor based on the slope of the energy profile and depends both on the direction of the local energy profile and on its slope. The effect is to dampen glissando perception capability proportionally to the slope of the energy profile if the pitch movement is aligned with an ascending energy movement. In the area of the phonetic syllables' nuclei, where spectral stability is expected ($E^0 \approx 0$), there will be little or no gamma correction on the perceived pitch estimated in Equation (6) while pitch movements aligned with falling energy glides will be more likely to be heard as glissandos. This is summarized in Figure 7.1.

Origlia et al. (2013) and Origlia and Cutugno (2014) have shown that utterances resynthesized with the obtained linear approximations of the pitch curve are perceived by human listeners, in most cases, as equivalent to the original ones. Moreover, the number of points used by the stylization algorithms decreases as the tonal perception model is improved without damaging perceptual equal-

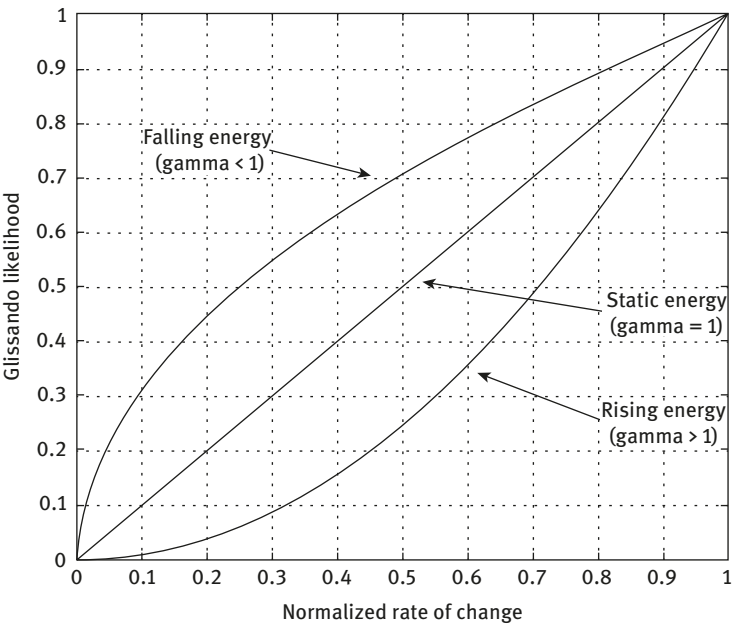

Figure 7.1: Glissando likelihood values are computed on the basis of energy movements in terms of gamma correction. In this figure, we report glissando likelihood value transformations for glissandos not exceeding Mertens' threshold.

ity. For details, we refer the reader to the cited works as our interest, in this discussion, is to make it clear that reduction processes, by influencing the acoustic patterns linked to phonetic syllables, indirectly influence the way pitch movements are perceived or at least influences the planning of pitch targets' positioning.

7.3.2 Segmentation and extra-lexical phenomena

Syllable segmentation can be very ambiguous from a linguistic perspective, as well, and for different reasons; however, the central concept remains the Sonority Sequencing Principle (SSP; Clements 2009), by means of which most borderline cases can be disambiguated. Sonority seems to be an embedded, deep-rooted element in the human brain for the decoding of linguistic meanings. Some authors even state that newborns' brain response to ill-formed syllables confirms the later tendency of privileging sonority constraints: "[. . .] neonates are sensitive to putatively universal restrictions on syllable structure. The observation of such regularities close to birth, before the onset of experience with language production, shows that sonority-related biases in human do not require extensive linguistic experience or ample practice with language production" (Gómez et al. 2014: 5839).

Nevertheless, some unexpected extra-lexical reduction phenomena can alter the syllabic segmentation: the SSP does not predict how connected speech will be structured and syllabified. We will now describe and examine these phenomena.

7.3.2.1 Reduction processes and related examples

Spontaneous speech is characterized by a large amount of unpredictable phenomena, acting independently from morphological or lexical constraints. At the top of the list, there is the process of coarticulation:

> A fundamental and extraordinary characteristic of spoken language, of which we speakers are not even conscious, is that the movements of different articulators for the production of successive phonetic segments overlap in time and interact with one another: as a consequence, the vocal tract configuration at any point in time is influenced by more than one segment. This is what the term "coarticulation" describes. (Farnetani and Recasens 1997: 316)

This process is especially relevant in connected speech: in fact, every segment can undergo coarticulation, because it is always confronted with a spoken context. Segmental changes due to coarticulation are then the most widespread forms of reduction; additionally, this process involves the whole segmental structure, because the tendency to reduce the articulatory effort is always present (Savy and Cutugno 1997). In the organization of articulatory movements, phonological expectations

are deeply forged by this process; according to Lindblom (1996: 1684), these processes are directly ascribable to so-called target undershoot:

> [...] listeners do not perceive and understand speech by way of articulation. In developing that view, we shall present evidence supportive of the following claims: (1) The listener and the speaking situation make significant contributions to defining the speaker's task; (2) that task shows both long- and short-term variations; (3) the speaker adapts to it; (4) the function of that adaptation is not to facilitate articulatory recovery, but to produce an auditory signal which can be rich or poor, but which, for successful recognition, must minimally always possess sufficient discriminatory power. With respect to coarticulation we can restate (3) and (4) by saying the following: (5) In the ideal case, the speaker will allow himself only so much coarticulation as the listener will tolerate. By definition, then, a successful signal contains just enough auditory information to perform its task: to be discriminatory.

In other words, articulatory target positions are not fully reached; in order to minimize articulatory effort, speakers try to find a compromise between targets and intelligibility, thus modifying the actual realizations in the signal.

However, coarticulation is not the only reduction phenomenon that takes place in spite of morphological and lexical constraints. According to Greenberg (1999: 162), when analyzing connected speech, the process of resyllabification has to be taken into account: "[...] there is the thorny issue of 'syllabification' (i.e., segmentation into syllabic entities) as well as 'resyllabification' (which deals with the reassignment of a phonetic constituent (rarely more than one) from one syllable to another. These issues are of concern when comparing the phonetic transcription with the canonical form in the lexicon."

Through this process, phonetic material – more precisely, consonantal material – is relocated to an adjacent syllable: as a consequence, the lexical form can be completely destroyed, as we can see in Figure 7.2.

In this case, the first two syllables of the German sentence *jetzt gehen sie in* 'now they go inside' were merged into a single unit, whose consonantal elements originally belonged to different segments.

Resyllabification can be caused by different mis-realizations: assimilation, elision and vowel deletion are all reduction phenomena that cause a restructuring of the syllabic segments.

As we can see in Figure 7.3, in the Italian word *questa* ['kwesta] 'this' the voiceless alveolar stop is assimilated to the preceding voiceless alveolar fricative: the disappearance of the stop causes the re-organization of the syllabic chain, originating a new, unexpected segmentation.

With respect to elision, Bigi et al. (2010: 3286) define it as follows: "Elision is the omission of one or more sounds (such as a vowel, a consonant, or a whole syllable), producing a result that is easier for the speaker to pronounce."

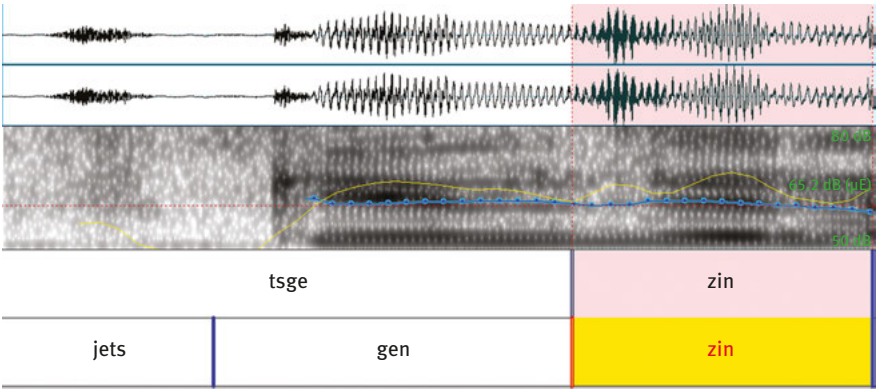

Figure 7.2: An example of resyllabification: jets + gen → tsge.

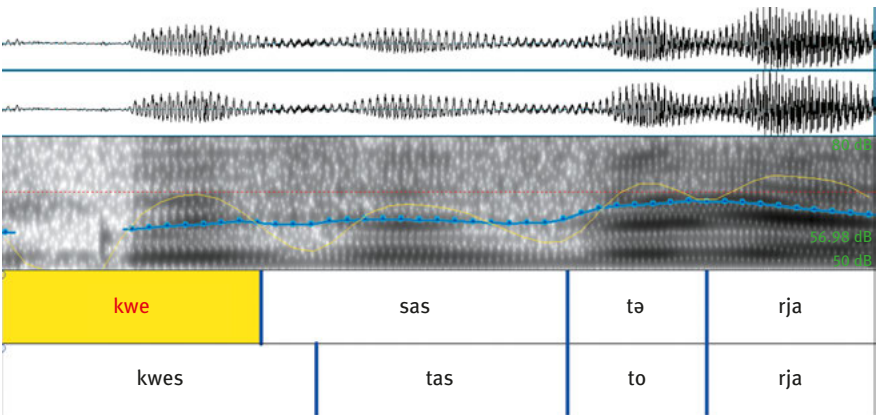

Figure 7.3: An example of assimilation: kwes + tas → kwe + sas.

For example, in both German and Italian, the last unstressed vowel is often omitted. In Italian this typically happens when it precedes another vowel; this is in order to avoid hiatuses, which are characterized by significant articulatory effort.

In Figure 7.4, we can see two examples of vowel deletion: in the Italian sentence *se ne accorge* 'he realizes' the unstressed final vowel [e] is left out from the actual realization; moreover, one of these examples corresponds to the typical Italian phenomenon of deletion of vowel before another vocalic phone (*ne accorge* → *n'accorge*).

This phenomenon is also essential in German, as Kohler and Rodgers (2001: 97) note: "Vowel reduction and deletion, and in particular deletion of schwa, play a pivotal role in reduction phenomena. Kohler (1990) points out that in a series of processes that derive reduced realizations from canonical forms, it is either a precondition or intermediate stage [. . .]"

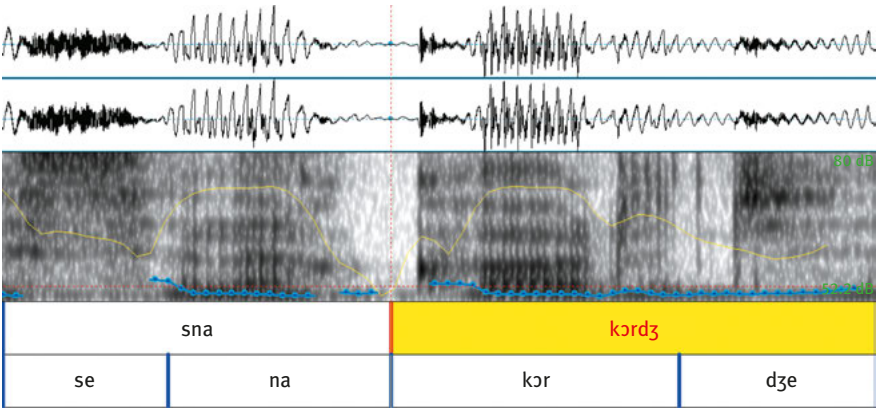

Figure 7.4: Elision examples: se + na → sna, kOr + dZe → kOrdZ.

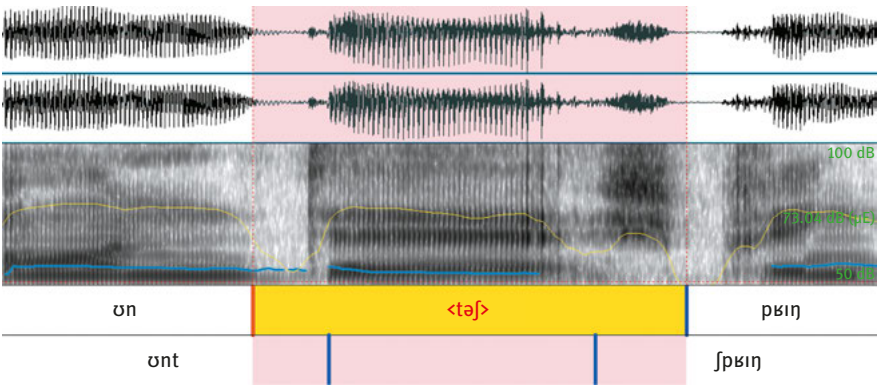

Figure 7.5: An example of hesitation annotation.

Although the current examples originate from only German and Italian, these phenomena are normally observable in many other languages.

We will now take a closer look at a few interesting examples found in our corpus. Spontaneous speech is characterized by repetitions, phatic signs, pauses and corrections: these phenomena organize and define oral communication.

Filled pauses often lead to resyllabification, because the inserted vowel builds a new nucleus; we can see an example of this phenomenon in Figure 7.5. This particular speaker uses the German word *und* 'and' as a phatic sign; in fact, he introduces every new sentence with this conjunction. Moreover, after this word he almost always produces a filled pause, interpretable as hesitation: for the annotation of these cases, we have used angle brackets.

Another important phenomenon of connected speech is repetition related to corrections. In the example in Figure 7.6, corrections are required: it is not "the dog" (Germ. *der Hund*) but "the frog" (Germ. *der Frosch*) that acts as subject of the sentence; as we can see, the speaker repeats the article two times when correcting the slip: it should be noted that the article is always pronounced in a reduced form, whether it is a repetition or not. Repetitions and corrections are, therefore, not necessarily related to a better pronunciation or a more understandable register.

Yet another phenomenon we observed in our corpus is the lenition of the voiced bilabial plosive [b] in German. When this phone appears in an intervocalic position, it is sometimes changed into an approximant [w]. An example is given in Figure 7.7.

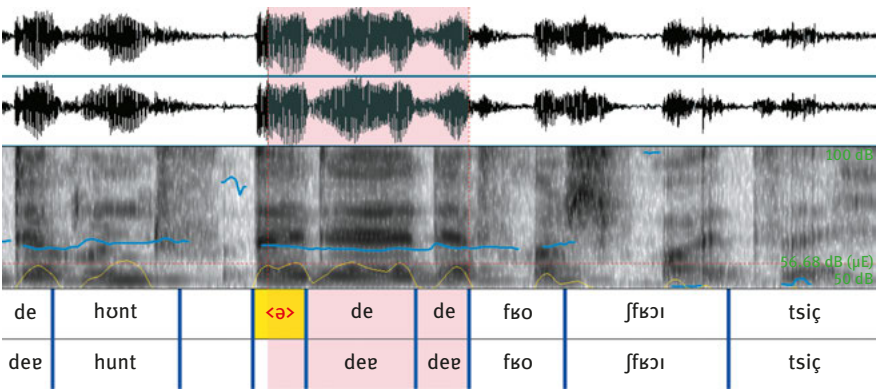

Figure 7.6: A reduced unit, in this case the article *der*, is kept reduced in the case of a correction/repetition.

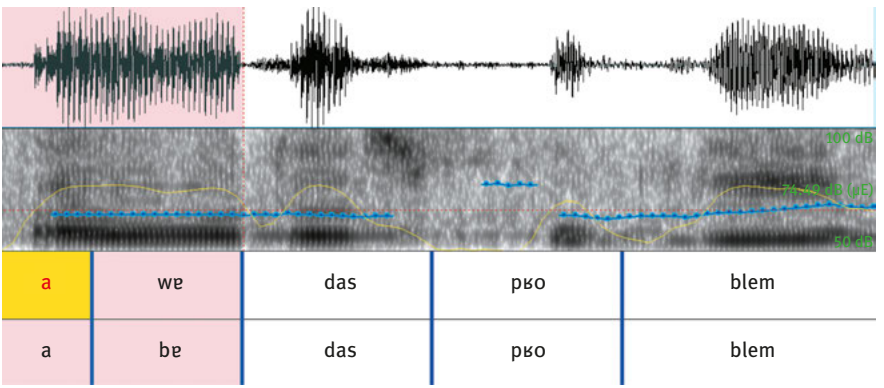

Figure 7.7: An example of lenition: /b/ → [w].

In order to allow an investigation of the impact reduction has from the point of view of syllable structures, a Praat (Boersma and Weenink 2011) script was used to convert manual annotations into CV annotations. The script was based on a simple conversion table used to identify and substitute symbols for vowels and consonants.

7.4 The case study

The analysis of connected speech in a real communicative situation is one of the most challenging topics in linguistic research today. This interest in the *normal* communicative situation has led to a new thread of studies, integrating theoretical approaches and empirical implementations: more and more attention is being paid to prosodic structure of the speech chain and in particular, in this view, the role played by the syllable is becoming more and more relevant.

On the basis of the syllable definitions given in Section 7.2 and building on previous similar work (see Origlia and Schettino 2014), we present here a series of measurements on Italian and German syllable sequences (both expected and realized) by using the software Praat. To obtain the segmentation, our first step was to divide the recordings into utterances. We segmented the signal so that there was at least one second of pause at the beginning and at the end of the audio-file; our main purpose was to avoid any form of dependency or influence between any two different chunks.

We analyze the relationship between expected and realized syllable structures in Italian and German to provide a general view on the kind of alterations that can be predicted versus only be observed in the signal.

7.4.1 Data and annotation criteria

7.4.1.1 The NOCANDO corpus

In this work, we use the NOCANDO (Non Canonical Construction Corpus) corpus (Brunetti et al. 2009, 2011) for our investigation. This corpus consists of a linguistic and multilingual taxonomy of noncanonical syntactic constructions; the available languages are Catalan, Spanish, Italian, English and German. The Italian and German speakers are mostly Erasmus students and Professors at the Universitat Pompeu Fabra in Barcelona (Brunetti et al. 2011); unfortunately, precise information about the regional origin, both for Italian and for German speakers, was not provided. NOCANDO contains spoken narratives based on a

picture book written by Mercer Meyer, *Frog Stories*. Recordings were elicited by asking the speakers to comment on these pictures, narrating the events as represented. This is, therefore, semi-spontaneous speech from informants who were aware that they were being recorded. Speakers were allowed to scroll the pages before starting the recording; an examiner was always there to guarantee the consistency of the procedure.

This corpus presents a number of advantages: (1) all the materials (audio files, transcriptions and annotations) are freely available; (2) it was recorded under the same conditions for all languages, so the coherence and consistency of methods allows a high degree of comparability between different languages and (3) given that this corpus was originally collected to study noncanonical syntactic constructions, which are often found in informal, unplanned speech, it is characterized by many prosodic and segmental phenomena that are typical of natural speech.

7.4.1.2 Contents

The NOCANDO corpus contains more than 16 hours of spontaneous speech. It contains speech coming from 68 speakers and the resulting narrative units are 222. The speech material is orthographically transcribed but no further annotation is available. In Table 7.2, we summarize quantitative information about the German and Italian subsets we selected for this study.

7.4.1.3 Annotation criteria

Our experiment consisted of a comparison between the predictions made by a phonological model versus the actual phonetic realization with specific reference to syllable segmentation aspects; the languages examined are German and Italian. In order to carry out the comparison, we have segmented and annotated the selected utterances using Praat (Boersma and Weenink 2011).

Table 7.2: Quantitative data about the NOCANDO corpus.

	Italian	German
Speakers	11	8
Recording time	361 s	375 s
# Phonological syllables	1,279	1,098
# Phonetic syllables	1,017	944

The syllable segmentation was carried out using the semi-automatic approach implemented in the Prosomarker tool (Origlia and Alfano 2012). The preliminary automatic segmentation proposed by the tool is provided by a modified version of the algorithm presented in Ludusan (2010).

Starting from the automatic segmentation, which is prone to errors, a human expert (one of the authors) manually corrected the generated segmentation. Errors caused by Prosomarker are mainly due to the definition of syllable upon which the tool is based. Given that it takes into account only energy movements as an indicator for syllable boundaries (Jespersen 1920), in some cases it was not able to correctly identify unit boundaries marked by nasal consonants, which are characterized by an amount of energy similar to vowels. While an alternative approach considering alterations in energy distribution can be applied to compensate for this error, at the present time, manual intervention is still required in these cases.

In this way, we have obtained a precise phonetic syllabification: the obtained units were segmented considering physical properties only (energy movements, pitch contour). In addition, each speech chunk was also segmented into phonologically expected syllables. Details about the method followed to obtain this segmentation are provided in Section 7.4.2.

In order to detect differences in the syllabic structure between the annotations on the phonological and the phonetic levels, the last step consisted in annotating the segments. We used Praat to create different layers aligned with the signal: on the first one we annotated the actually realized syllables, whereas the second one was used to annotate the phonologically expected ones. As a consequence, the first layer was named "phonetic tier," while the second was labeled as "phonological tier." We assigned a label to each speech sound in the phonetic level and to each phoneme in the phonological level; in order to do this, we used a set of labels simplified from the International Phonetic Alphabet.

7.4.2 Phonological rules

In this section, we describe some of the phonological rules used both for German and for Italian. Our aim is to estimate which syllables on the phonetic layer can be predicted considering phonological rules only and in which cases a mismatch between predictions and realizations can be observed. The phonological rules used for the segmentation of expected syllables are taken from Eisenberg (1998) for German and Bertinetto (2010) for Italian.

Phonotactic principles and phonological and rhythmic constraints can be easily inferred from the bibliography; nevertheless, some phonological phenomena of these languages have to be examined more in more detail.

One of these phenomena is the behavior of /s/+consonant at the beginning of a word or a sentence. In fact, this phenomenon is a problem for the statement made by some authors that syllabification should be linked. According to Clements (2009), "[…] syllabification conforms universally to the Sonority Sequencing Principle according to which the segments in a syllable rise in sonority from the margin to the peak. Segments that cannot be gathered into syllables in this way remain extrasyllabic, linking to higher levels of prosodic structure such as the foot, the prosodic word, or the prosodic phrase." (p. 165) However, the cluster /s/+consonant at the beginning of a word or a sentence explicitly violates the SSP, as the degree of sonority in the onset decreases instead of increasing.

Concerning Italian, in Bertinetto (1999) an in-depth analysis of the problem is carried out, and conclusions state that the two phones cannot be tautosyllabic – that is to say, they cannot belong to the same syllable: either they are heterosyllabic or the problem is undecidable. In this work, many variables are examined: the kind of article appearing before the cluster, synchronic and diachronic changes in the Italian language, diatopic variants and assimilation processes are all considered elements. Synchronic and diachronic changes refer to variation along time, while diatopic variants refer to regional differences; concerning assimilation, it. Assimilation is an alteration of the phonetic material in which a phone influences another adjacent one so deeply that the second phone acquires one or more properties of the first. However, the appearance of /s/ before another consonant in the onset cannot be explained unless we consider it heterosyllabic and/or extrasyllabic. The same remarks are made for German: in fact, we can find counterexamples for the SSP, for example /ʃ/ in words such as *Strumpf* 'sock' or *Sprung* 'jump'; in these cases, it is marked as *Nebensilbe* (Vennemann 1982) or it is part of a phenomenon called *extrasyllabicity* (Wiese 1991).

In Italian, the situation of geminates is quite interesting: every consonant except /z/, /j/ and /w/ can be geminated both when it appears in an intervocalic position and when it appears between a vowel and /l/, /r/, /j/ and /w/. There is no conclusive agreement on the phonological status of geminates, and it must be admitted that it is not clear if they should be considered monophonemic or not (Loporcaro 1996; Muljačić 1972).

Moreover, a consonant in the onset must be a geminate if the preceding word is a stressed monosyllable or ends with a stressed vowel; this phenomenon is known as syntactic gemination and can also be recognized in the univerbation of some words, for example, *affinché*, *appena* and *davvero*. The doubling of the initial consonant does not take place when the word begins with the cluster [s]+consonant (*['das'spat:sjo]); another exception is encountered in the cases of /j/ and /w/: this fact confirms the articulatory weakness of both /s+consonant/

clusters and approximants. For the current study, we decided to ignore length and consider geminates as a distinct phone on the phonetic level of analysis.

Concerning the phonological annotation, German actually represents a great challenge, because its phonology/phonetics includes some critical points:

- The first problem is the presence of the so-called syllabic consonants, that is to say, consonants that can work as nucleus. From a phonological perspective, vowels are expected but acoustically difficult to measure; they are merged with the consonants (usually sonorants) by means of coarticulation. From a phonetic, acoustic, and perceptual perspective, on the other hand, vowel cues are simply expressed by means of sonority traces spread inside the syllabic consonantal cluster. Given that our study takes an acoustic perspective, when considering the consonant structures' onset and coda we annotated these syllables as CC on the phonological level, too.
- Likewise, we decided to consider vocalized consonants –for example /ɾ/ → [ɐ] – as simple vowels, because there are no articulatory and perceptual differences between them and other vowels.
- Some problems are linked to the German phenomenon known as *Auslautverhärtung* (final-obstruent devoicing). In some cases, it works against the Maximal Onset Principle; in fact, in accordance with this Principle, consonants should not be put in the coda, but in the onset of the following syllable instead. Nevertheless, morpho-phonological devoicing is typical of consonants appearing in the coda of the syllable. The two phenomena, then, are in conflict. For purposes of the current study, devoiced consonants were segmented together with the following syllable, in the onset.

7.4.3 Segmentation criteria

Bearing in mind what we have reported in Section 7.3, we have followed the rules summarized below for the syllabification:

- If a phone was wrongly segmented as an independent syllable by the Prosomarker tool, we merged it with the syllable it belonged to.
- When an unexpected phone was realized, the actual realized structure was reported on the phonetic level, while on the phonological level the expected syllable was annotated.
- Long pauses were annotated as silence.
- When a phonetic syllable was not detected because of creaky voice or other qualitative problems, its presence was manually added in the annotation layer.
- In the case of a syntactic gemination, the consonant was considered a distinct one, although long; long vowels also were labeled as distinct phones.

- We included schwa as a candidate for the syllabic nucleus since acoustically, there are no differences between [ə] and other vowels.
- Both phonetic and phonological syllables had to comply with the SSP: that is to say, the nucleus is always the most sonorous part of the syllable; all the preceding phones express increasing sonority, while the phones of the coda are characterized by decreasing sonority.
- Some consequences follow from the preceding statement, and in particular: when a fricative is followed by a plosive in the onset, the fricative actually belongs to the coda of the preceding syllable. This rule has only few exceptions, among which the cluster [s]/[ʃ]+plosive at the beginning of a speech chain. In this case some authors talk about extrasyllabicity (Wiese 1991; cf. also discussion above)
- In accordance with the Maximal Onset Principle, syllables had to be segmented prioritizing the presence of consonants in the onset and not in the coda; that is to say, when the phonotactic rules of a language cannot discern if a phone has to be put in the coda of the preceding syllable or in the onset of the following unit, the second choice is privileged.
- When nasals or fricatives appear at the end of a syllable and are directly followed by a vowel, they must build a syllabic unit together (Nespor 1993).
- Diphthongs and hiatuses were segmented according to phonetic and articulatory principles: that is to say, when we did not find a clear distinction in the signal between the two vowels, they were annotated as a single long phone.
- Syllables are considered as a nonlexical phenomenon connected to the continuous speech chain. Syllable segmentation process will take into account phenomena which normally occur across word boundaries, such as apheresis (phone or syllable deletion) or synalephe (vowel merging), which will be discussed in detail in the next section.

7.5 Data analysis

In this section, we will investigate the relationship between the syllable annotations on the phonetic and the phonological tiers, examining the data and the outcomes obtained in the German and Italian syllabification processes.

First, we will carry out a simple statistical analysis: we will count how many times syllables were not produced with their full phonological form because of reduction and coarticulation processes.

Second, we will make a qualitative analysis of these reductions: we will check which consonants are more often reduced and which clusters prove to be resistant against reduction. Moreover, we will also analyze the most frequent alterations

with respect to phonetic categories and we will investigate different syllabic types (e.g. CV, CVC or CC syllables) and the relative resistance degree against reduction.

Finally, we will discuss exceptions and special cases, in order to depict the situation from a linguistic perspective. We will show a wide set of cases in which the phonetic and phonological annotation levels do not match, and we will propose a classification to make it possible to concretely evaluate the relationship between the two levels.

When comparing syllabic units in the phonetic and phonological annotation layers the following three cases can be encountered:

- *Match:* we have marked as Match all the phonetic syllables whose structure and segmentation completely comply with the phonological counterpart;
- *Correspondence:* this label was used for every phonetic syllable whose boundaries were the same as the phonological level, but whose internal structure was different from the expected one;
- *Collapse:* every phonetic unit resulting from the merging of various phonological syllables by means of resyllabification was marked with this label.

The label Correspondence, then, refers to all the cases in which the structure of the phonetic syllable is unexpected. In these cases, the syllabification is exactly the same on both the phonological and phonetic annotation levels, that is to say, the unexpected realizations cause no reorganization of the syllabic sequence. Collapses, on the other hand, describe cases in which a resyllabification takes place.

An example of both types of unexpected realizations (correspondences and collapses) can be observed in Figure 7.8: In this sentence, in fact, there are two cases in which the phonological and the phonetic annotation levels do not

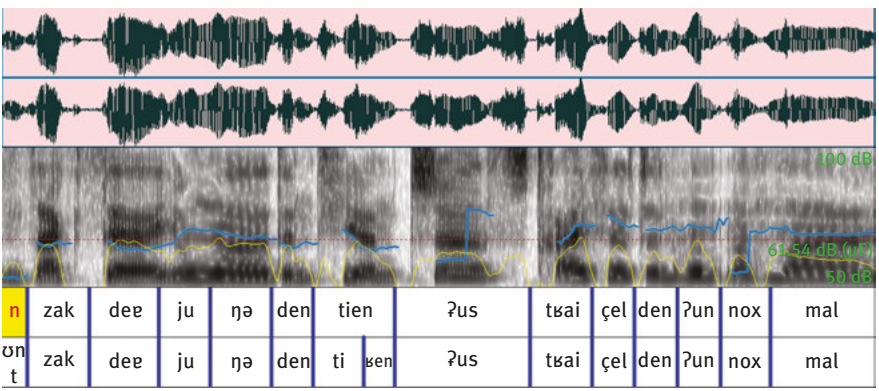

Figure 7.8: An example of a Correspondence (ɔnt → n) and of a Collapse (ti + ʁen → tien).

correspond; in the first case, what should have been [ʊnt] has been simply pronounced as [n]: the syllabification undergoes no change, but the structure of the syllable is deeply transformed; in the second case, the deletion of the fricative [ʁ] leads to a restructuring of the chain segmentation: two different syllables are merged into one, thus causing what is known as resyllabification.

7.5.1 The Italian data

In Figure 7.9, we present the situation in the Italian data. As can be clearly seen, a difference between the phonological and the phonetic annotation levels exists, meaning that not every phonologically expected syllable was actually realized. This allows us to confirm that reduction and coarticulation processes are very important in connected speech, because they deeply change the way in which the speech chain is structured, produced and perceived.

We also portray the percentage of matches, correspondences and collapses. Clearly, the phonological predictions and the phonetic realization do match to a great degree; nevertheless, approximately a quarter of the phonetic syllables do not correspond to phonological expectations.

Figure 7.10 shows the number of reduction processes involving each different phone class, that is to say, the qualitative properties of the discordances. With this histogram, we try to evaluate which phones are most often reduced. We have divided the phones in consonantal classes, in order to investigate what type of deletions take place in connected speech; in this taxonomy, *syls* expresses the complete deletion of a syllable.

In the Italian data, it appears that deletions of vowels and entire syllables are the most common forms of reduction. As a consequence, it could be said that the nucleus is of particular importance in the Italian speech chain. However, the reduction of a vowel does not always cause the deletion of an entire syllable: for example, if we have two vowels in a nucleus (e.g. a hiatus), the deletion or reduction of one of them produces no resyllabification, because the other one prevents

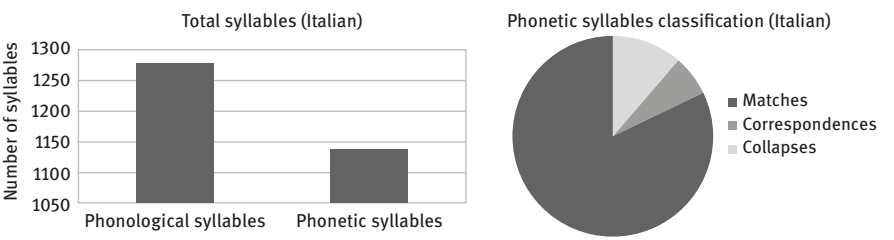

Figure 7.9: Italian data distributions.

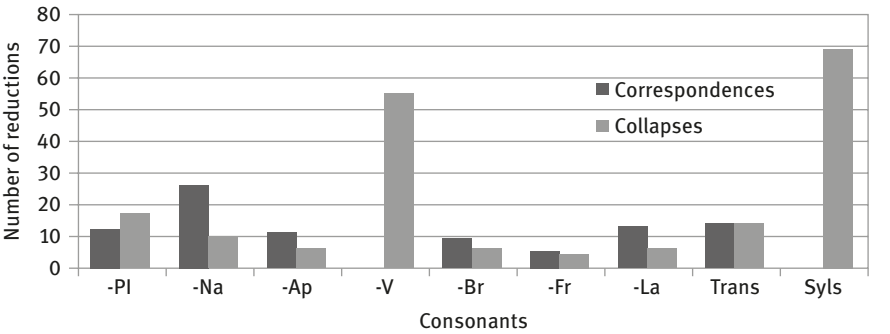

Figure 7.10: Impact of reduction processes on phone classes in Italian. The report indicates the number of deleted plosives (-Pl), nasals (-Na), approximants (-Ap), vowels (-V), fricatives (-Fr), vibrants (-Br) and laterals (-La). The histogram also reports cases of transformed phones (Trans) and syllable deletions (Syls).

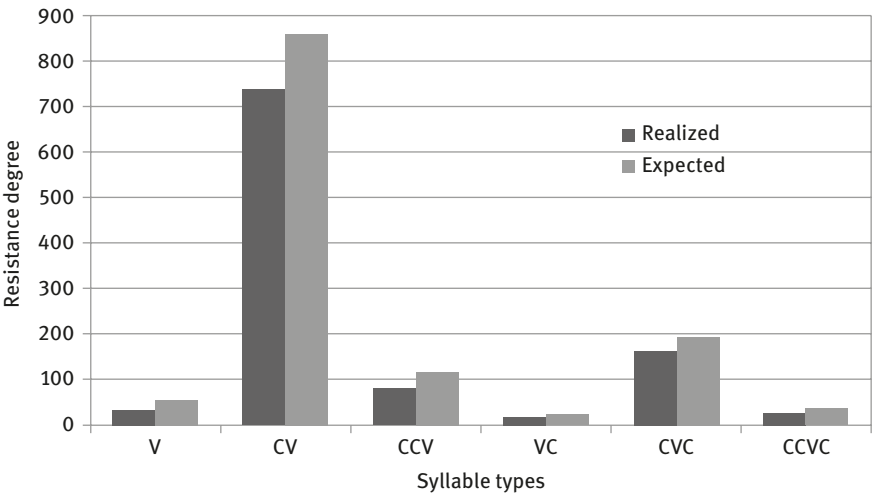

Figure 7.11: Reduction impact on syllable structures in Italian.

the syllable from collapsing. As opposed to vowels, consonants are reduced more rarely: moreover, consonant reduction results in most cases in correspondence syllables, rather than collapses.

Given that the structure of syllables can be very variable, we investigated which clusters are more capable of resistance against reduction processes. Using the results of the Praat script generating syllable structures from the manual annotations described in the previous section, in Figure 7.11 we show that CV- and CVC-syllable types are the most widespread as well as the most resistant against reduction.

Table 7.3: Realization statistics for Italian syllables.

	V	CV	CCV	VC	CVC	CCVC	CCCV	CVCC
Realized	35	735	80	16	159	25	0	0
Expected	55	855	116	22	192	34	4	1
Percentage	63.64	85.96	68.97	72.73	82.81	73.53	0	0

Table 7.3 collects the exact percentages. It can be clearly seen that vowels show a lower degree of resistance, while CV and CVC syllables are both the most widespread and safeguarded syllabic types.

7.5.2 The German data

In Figure 7.12, the distribution of data in German is shown. As in the case of Italian, phonetic realizations and phonological expectations do not always correspond, considering that we have 1,098 syllables on the phonological annotation tier and 1,024 on the phonetic.

Once again, reduction processes play a crucial role in connected speech; almost a quarter of the phonetic syllables do not match with phonological expectations. Nevertheless, it must be noticed that, in this case, correspondences are more widespread than collapses, as opposed to in Italian.

If we examine the details about reduced consonants in Figure 7.13, we see that plosives participate in a much more dominant way in reduction processes in German, together with vowels, confirming their central role in coarticulation and reduction phenomena. Fricatives and nasals are reduced quite often, while other consonants are only rarely omitted.

In Figure 7.14, we show the resistance degree of German syllable types. As can be clearly seen, CV and CVC syllables are once again both the most widespread and the most resistant; Table 7.4 gives further details.

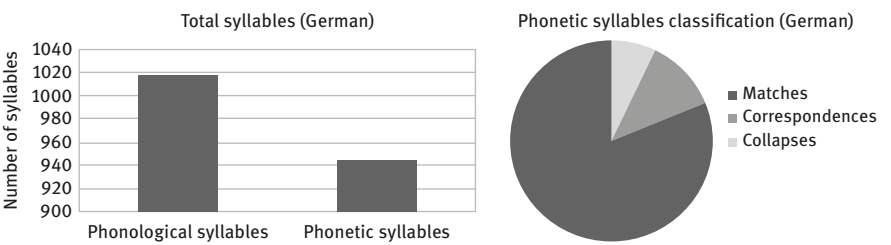

Figure 7.12: German data distributions.

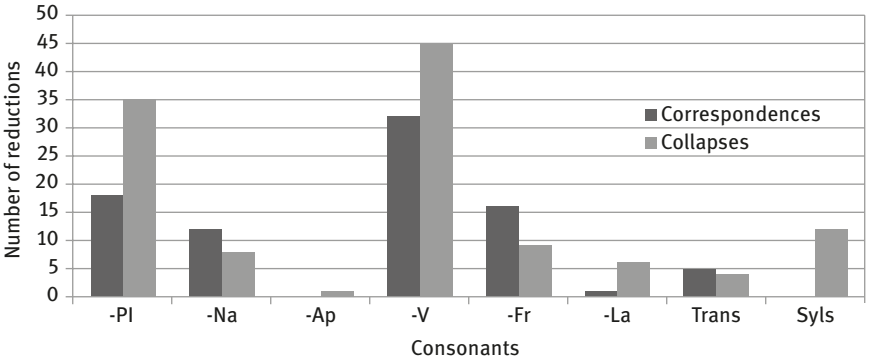

Figure 7.13: Impact of reduction processes on phone classes in German. The report indicates the number of deleted plosives (-Pl), nasals (-Na), approximants (-Ap), vowels (-V), fricatives (-Fr) and laterals (-La). The histogram also reports cases of transformed phones (Trans)

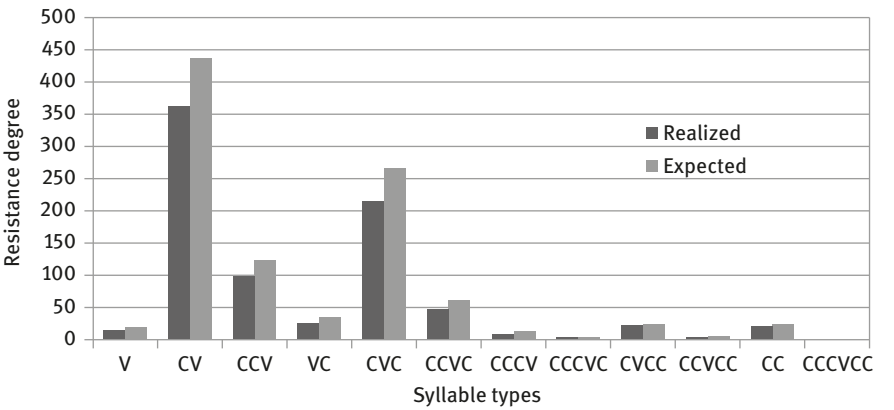

Figure 7.14: Reduction impact on syllable structures in German.

The percentages related to CCCVC, CCVCC and CCCVCC syllables show that these constructions are relatively rare: given that the occurrences are not very frequent (4, 6 or 1 instances), our data are not sufficient to derive a valid interpretation in these cases.

Concerning the other cases, the first thing that can be clearly gathered from this table is the fact that in German many more consonantal clusters are allowed both in the onset and in the coda; as a consequence, the possible syllable types are numerous in comparison with what Italian phonotactics allows; nevertheless, CV and CVC syllables remain the most widespread and also very resistant.

Table 7.4: Realization statistics for German syllables. Realized syllables (R), expected syllables (E) and percentages (P) are reported.

	V	CV	CCV	VC	CVC	CCVC	CCCV	CCCVC	CVCC	CCVCC	CC	CCCVCC
R	16	361	98	25	216	48	9	4	23	4	22	1
E	19	437	124	36	266	62	13	4	25	6	24	1
P	84.2	82.6	79	69.4	81.2	77.4	69.2	100	92	66.7	91.7	100

However, the biggest difference lies in the V syllables: they are not very common (19 instances), but at the same time they are almost always realized.

Another interesting observation concerns the behavior of CC syllables: as we have already stated, sonorants in German can build the nuclear part of syllables: a syllable can be composed by consonants only. This pattern seems to be very resistant: CC syllables are almost always safeguarded from reduction processes (91.67%).

7.5.3 Cross-linguistic comparison

The percentage of disagreement is similar in both languages: about 20% of phonetically realized syllables do not correspond to phonological expectations. Nevertheless, mismatches are qualitatively different: in Italian we found more collapses than correspondences, while in German the tendency is the opposite.

Table 7.5 shows that Italian and German have rather different qualitative properties for what mismatches concern: Italian mismatches often cause resyllabification – a restructuring of adjacent syllables – whereas in German it is the internal syllable structure that undergoes most of the reduction processes. This difference suggests that the rhythm of an Italian sentence is re-planned more frequently than in its German counterpart, because syllables are often completely deleted. In German, on the contrary, the syllabic sequence is rarely restructured in connected speech, while the internal syllabic structure (i.e. phones) is more often reduced.

Table 7.5: Qualitative differences in the mismatches comparison between Italian and German.

	Mismatches	Correspondences	Collapses
Italian	230	85 (37%)	145 (63%)
German	193	120 (62%)	73 (38%)

The last point is confirmed by an additional peculiarity of the Italian data. As we have seen in Figures 7.10 and 7.14, an entire syllable is deleted more often than a vowel in Italian, contrary to what occurs in German. Moreover, German vowels can also be deleted in syllables marked as correspondences: the syllable structure is thus often the target of German reduction processes, regardless of whether it involves consonants or vowels.

Concerning the degree of resistance, both Italian and German CV and CVC syllables proved to be the most widespread and resistant to reduction. In German, however, many more consonantal clusters are allowed in the onset and in the coda. This is consistent with what was stated in Greenberg (1999) concerning the resistance to reduction of complex syllable onsets.

It seems then that coarticulation processes are more widespread in German, above all with reference to vowels and unstressed syllables; the Italian speech chain, on the contrary, often undergoes elision and deletion: coarticulation appears to be less widespread.

7.5.3 Lexical stress and reduction

To enrich our analysis, we now inspect the relationship between lexical stress and reduction processes.

As we can see in Table 7.6, both stressed and unstressed syllables can undergo reduction processes; nevertheless, there are interesting and meaningful differences. Most of the reductions for Italian take place in lexical words' unstressed syllables, while in German weak forms are most often involved. However, given that weak forms are reduced only when unstressed (see Kohler 1990), we could combine them with unstressed syllables, obtaining Table 7.7.

If we do this, we can then conclude that German has the tendency to reduce unstressed syllables, whereas in Italian these tendency is less defined. These two languages should not be seen as opposites, but rather as falling on a continuum.

Table 7.6: Mismatches related to stressed/unstressed syllables.

	Italian	German
Reduction processes	241	195
Stressed syllables	85 (35%)	45 (23%)
Unstressed syllables	106 (44%)	46 (24%)
Weak forms	50 (21%)	104 (53%)

Table 7.7: Mismatches related to stressed/unstressed syllables when considering weak forms in German as unstressed.

	Stressed (%)	Unstressed (%)
Italian	35	65
German	23	77

7.6 Discussion

In this section, we will make a detailed comparison of the relationship between phonological predictions and phonetic realization in German and Italian, with specific reference to reduction processes. Moreover, we will interpret the behavior of German and Italian syllables from an empirical perspective.

7.6.1 Relationship between phonological and phonetic syllables

About 20% of the phonologically expected syllables are reduced or deleted in connected speech, both in German and in Italian; the reasons for these reductions, however, can be different. In general, it can be said that reduction processes usually correspond to the physiological need to coarticulate phones in order to decrease the articulatory effort. This need results in the deletion or reduction of phonetic material, thus invalidating phonological expectations about syllabic structures. Indeed, unforeseeable changes are found both in the syllabic and segmental structures. Our data seems to indicate that German syllables are somehow more resistant to reduction: the percentage of realized units is approximately 70%, at least, whereas in Italian it can sink to almost 60%. Appearances, however, can be deceptive: the big difference between Italian and German is that in the latter, reduction processes are so widespread that they have apparently been included in the phonological rules of the language. Examples of this are: (1) [ɐ] and [ə] cannot be stressed, (2) the accent in native words is always put on the first syllable and (3) sonorants can build a syllabic nucleus. These situations cause a wide incidence of reduction, namely the deletion of the non-accented vowel in disyllabic words. This phenomenon almost always takes place with conjugated verb forms, which are somehow predictable and for this reason easily reduced. As a consequence, verb forms such as *kommen, fliegen, lieben* 'come, fly, love' are articulated as ['kɔmn], ['fli:gn] and ['li:bn], respectively.

Reduction processes in German, therefore, have been recognized and accepted as part of the phonology; Kohler (1974, 1977, 1990) examines four different processes that are essential in German connected speech:

- /r/-vocalization
- weak forms
- elision
- assimilation

In reference to /r/-vocalization, we have already said that /r/ can be pronounced as [ɐ] when it appears alone in the coda; the pronunciation [ɐ] for words ending in -er is meanwhile considered normal. For this reason, we have treated this reduction as expected deletion, and as a consequence, there were no differences in the annotation between the phonological and the phonetic levels. Concerning weak forms, these words are significantly different from other lexical categories, because they can be realized in different reduced forms depending on their position in the sentence (Selkirk 1996).

However, the problem with weak forms is that their behaviour is quite hard to predict, precisely because they can have different reduced forms depending on the context. Actually, they build a special category that is very closely linked to reduction processes. The degree of reduction of function words, for example, correlates with the communicative situation and with the stylistic level (Kohler 1990).

Reduced forms, then, depend on the context and cannot be foreseen; as a consequence, it is not possible to predict when and how a function word will be reduced. Consequently, we decided to accept only completely foreseeable forms on the phonological annotation level, for example, the realization [m] of the article *dem* 'the' (dat.) when it appears after a preposition like *zu* 'to': the merging of these words is accepted in every German grammar, as Kohler (1990) reports.

With respect to elision, German native words have a tendency to be monosyllabic: the elision of unstressed schwa in the possible second syllable can be seen as a validation of this inclination.

Another form of elision is the reduction of geminates to simple consonants, a typical phenomenon in German: the only case in which a geminate consonant is not reduced is when the elision would lead to the complete deletion of a function word, as for example in the word 98 [axtnnɔɪntsɪç] (achtundneunzig) where the [nn] cannot be reduced because it would coincide with the word indicating the sum 8.90 [axtnɔɪntsɪç] acht neunzig (see Kohler 1990).

Another elision phenomenon in German is the loss of aspiration: every plosive that appears in the coda before another plosive regularly loses this feature. Actually, this phenomenon is not so relevant, because the final plosive aspiration is already weaker than in the initial position; it is important only because through this process geminates are created, which can be then further reduced.

In our corpus glottal stops are also frequently reduced or elided, probably because of the increased articulatory effort they entail.

We have considered assimilation as a predictable reduction phenomenon, thus also reporting it in the phonological annotation. Many phones undergo assimilation in German, such as nasal consonants, for which the existence of an archiphoneme /N/ can be postulated; the same is also the case for [s] and [z] and the archiphoneme /S/. Nevertheless, some types of assimilation are not foreseeable, for example when two juxtaposed plosives merge together in a context in which the correct realization of both of them is expected. For this reason, we evaluated every case individually when taking assimilation into account.

In Italian the assessment was easier, because reduction processes are usually not phonologically expected: this fact can also explain why the percentage of agreement between phonological and phonetic levels is higher in German than in Italian.

7.7 Conclusions

In this work, we have presented an analysis of reduction processes in connected speech, comparing German and Italian; we have compared a set of phonologically predicted syllables with their phonetically realized counterparts in order to understand what type of syllabification better reflects reductions. Now, we describe the outcome of our analysis.

The degree of correspondence between phonologically predicted and phonetically realized syllables is higher in German (93.35%) than in Italian (88.56%): for this reason, we argue that German syllabification is more easily predictable on the base of its phonological rules than Italian syllabification.

Concerning reduction processes linked with speech rhythm and syllabification, many questions still remain unanswered; however, our empirical work shows that, despite our articulatory perspective for the phonological annotation, the phonological and the phonetic annotation levels are significantly different because of reduction processes. These phenomena have different connotations in Italian and German: in Italian it appears to be more difficult to coarticulate phonemes (collapses are much more widespread), that is to say, it is more simple to delete the entire syllable than to reduce the internal structure. In German, on the contrary, it seems that the degree of coarticulation can be very high: we find even syllables in which vowels and sonorants are assimilated and the consonant can assume the role of syllabic nucleus.

Concerning stress, in both languages reductions are mainly linked to unstressed syllables, as expected. German, however, has shown the tendency to strongly reduce weak forms and unstressed syllables: reduction processes are thus modeled on the rhythmic patterns of the speech chain.

As discussed in Sections 7.2 and 7.3.1, this kind of investigation impacts on technological approaches to speech processing as it involves the definition of a unit of analysis, the syllable, that can be used as a link between linguistics and computer science. Modeling the relationship existing between expected syllabic structures and the alterations that can be observed in the signal may prove invaluable to automatic speech processing systems as they may find solid theoretical grounds in linguistic research. As a final note, we would like to point out that the effort toward shared processing models should be twofold. As we highlighted in Section 7.3.1, in fact, systematic errors by automatic syllabification algorithms considering just the energy profile to detect syllable boundaries may be corrected by using phonetic models taking into account the full range of spectral changes that introduce a boundary.

To summarize, even from the technological point of view a deeper investigation on the concept of syllable is necessary to clarify the acoustic profile of this unit. While in most cases it is possible to predict what will be observed in the signal, unpredictable segmental structures observed in speech are, indeed, organized in an acoustically self-consistent way, described by the SSP. While the majority of the cases fall under predictable phenomena, the portion of unpredicted cases we found is still significant both in Italian and in German. The role of the phonetic syllable, in this sense, may be important because it would constitute a unit that would not be considered a mis-product of the articulation process but a self-existing, acoustically defined, element. On the other hand, the description of the phonetic syllable should be improved because, as we reported in Section 7.3.1, an entire class of sounds, specifically nasal sounds, eludes an acoustic description based on energy movements only.

Technological approaches, in this kind of research, also find application in investigating rhythmic planning strategies in speech and their influence on message interpretation. Being completely independent from the semantic content of the utterances and relying only on acoustic parameters, they provide an objective representation that can be used to investigate, for example, prominence patterns. By improving acoustic models of both syllables and prominence patterns, results obtained by automatic approaches may help researchers working in the field of linguistics improve theoretical models related to these phenomena. Strategies used by speakers to emphasize or de-emphasize specific units may be investigated with these methods. In this view, unexpected alterations found in the speech chain may be used to provide important information for speech interpretation, since they would act as a counterpart for prominent units. Specifically, they may be considered an explicit signal of a lower importance of the involved unit during the decoding/interpretation phase with respect to expected or emphasized units. This means that reduced forms should not be considered aberrations or *abnormal* expressions of speaker behaviour but as the most natural form of

speech coding. If this position can be accepted, the natural consequence is that phoneticians should not continue to fight against deviations when they study spontaneous speech; on the contrary, they should reevaluate the importance of some studies based on artificial, laboratory speech.

References

Bell, Alan & Joan Bybee Hooper 1978. Issues and evidence in syllabic phonology. In Alan Bell & J. Bybee Hooper (eds.), *Syllables and segments*, 3–22. Amsterdam: North-Holland publishing company.

Bertinetto, Pier Marco 1999. La sillabazione dei nessi /sc/ in italiano: un'eccezione alla tendenza 'universale'. In Paola Benincà, Laura Vanelli & Laura Mioni (eds.), *Fonologia e morfologia dell'italiano e dei dialetti d'Italia: Atti del XXXI Congresso Internazionale di Studi della Società di Linguistica Italiana*, 71–96. Roma: Bulzoni.

Bertinetto, Pier Marco 2010. Fonetica italiana. Quaderni del Laboratorio di Linguistica della Scuola *Normale Superiore* 9 (1). 1–30. http://linguistica.sns.it/QLL/QLL10/Bertinetto_Fonetica_italiana.pdf (accessed 27 October 2016).

Bigi, Brigitte, Christine Meunier, Irina Nesterenko & Roxane Bertrand 2010. Automatic detection of syllable boundaries in spontaneous speech. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, 3285–3292. Paris: European Language Resources Association

Boersma, Paul & David Weenink 2011. Praat: Doing phonetics by computer [computer program].

Brunetti, Lisa, Stefan Bott, Joan Costa, Enric Vallduvì 2009. Nocando: A multilingual annotated corpus for the study of information structure. *Paper presented at the 5th Corpus Linguistics Conference*, University of Liverpool, 20–23 July.

Brunetti, Lisa, Stefan Bott, Joan Costa & Enric Vallduvì 2011. A multilingual annotated corpus for the study of information structure. In Marek Konopka, Jacqueline Kubczak, Christian Mair, František Štícha & Ulrich H. Waβner (eds.), *Grammar and Corpora 2009: Third International Conference*, *2011*, 305–327. Tübingen: Narr Verlag.

Clements, George N. 2009. Does sonority have a phonetic basis. In Eric Raimy & Charles E. Cairns (eds.), *Contemporary views on architecture and representations in phonological theory*, 165–175. Cambridge and London: MIT Press.

D'Alessandro, Christophe & Piet Mertens 1995. Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language* 9 (3). 257–288.

De Saussure, Ferdinand 2015 [1967]. *Corso di linguistica generale: introduzione, traduzione e commento di Tullio De Mauro*. Bari: Laterza.

Eisenberg, Peter 2013 [1998]. *Grundriß der deutschen grammatik: Band 1: Das Wort*. Stuttgart: Metzler.

Farnetani, Edda & Daniel Recasens 1997. Coarticulation and connected speech processes. In William J. Hardcastle & John. Laver (eds.), *The handbook of phonetic sciences*, 371–404. Oxford: Wiley.

Feth, Lawrence Lee 1972. Combinations of amplitude and frequency differences in auditory discrimination. *Acustica* 26. 67–77.

Ghosh, Prasanta Kumar & Shrikanth Narayanan 2009. Pitch contour stylization using an optimal piecewise polynomial approximation. IEEE Signal Processing Letters 16 (9). 810–813.

Gómez, David Maximiliano, Iris Berent, Silvia Benavides-Varela, Ricardo A. H. Bion, Luigi Cattarossi, Marina Nespor and Jacques Mehler 2014. Language universals at birth. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 111 (16). 5837–5841.

Greenberg, Steven 1999. Speaking in shorthand–a syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29 (2). 159–176.

t'Hart, Johan, René Collier & Antonie Cohen 1990. A perceptual study of intonation: An experimental-phonetic approach. Cambridge: Cambridge University Press.

Hirst, Daniel & Robert Espesser 1993. Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix-en-Provence* 15. 75–85.

House, D. 1990. Tonal perception in speech. Lund: Lund University Press

House, David 1996. Differential perception of tonal contours through the syllable. In: H. Timothy Bunnell & William Idsardi (eds.), *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP)*, 2048–2051. New Castle: Citation Delaware.

Jespersen, Otto 1920. Lehrbuch der Phonetik. Leipzig and Berlin: B.G. Teubner.

Jones, Rhys James, Simon Downey and John S. D. Mason 1997. Continuous speech recognition using syllables. In George Kokkinakis, Nikos Fakotakis, Evangelos Dermatas (eds.), *Proceedings of the Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, 1171–1174. http://www.isca-speech.org/archive/archive_papers/eurospeech_1997/e97_1171.pdf (accessed 27 October 2016)

Kohler, Klaus J. 1974. Koartikulation und steuerung im deutschen. In Paul Grebe & Ulrich Engel (eds.), *Sprachsystem und Sprachgebrauch. Festschrift für Hugo Moser*, Teil 1, 172–192. Düsseldorf: Schwann.

Kohler, Klaus J. 1977. Investigating coarticulation. In Wolfgang U. Dressler & Oskar E. Pfeiffer (eds.), Phonologica 1976, Akten der dritten Internationalen Phonologie-Tagung (Innsbrucker Beiträge zur Sprachwissenschaft), 243–247. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck.

Kohler, Klaus J. 1990. Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In: William J. Hardcastle & Alain Marchal (eds.), *Speech production and speech modelling*, 69–92. Dordrecht: Kluwer Academic Publishers.

Kohler, Klaus J. 1998. The phonetic manifestation of words in spontaneous speech. In Danielle Duez (ed.), *Sound patterns of spontaneous speech*, 13–22, ISCA Archive. http://www.isca-speech.org/archive_open/archive_papers/sposs/spos_013.pdf (accessed January 2018).

Kohler, Klaus J. & Johnatan Rodgers 2001. Schwa deletion in German read and spontaneous speech. In Klaus J. Kohler (ed.), *Sound patterns in German read and spontaneous speech: Symbolic structures and gestural dynamics (AIPUK 35)*, 97–123. Kiel: Institut für Phonetik und digitale Sprachverarbeitung.

Kozhevnikov, Valeriĭ Aleksandrovich & Li͡udmila Andreevna Chistovich 1965. *Speech: Articulation and perception*. Washington DC: U.S. Joint Publications Research Service.

Laver, John 1994. *Principles of phonetics*. Cambridge: Cambridge University Press.

Lindblom, Björn 1990. Explaining phonetic variation: A sketch of the h&h theory. In William J. Hardcastle & Alain Marchal (eds.), *Speech production and speech modelling*, 403–439. Dordrecht: Kluwer Academic Publishers.

Lindblom, Björn 1996. Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America* 99 (3). 1683–1692.

Loporcaro, Michele 1996. On the analysis of geminates in Standard Italian and Italian dialects. In: Bernhard Hurch and Richard Rhodes (eds.). *Natural Phonology: The State of the Art.*

Papers from the Bern Workshop on Natural Phonology, 153–187. New York and Amsterdam: Mouton de Gruyter [preprint in Quaderni del Laboratorio di Linguistica della Scuola Normale Superiore di Pisa 8: 125–147].

Ludusan, Ioan Bogdan 2010. Beyond short units in speech recognition: A syllable-centric and prominence-based approach. Napoli: Università degli Studi di Napoli "Federico II" dissertation.

Maiwald, Di 1967. Ein funktionschema des Geh¨ors zur Beschreibung der Erkennbarkeit kleiner Frequenz- und Amplitudea¨nderungen. Acustica 18. 81–92.

Martin, Philippe 2010. Prominence detection without syllabic segmentation. In *Proceedings of Speech Prosody* [online only]. http://speechprosody2010.illinois.edu/papers/102009.pdf

Mermelstein, Paul 1975. Automatic segmentation of speech into syllabic" units. Journal of the Acoustical Society of America 58 (4). 880–883

Mertens, Piet 2004. The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In Bernard Bel & Isabelle Marlien, *Speech Prosody* 2004, 549–552. [ISCA Archive]. http://www.isca-speech.org/archive/sp2004/papers/sp04_549.pdf (accessed January 2018).

Mohammadi, Gelareh, Antonio Origlia, Maurizio Filippone & Alessandro Vinciarelli 2012. From speech to personality: mapping voice quality and intonation into personality differences. In: Noboru Babaguchi, Kiyoharu Aizawa, John Smith (eds.), *MM'12 – ACM Multimedia Conference*, 789–792. New York: ACM.

Møller, Aage R. 1974. Dynamic properties of cochlear nucleus units in response to excitatory and inhibitory tones. In Eberhard Zwicker & Ernst Terhardt (eds.), *Facts and models in hearing*, 227–240. Belin Heidelberg New York: Springer Verlag.

Muljačić, Žarko 1972 [1969]. *Fonologia della lingua italiana*. Bologna: Il Mulino.

Nam, Hosung, Louis Goldstein & Elliot Saltzman 2009. Self-organization of syllable structure: A coupled oscillator model. In François Pellegrino, Egidio Marsico, Ioana Chitoran, Christophe Coupé (eds.), *Approaches to phonological complexity*, 297–328. Berlin and New York: Walter de Gruyter.

Nespor, Marina 1993. *Fonologia*. Bologna: Il Mulino.

Origlia, Antonio & Iolanda Alfano 2012. Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabification. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, 997–1002. Istanbul: European Language Resources Association (ELRA).

Origlia, Antonio, Giovanni Abete & Francesco Cutugno 2013. A dynamic tonal perception model for optimal pitch stylization. *Computer Speech and Language* 27 (1). 190–208.

Origlia, Antonio & Francesco Cutugno, F. 2014. A simplified version of the OpS algorithm for pitch stylization. In Nick Campbell, Dafydd Gibbon and Daniel Hirst (eds.), *Social and Linguistic Speech Prosody: Proceedings of the 7th International Conference on Speech Prosody*, 992–996. Dublin: Science Foundation Ireland.

Origlia, Antonio & Valentina Schettino 2014. Automatically detecting syllables: a two-way bridge between linguistics and technology. In Antonio Romano, Matteo Rivoira, Ilario Meandri (eds.), *Aspetti prosodici e testuali del raccontare: dalla letteratura orale al parlato dei media*, 293–304. Alessandria: Dell'Orso.

Origlia, Antonio, Francesco Cutugno & Vincenzo Galatà 2014. Continuous emotion recognition with phonetic syllables. *Speech Communication* 57. 155–169.

Origlia, Antonio, Vincenzo Galatà & Francesco Cutugno 2015. Introducing context in syllable based emotion tracking. In Péter Baranyi, Adam Csapo & Gyula Sallai (eds.), *Cognitive Infocommunications (CogInfoCom)*, 337–342. Heidelberg Berlin: Springer International Publishing Switzerland.

Petrillo, Massimo & Francesco Cutugno 2003. A syllable segmentation algorithm for English and Italian. In *Proceedings of Eurospeech 2003*, 2913–2916. [online only, ISCA Archive] http://www.isca-speech.org/archive/archive_papers/eurospeech_2003/e03_2913.pdf (accessed January 2018).

Pike, Kenneth Lee 1945. *The intonation of American English*. Ann Arbor: University of Michigan Press.

Ravuri, Suman and Daniel P. W. Ellis 2008. Stylization of pitch with syllable-based linear segments. In: IEEE [ISCA Archive], *International Conference on Acoustics, Speech,and Signal Processing: Proceedings*, 3985–3988. [online only] https://core.ac.uk/download/pdf/27293035.pdf (accessed January 2018).

Roach, Peter 2000 [1983]. English phonetics and phonology: A practical course. Cambridge: Cambridge University Press.

Rossi, Mario 1978. Interactions of intensity glides and frequency glissandos. *Language and Speech* 21 (4). 384–394.

Savy, Renata & Francesco Cutugno 1997. Hypospeech, vowel reduction, centralization: how do they interact in diaphasic variations. In Bernard Caron (ed.), *Proceedings of the XVIth International Congress of Linguists*, 1–13. Oxford: Pergamon-Elsevier.

Selkirk, Elisabeth 1996. The prosodic structure of function words. In James L. Morgan & Katherine Demuth (eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, 187–214. Mahwah: Lawrence Erlbaum Associates.

Stetson, Raymond Herbert 1951. *Motor phonetics : A study of speech movements in action*. Amsterdam: North Holland.

Vennemann, Theo 1982. Zur Silbenstruktur der deutschen Standardsprache. In Theo Venneman & Deutsche Gesellschaft für Sprachwissenschaft (eds.), *Silben, Segmente, Akzente*, 261–305. Berlin: Mouton de Gruyter.

Wang, Dagen & Shrikanth Narayanan 2005. Piecewise linear stylization of pitch via wavelet analysis. In ISCA Archive [online only], *Proceedings of the European Conference on Speech Communication and Technology*, 1–4. http://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_3277.pdf (accessed January 2018).

Wiese, Richard 1991. Was ist extrasilbisch im deutschen und warum? *Zeitschrift für Sprachwissenschaft* 10 (1). 112–133.

Wu, Su-Lin, Michael L. Shire, Steven Greenberg & Nelson Morgan 1997. Integrating syllable boundary information into speech recognition. In IEEE (ed.) [online only], *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 987–990. https://www.researchgate.net/profile/Steven_Greenberg5/publication/3693504_Integrating_syllable_boundary_information_into_speech_recognition/links/09e4151114444e226e000000.pdf (accessed January 2018).

Wypych, Mikołaj 2006. Automatic pitch stylization enhanced by top-down processing. In Rudiger Hoffmann & Hansjörg Mixdorff (eds.), *Speech Prosody 2006*, ISCA Archive [online only], http://www.isca-speech.org/archive/sp2006/papers/sp06_217.pdf (accessed January 2018).

Zwicker, Eberhard 1962. Direct comparisons between the sensations produced by frequency modulation and amplitude modulation. *Journal of the Acoustical Society of America* 34. 1425–1430.

Carol Espy-Wilson, Mark Tiede, Vikramjit Mitra, Ganesh Sivaraman, Elliot Saltzman and Louis Goldstein

8 Speech inversion using naturally spoken data

Abstract: Speech acoustic patterns vary significantly as a result of coarticulation and lenition processes that are shaped by segmental context or by performance factors such as production rate and degree of casualness/precision. Such processes have the most dramatic effect on the acoustic properties of the speech signal that relate to manner and place of articulation, and the resultant acoustic variability continues to offer serious challenges for the development of automatic speech recognition (ASR) systems that can perform well with minimal constraints. For example, conventional ASR systems attempt to account for coarticulatory effects through tri- or quin-phone and cross-word models; however, it is inherently difficult to quantify a fixed scope for coarticulatory effects. Articulatory phonology (AP) provides a unified framework for understanding how spatiotemporal changes in the pattern of underlying speech gestures can lead to corresponding changes in the extent of intergestural temporal overlap and in the degree of gestural spatial reduction; in turn, these changes in overlap and reduction create acoustic consequences that are typically reported as assimilations, insertions, deletions and substitutions. We have made important progress in developing a speech inversion (SI) system based on a computational model of AP and have shown that such a system can greatly improve the robustness of ASR systems to noise. These encouraging results have been obtained even though we have had to make use of synthetically generated speech and articulatory data to develop our SI system, as there are to date no natural speech databases with the kind of articulatory annotations needed.

Keywords: speech recognition, speech inversion, acoustic-to-articulatory mapping, coarticulation, lenition, articulatory phonology, task-dynamics, articulatory gestures, vocal tract variables

Carol Espy-Wilson, University of Maryland
Mark Tiede, Haskins Laboratories
Vikramjit Mitra, Speech Technology and Research Lab
Ganesh Sivaraman, University of Maryland
Elliot Saltzman, Boston University
Louis Goldstein, University of Southern California

https://doi.org/10.1515/9783110524178-008

In this study, we focus on the capabilities of the SI system for modeling changes in temporal overlap and spatial magnitude of gestures. Specifically, we address the following questions: (1) With proper contextualization, can our SI system uncover gestures "hidden" acoustically by increases in overlap (coarticulation) and/or decreases in magnitude (lenition)?; (2) Is undershoot of articulatory targets accurately reflected in the output of the SI system?; (3) Will the use of naturally spoken data (e.g., concurrently recorded speech acoustics and kinematics) for training the SI system result in AP gestural trajectories that accurately reflect articulatory movements during and between gestures?; and finally (4) What is the best methodology for training the SI system with naturally spoken speech and articulatory data? The implications of successfully answering these questions are significant since, if our SI system is able to "uncover" seemingly hidden gestures, then the robustness and accuracy of ASR systems will be vastly improved. Furthermore, such results will also provide the means for improving a variety of speech applications and leading, for example, to the strengthening of speech pronunciation tools in the classroom and clinic, and to the development of more natural sounding articulatory speech synthesizer that will better reflect idiosyncratic individual differences between speakers.

8.1 Introduction

Speech acoustic patterns vary significantly as a result of coarticulation and lenition processes that are shaped by segmental context or by performance factors such as production rate and degree of casualness/precision. Such processes have the most dramatic effect on the acoustic properties of the speech signal that relate to manner and place of articulation, and the resultant acoustic variability continues to offer serious challenges for the development of automatic speech recognition (ASR))systems that can perform well with minimal constraints. For example, conventional ASR systems attempt to account for coarticulatory effects through tri- or quin-phone and cross-word models; however, it is inherently difficult to quantify a fixed scope for coarticulatory effects. Articulatory phonology (AP) provides a unified framework for understanding how spatiotemporal changes in the pattern of underlying speech gestures can lead to corresponding changes in the extent of intergestural temporal overlap and in the degree of gestural spatial reduction; in turn, these changes in overlap and reduction create acoustic consequences that are typically reported as assimilations, insertions, deletions and substitutions. We have made important progress in developing a speech inversion (SI) system based on a computational model of AP and have shown that

such a system can greatly improve the robustness of ASR systems to noise. These encouraging results have been obtained even though we have had to make use of synthetically generated speech and articulatory data to develop our SI system, as there are to date no natural speech databases with the kind of articulatory annotations needed.

In this chapter, we focus on the capabilities of the SI system for modeling changes in temporal overlap and spatial magnitude of gestures. Specifically, we address the following questions: (1) With proper contextualization, can our SI system uncover gestures "hidden" acoustically by increases in overlap (coarticulation) and/or decreases in magnitude (lenition)?; (2) Is undershoot of articulatory targets accurately reflected in the output of the SI system?; (3) Will the use of naturally spoken data (e.g., concurrently recorded speech acoustics and kinematics) for training the SI system result in AP gestural trajectories that accurately reflect articulatory movements during and between gestures?; and finally (4) What is the best methodology for training the SI system with naturally spoken speech and articulatory data? The implications of successfully answering these questions are significant since, if our SI system is able to "uncover" seemingly hidden gestures, then the robustness and accuracy of ASR systems will be vastly improved. Furthermore, such results will also provide the means for improving a variety of speech applications and leading, for example, to the strengthening of speech pronunciation tools in the classroom and clinic, and to the development of more natural sounding articulatory speech synthesizer that will better reflect idiosyncratic individual differences between speakers.

8.1.1 AP and the task-dynamic model of speech production

The conceptual framework of our modeling efforts is provided by AP (Browman and Goldstein 1988, 1989, 1992; Nam et al. 2004), which views speech as a spatiotemporal constellation of vocal-tract constriction actions (e.g., lip closure for /b/, tongue tip closure for /d/) called *gestures*. Each gesture in a lexical entry is defined as a critically damped second-order dynamical system with its own set of invariant, context-independent dynamic parameters (constriction target, stiffness and damping), that controls one of the constricting organs (end-effectors) of the vocal tract: lip aperture (LA), lip protusion (LP), tongue tip (TT), tongue body (TB), velum (VEL) and glottis (GLO). The gestural dynamic parameters are crucial in distinguishing utterances in a gesture-based lexicon, for example, gestural stiffness distinguishes consonants from vowels, since articulatory motions for consonants, which are parameterized as gestures with higher stiffness, are faster than those of vowels. The gestural targets for the constrictions of these end-effectors are defined

	Tract variable	Articulators involved
LP	Lip protrusion	Upper and lower lips, jaw
LA	Lip aperture	Upper and lower lips, jaw
TTCL	Tongue tip constrict location	Tongue tip, tongue body, jaw
TTCD	Tongue tip constrict degree	Tongue tip, tongue body, jaw
TBCD	Tongue body constrict location	Tongue body, jaw
TBCD	Tongue body constrict degree	Tongue body, jaw
VEL	Velic aperture	Velum
GLO	Glottal aperture	Glottis

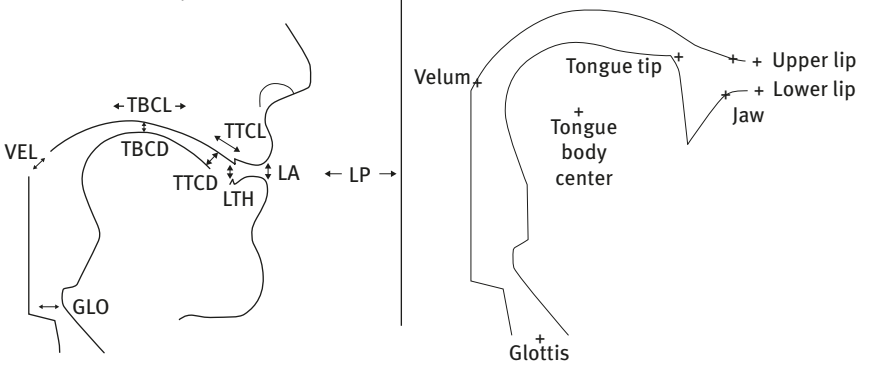

Figure 8.1: Tract variables (left); model articulators (right).

in a set of *tract variables* (TVs; Figure 8.1, left column), and each TV has its own set of associated *model articulators* (Figure 8.1, right column). TVs define a set of task-space coordinates that are intrinsic to vocal tract geometry, and that characterize vocal tract shape in terms of constriction degrees and locations (Figure 8.1).

The timing of the gestures must be planned so that they unfold in the vocal tract over time with appropriate temporal patterning, consistent with the language's phonology. This temporal plan cannot be as simple as triggering phonological segments in sequence, because the stable patterns of gestural triggering that we observe show that gestures composing a single segment can be triggered sequentially and that gestures belonging to a sequence of segments can be triggered simultaneously. The timing plan for a given utterance has been modeled in AP as a network of gestural *planning oscillators* in which each oscillator defines a node and is responsible for triggering the activation of its associated gesture(s), and in which each edge connects a pair of oscillators and defines a coupling function that specifies a target relative phase between those oscillators (Browman and Goldstein 2000; Nam and Saltzman 2003, Nam et al. 2009; Saltzman and Byrd 2000; Saltzman et al. 2008;). Planning networks are represented by *coupling graphs* (see Figure 8.2), and subgraphs representing words define the stored lexical entries of the model.

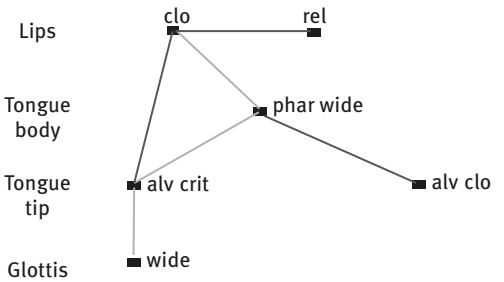

Figure 8.2: Coupling graph for constriction gestures in "spat."

Simulated utterances are generated using TADA (TAsk Dynamic Application; Nam et al. 2004), which is a computational implementation of AP developed at Haskins Laboratories. The TADA model produces speech from text input (orthography or broad phonetic transcription) by first creating a coupling graph for an utterance which is used to parameterize a corresponding planning oscillator network. When the oscillators settle into a pattern of stable relative phases, a set of gestural *activation* waves are triggered that define the utterance's *gestural score* (Figure 8.3); each activation wave component of the gestural score reflects the strength with which its associated gesture (e.g., lip closure) "attempts" to shape vocal tract movements at any given point in time. The gestural score is input to the task-dynamic model of motor control and coordination of Saltzman and Munhall (1989), which produces the utterance's set of TV (Figure 8.4, bottom four rows) and model-articulator kinematic trajectories. These trajectories, together with sound sources, are input to HLsyn™, a parametric quasi-articulatory synthesizer (Sensimetrics Inc.; Hanson and Stevens 2002) to produce synthetic speech (Figure 8.4, top row). A summary flowchart of the speech production model is shown in Figure 8.5.

Quantitative changes in the temporal patterning of the gestures as a function of prosody (phrasing and stress/accent) and speech rate can produce extreme,

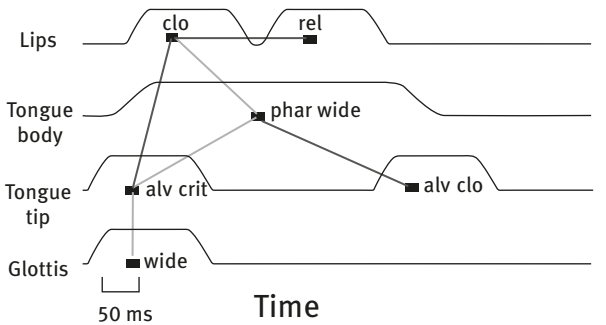

Figure 8.3: Gestural scores generated by TADA from the coupling graph shown in Figure 8.2.

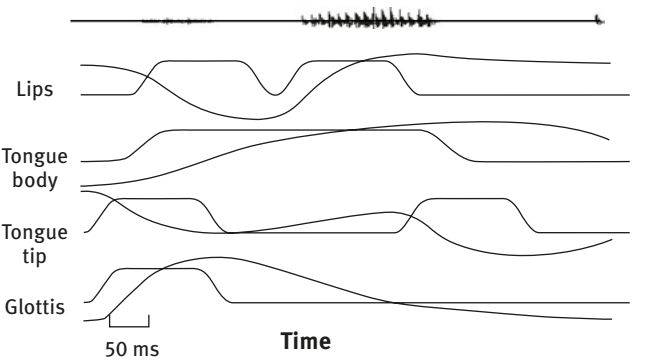

Figure 8.4: Tract variables, gestural scores and synthetic speech generated by TADA.

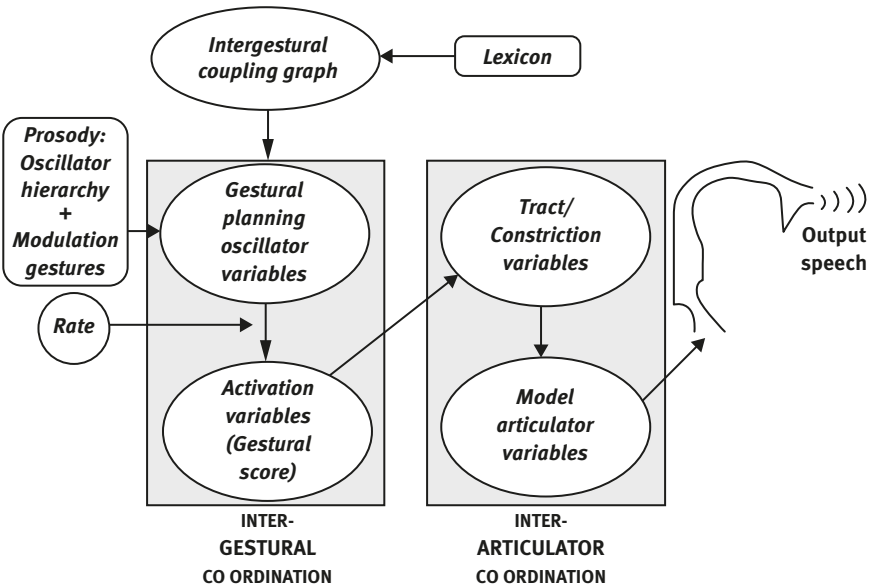

Figure 8.5: Organization of the task-dynamic model.

apparently qualitative changes in the resulting sound wave. Such changes include reduction/expansion of the temporal (and spatial) extent of gestures and changes in the temporal lag between neighboring gestures. For example, in unstressed and unaccented contexts that are remote from phrase boundaries, temporal lags can decrease so as to create new instances of gestural overlap. An example is shown in Figure 8.6. The sequence "seven plus" is spoken carefully on the left, with a prosodic boundary between the words, and in a fast conversational style on the right, without any boundary. As the transcriptions indicate, the last syllable

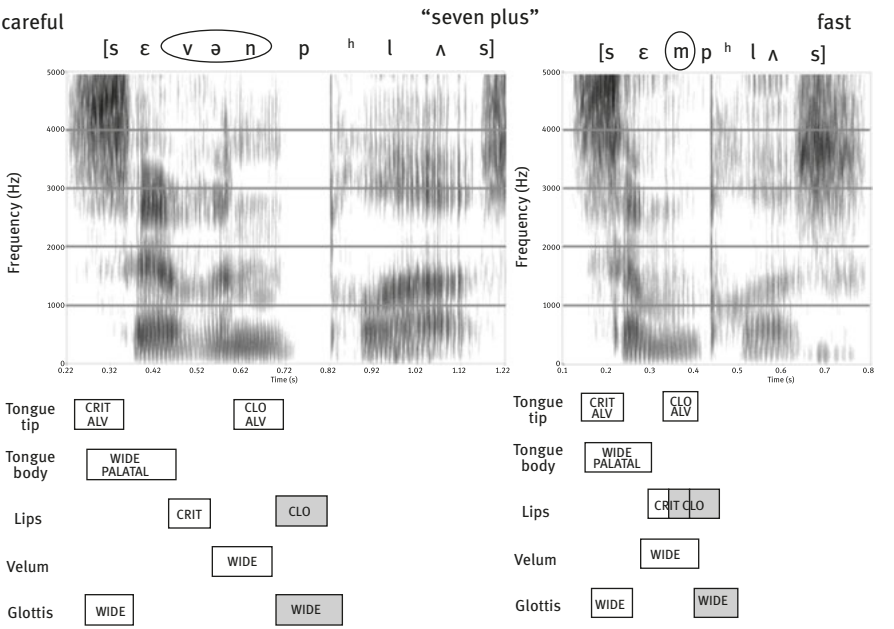

Figure 8.6: "seven plus" spoken carefully (left) and fast (right). Top: phonetic transcriptions; Middle: spectrograms; Bottom: schematic gestural score.

of "seven" has apparently been replaced in the fast form by [m]. However, the gestural scores (approximated from articulatory data in similar examples) show that all of the same gestures are produced in both instances. *This persistence property of gestures is why coarticulation is better modeled in a gesture-based recognition system*. In the fast form, the initial lip closure gesture for "plus" (shown in gray) now overlaps the lip gesture for the /v/ in "seven," and blends with it. The lips do not open between them, so the reduced vowel in that syllable is elided, and the TT gesture is acoustically obscured by the lip gestures.

In fact, the TADA model behaves in exactly this way (i.e., displaying increased intergestural overlap or "sliding," and kinematic reduction of gestures that are acoustically hidden by such sliding) when speech rate is increased by globally scaling the frequencies of all the oscillators in the planning oscillator ensemble (see Figure 8.5). The reason for this behavior is that such frequency scaling shortens the durations of gestural activation waves, which creates kinematic undershoot, that is, a gesture activated for a shorter time interval falls correspondingly short spatially relative to when it is produced during careful, slower speech. Additionally, such frequency scaling results in a reduced interval between activation onsets of successive gestures. Since the intrinsic dynamic parameters of the gestures remain constant, the time taken to relax back to its post-release position

does not change, and the tail end of a leading gesture can increase its overlap with the rising phase of a final gesture (intergestural "sliding"), and "hide" the acoustic consequences of an intervening (reduced) gesture.

8.1.2 Articulatory information and speech technologies

The use of articulatory information in any speech application necessitates either recording such information directly from the speakers during speech (model-free approach) or inferring such information through a model that takes speech acoustics as input (model-dependent approach). This section will illustrate several methods used to obtain articulatory data, with special emphasis on the representations defined in the AP framework that were discussed in the previous section.

8.1.2.1 Articulatory kinematics

Articulatory trajectories are measured during speech production to provide time-varying positional information for the constricting organs or articulators within the vocal tract. The most direct way to capture such articulatory information is by tracking flesh points associated with different speech articulators and recording their movements while speech is generated. Such flesh-point articulatory trajectories have been exhaustively studied in the literature. Figure 8.7 shows the pellet placements for the University of Wisconsin X-ray microbeam (XRMB) dataset (Westbury 1994).

To use articulatory information in speech processing applications, one has to either observe such trajectories directly while the speaker is producing speech or infer the trajectories indirectly from the speech signal. While direct measurement provides accurate articulatory information, it necessitates placement of tracking sensors within the mouth, requiring specialized equipment and potentially perturbing the speaker's usual production patterns. As an alternative, several studies have attempted to estimate articulatory information from the speech signal, a line of research commonly known as "acoustic-to-articulatory inversion" or simply speech inversion (SI).

8.1.2.2 Speech inversion

SI has been widely researched in the last 35 years. An early approach due to Atal et al. (1978) used four articulatory parameters: length of the vocal tract, distance of the maximum constriction region from the glottis, cross-sectional area at the maximum (i.e., narrowest) constriction region and the area of the mouth

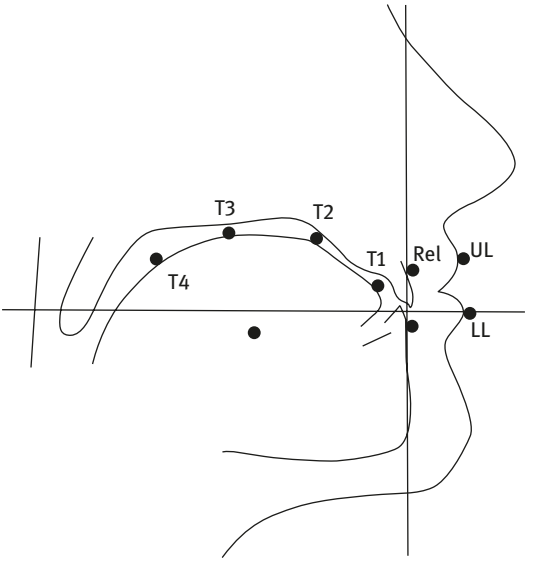

Figure 8.7: Placement of eight pellets along the vocal tract in the XRMB dataset. Also shown are the reference axes for the pellet positions.

opening. Rahim et al. (1991, 1993) used an articulatory synthesis model to generate a database of articulatory-acoustic vector pairs. The acoustic data consisted of 18 Fast-Fourier Transform-derived cepstral coefficients, whereas the articulatory data is composed of 10 vocal tract areas and a nasalization parameter. They trained multi-layered perceptron (MLP) neural networks to map from acoustic data to the vocal tract area functions. The articulatory-acoustic data pairs were obtained by random sampling over the manifold of reasonable vocal tract shapes within the articulatory parameter space of Mermelstein's (1973) articulatory model. However, a limitation of their approach was an inadequate random sampling strategy, as such sampling may select physiologically plausible articulatory configurations that are nonetheless uncommon or unused in typical running speech. To address this, Ouni and Laprie (1999) sampled the articulatory space such that the inversion mapping was piece-wise linearized. Their sampling was based upon the assumption that the articulatory space is contained within a single hypercube, sampling more aggressively in regions where the inversion mapping is complex and less elsewhere.

Use of neural networks for SI has become popular since the pioneering work of Papcun et al. (1992). They used MLPs to perform SI on the XRMB data to obtain the vertical (y-coordinate) motions of three articulators (lower lip, TT and tongue dorsum) for six English stop consonants. Their data were recorded from three male native American English speakers, who uttered six nonsense words. The

words had repeated [-Cə-] syllables, where "C" belonged to one of the six English oral stop consonants /p,b,t,d,k,g/. The MLP architecture was chosen based upon trial-and-error and the optimization of the architecture was based upon minimizing the training time and maximizing the estimation performance. The network was trained using a standard backpropagation algorithm. An important observation noted in their study was that trajectories of articulators considered critical for the production of a given consonant demonstrated higher correlation coefficients than those that were considered noncritical to the production of that consonant. This result was termed the "critical articulator phenomenon," leading them to conclude that for a given consonant, the dynamics of its critical articulator were more constrained than those of the noncritical ones. This observation was further supported by Richmond (2007), whose SI system used mixture density networks (MDN) and showed that the conditional probability density functions (pdf) of critical articulators show less variance compared to noncritical articulators. This work also showed that MDNs tackle the *nonuniqueness* problem of SI more appropriately than other modeling techniques. Nonuniqueness refers to the one-to-many mapping that exists from speech acoustics to vocal tract configurations (i.e., similar acoustic patterns can result from different vocal tract configurations), and is a critical issue in the acoustic-to-articulatory inversion of speech.

8.1.2.3 A first step: applying SI to synthetically generated acoustic and articulatory data

The training of SI systems requires concurrent sets of speech acoustic and articulatory trajectories for the chosen set of training utterances. Because of the difficulties associated with direct observation of the vocal tract during speech, datasets of this type remain sparse and among those that do exist, differences in recording systems and methods make it difficult to generalize across such diverse data sources. Given this difficulty, synthetic datasets have been used (Mitra et al. 2010a, 2014) to provide ground-truth mappings between acoustic and articulatory data for applications such as speech recognition and human emotion recognition. We have produced such synthetic datasets using the Haskins Laboratories TADA model (Nam et al. 2004; see Section 8.1.1). Given English text or ARPABET, TADA generates output in the form of formants and TV time functions. Synthetic speech acoustics is then generated using HLsyn™ (a parametric quasi-articulatory synthesizer developed by Sensimetrics Inc. (Hanson and Stevens 2002) using parameters generated by TADA as inputs. Figure 8.8 shows the flow diagram of how synthetic speech data can be generated using TADA and HLsyn.

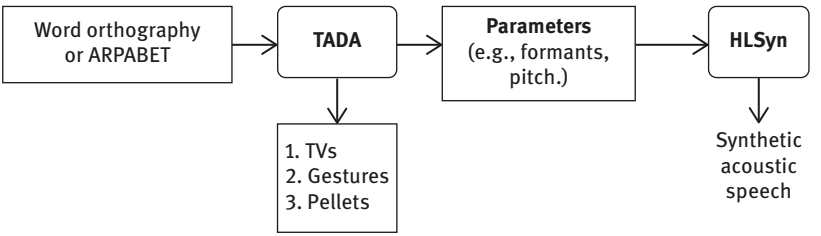

Figure 8.8: Synthetic speech and articulatory information generation using TADA and HLSyn.

Synthetic speech has been used to train SI models. While actual articulatory data (collected using electromagnetic articulometry [EMA; Ryalls and Behrens 2000] or other methods) represent articulator positions in terms of the Cartesian coordinates of transducers, their counterpart TVs are defined as relative positional measures that specify the constriction degrees and locations of gestures in the vocal tract and are provided directly during simulations using the TADA model. The benefits of using TVs as opposed to the spatial coordinates of transducers are twofold. First, the TVs specify the salient features (McGowan 1994) of the vocal tract area functions more directly than flesh-point positions of the articulators. Second, because the TVs are relative measures (as opposed to absolute flesh-point measures) they are more effective in addressing the nonuniqueness problem in SI.

We have explored different machine learning techniques for SI using TVs as articulatory representations (Mitra et al. 2010b), and, in all techniques, have observed that exploiting the correlations among TV motion patterns has helped to improve estimation accuracy. For example, in producing /t/ the TT is the "critical articulator" since it is responsible for making the alveolar constriction; similarly, for producing vowels and velic consonants, the TB is the "critical articulator". (Note: in our framework, TT is the TV used to produce alveolars, and TB is the TV used to produce vowels and velic consonants.) However, even though TB is the TV used to produce vowels and velic consonants and is not active during the production of /t/, the motion patterns of TT ("critical" for /t/) and TB ("noncritical" for /t/) are inherently coupled biomechanically with one another and their motions are correlated during the production of /t/. These correlations can be detected and used by a hierarchical support vector regression (H-SVR) model shown in Figure 8.9, to infer their underlying hierarchical structure; in turn, this hierarchical structure was shown to improve SI performance compared to a nonhierarchical SVR model (Mitra et al. 2009). Figure 8.9 shows the block diagram of the H-SVR model for SI.

Further, in Mitra et al. (2010b) an artificial neural network (ANN)-based SI architecture was presented, where a single ANN was trained to predict the motions of all the TVs. Interestingly, this single ANN showed much better performance

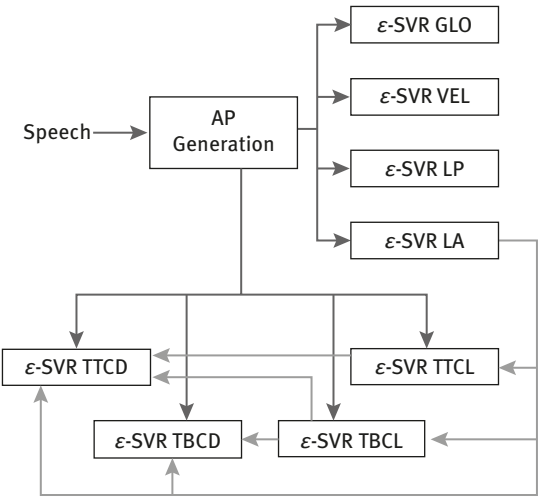

Figure 8.9: H-SVR architecture for estimating TVs from speech (Mitra et al. 2009). Note the hierarchy enforced in the architectures. The TV estimators with incoming black arrows are first decoded with input acoustic features. The TV estimators with incoming grey arrows are the ones in the second level of the hierarchy, which not only uses acoustic features as input but also the outputs from other TV estimators.

compared to the H-SVR. The rationale behind the success of the ANNs was that it leveraged the inherent correlation amongst the different TVs more effectively than the H-SVR. While in the H-SVR the hierarchy was extrinsically defined, the ANN did not have a hierarchy but intrinsically exploited the correlation among the data through sharing of hidden neurons in a fully connected network. SI, that is, going from acoustic observations to articulatory representations, is known to be an ill-posed inverse problem as it is not only nonlinear but also nonunique. In an interesting study by Qin and Carreira-Perpiñán (2007), the authors claim that nonuniqueness may not be as pronounced as nonlinearity for SI. In a separate study by Richmond (2001), both nonlinearity and nonuniqueness are addressed through time-contextualizing the acoustic space, hence mitigating the confusion in speech-to-articulator mapping. Experimental analysis (Mitra et al. 2010b) revealed that the use of contextualized acoustic input over a window up to 200ms in length and the use of nonlinear activations (such as tan-sigmoid functions) increased the reliability of ANNs for SI by quite a substantial margin, compared to using no-context and linear activation. Also, ANNs are well-known nonlinear function approximators and hence have been the algorithm-of-choice for SI researchers (Mitra et al. 2010b, 2014; Richmond 2001). Figure 8.10 shows a typical ANN architecture for estimating articulatory features from speech.

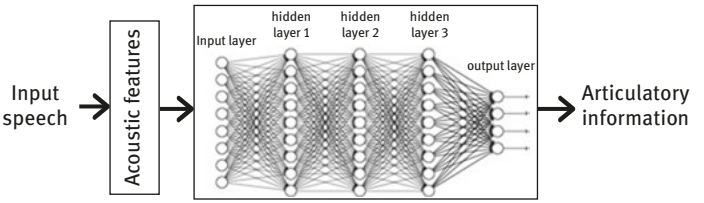

Figure 8.10: Architecture of the ANN-based direct inverse model, where input acoustic features after time contextualization is mapped to output articulatory features.

As mentioned previously, MDNs (Bishop 1994) have also been proposed for SI to address the problem of nonuniqueness in the acoustic-to-articulatory mapping (Richmond 2001). MDNs (shown in Figure 8.11) obtain articulator trajectories as conditional probability densities of the input acoustic parameters, where the pdf's of the critical articulators show very small variance compared to the noncritical articulator trajectories. For training MDNs, first a Gaussian mixture model (GMM) is trained and then its parameters (means, variances and weights represented as μ, σ^2 and α) are used as a target for training the ANN. Instead of making absolute decisions, MDNs provide a probabilistic estimate of the articulator space conditioned on the input, hence they address nonuniqueness more directly than ANNs. However, in Mitra et al. (2010b), we observed that for TV estimation, ANNs with multiple hidden layers using nonlinear activations outperformed MDNs, indicating that for SI using TVs, nonlinearity is a more severe problem than nonuniqueness; therefore, algorithms that effectively address nonlinearity can provide reasonable performance on the inversion task.

Nonuniqueness can also be handled by approaches such as distal supervised learning (DSL; Jordan and Rumelhart 1992), shown in Figure 8.12. In the DSL paradigm, two models are placed in cascade with one another: (1) the forward model (which generates acoustic features [y] given articulatory [x] trajectories, and involves an x-to-y mapping that is many-to-one and unique) and (2) the inverse model (which generates articulatory trajectories from acoustic features, and involves a y-to-x mapping that is one-to-many and nonunique). Given an

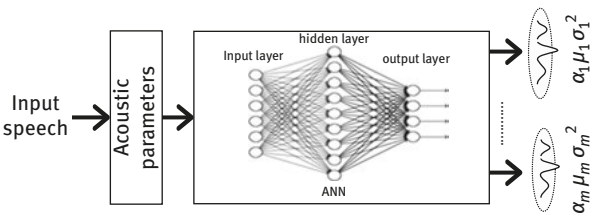

Figure 8.11: MDN for TV estimation. The MDN maps input acoustic features to the parameters of a GMM model learned from the data.

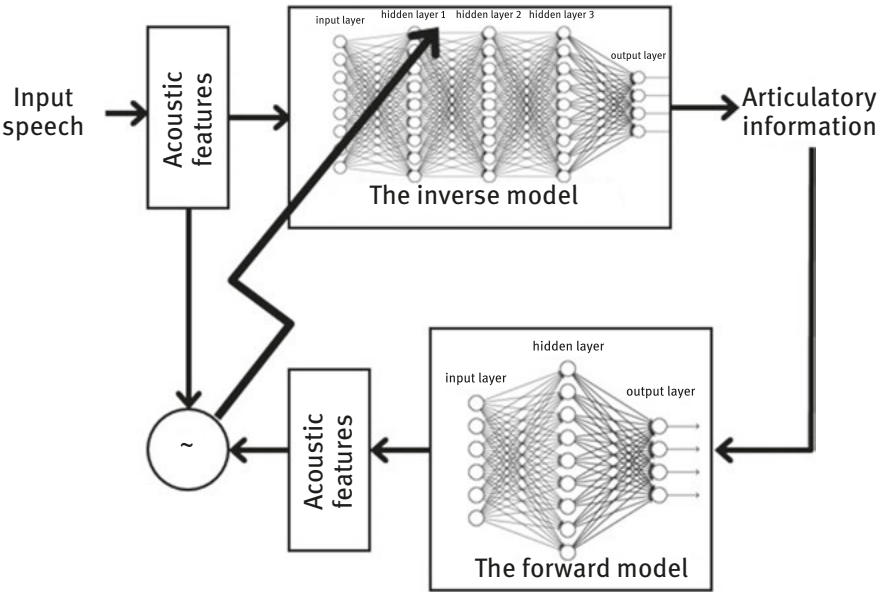

Figure 8.12: The DSL approach for obtaining acoustic to TV mapping. First, the forward model is trained, which maps articulatory features to acoustic features and then the inverse model is trained by placing it in cascade to the forward model and keeping the parameters of the forward model fixed.

observed set of $[x, y]$ pairs, DSL first learns the forward model, which is unique but not necessarily perfect. DSL then learns the inverse model by placing it in cascade with the forward model as shown in Figure 8.12. The DSL architecture can be interpreted as an "analysis-by-synthesis" approach, where the forward model is the synthesis stage and the inverse model is the analysis stage. In the DSL approach, the inverse model is trained (its weights and biases updated) using the error that is backpropagated through the forward model whose previously learned weights and biases are kept constant. For the inverse model, Jordan and Rumelhart (1992) defined two different approaches, a local optimization approach and an optimization along the trajectory approach. The local optimization approach necessitates using an online learning rule, whereas the optimization along trajectory requires recurrency in the network (hence, error minimization using backpropagation in time), both of which significantly increase the training time and memory requirements. For TV estimation (Mitra 2010), a global optimization approach was proposed which uses the tools of DSL as in Jordan and Rumelhart (1992), but uses batch training in the feed-forward network. Although the results were encouraging, it suffered from inaccuracies in the forward model, which could be overcome by using more training data.

The estimated TVs from the models discussed above are typically noisy and it was observed (Mitra 2010) that smoothing the trajectories using Kalman filtering not only ensures that the estimated articulatory trajectories will display their well-known low-pass characteristics (Hogden et al. 1998) but also improves estimation accuracy (Mitra 2010). However, smoothing estimated TV trajectories using a Kalman smoother is an ad-hoc process that can be improved upon by the use of autoregressive architectures. In Mitra (2010), an autoregressive artificial neural network (AR-ANN) shown in Figure 8.13 that used time-contextualized input feature vectors and a feedback loop connecting the output directly to the input was explored. In that study individual two-hidden layer AR-ANN models were trained separately for each TV. Although the obtained results were quite impressive and the estimated TVs were fairly smooth and less noisy in nature, the AR-ANN models failed to outperform the multi-layered ANN architecture. Note that training an AR-ANN requires dynamic backpropagation in time, which is quite expensive computationally. If one single AR-ANN is to be trained for all the TVs, then many feedback loops are needed for the model, increasing the training complexity. However, if a single model is trained for each TV, then the model fails to leverage the correlations among the TVs, which reduces the TV estimation accuracy compared to the multi-layered ANNs.

Table 8.1 shows a direct comparison of the different approaches that have been explored for TV estimation, where the performance metric is Pearson's product moment correlation (PPMC), computed between the estimated TVs and the ground truth TVs. More details about the experiments and the dataset can be found in Mitra (2010). Note that all the models except AR-ANN have their output smoothed by a Kalman smoother. In summary, the multi-layered ANNs are found to give reasonable performance, providing the best correlations with respect to the ground truth TVs most of the time. The advantages of ANNs are that they are fairly simple to train and test and are found to be quite robust across datasets and acoustic noise conditions (Mitra et al. 2010b).

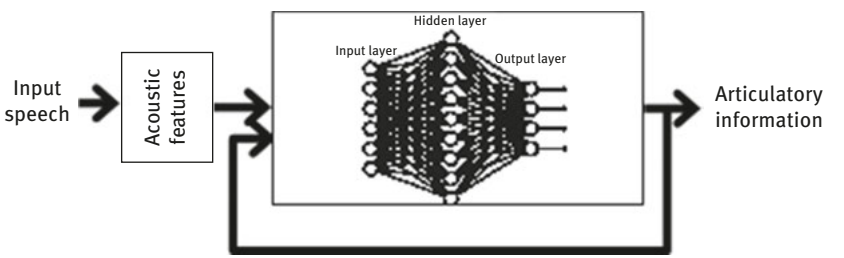

Figure 8.13: AR-ANN-based TV-estimation model. The input to the network consists of (a) speech parameterized as acoustic features after time contextualization and (b) feedback from the output of the network. The output targets are the articulatory features.

Table 8.1: PPMCs from the different TV-estimation architectures using Mel-Frequency Cepstral Coefficients (MFCC) as the acoustic feature.

	SVR	ANN	AR-ANN	DSL	MDN
GLO	0.943	0.965	0.985	0.980	0.819
VEL	0.933	0.966	0.896	0.967	0.948
LA	0.722	0.894	0.847	0.917	0.866
LP	0.743	0.927	0.518	0.788	0.748
TBCL	0.872	0.968	0.930	0.964	0.949
TBCD	0.872	0.962	0.932	0.948	0.917
TTCL	0.851	0.951	0.912	0.949	0.942
TTCD	0.898	0.949	0.905	0.930	0.939

The ANN-based TV estimator is also found to be robust to noise (Mitra et al. 2010a) when it was trained with clean synthetic speech and then tested with noisy speech at different signal-to-noise ratios (SNRs). Figure 8.14 shows a plot of the PPMC values for different TV estimates at different SNRs for subway-noise corrupted data. The PPMC values decrease with lower SNR values but the results still show moderate correlation scores at even zero or less dB SNRs.

8.1.2.4 The next step: incorporating TVs derived from synthetic speech into the ASR process

Given the high correlations found between the TVs derived from the SI system trained with synthetic data and the TVs generated by TADA, we explored whether the derived TVs will improve the accuracy of speech recognition systems (Mitra, 2013). An advantage of dealing with synthetic speech data is that there is no constraint in terms of the volume of data that can be generated. In a recent study (Mitra et al. 2013), we used the whole Carnegie Mellon University (CMU) English dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict; containing more than 111K words) to generate a large vocabulary English synthetic articulatory and acoustic data using TADA. The data was used to train a deep neural network (DNN)-based SI system (Mitra et al. 2015) having six hidden layers with tan-sigmoidal activation functions. The network was trained in a greedy layer-wise-learning manner. Table 8.2 presents the PPMCs obtained on a held-out test dataset.

The DNN-based TV estimator was used to train and test a DNN acoustic model (Mitra et al. 2015) for English telephone speech continuous speech recognition task. The Aurora-4 (Hirsch, 2001) English continuous speech recognition dataset is created from the standard 5K Wall Street Journal (WSJ0) database and has 7,180 training utterances of approximately 15 h total duration, and 330 test utterances,

Table 8.2: PPMCs between the ground-truth synthetic TVs generated by TADA and those derived using the DNN-based TV-estimator trained using large-vocabulary synthetic speech data generated by TADA using the CMU dictionary (Mitra et al. 2014).

GLO	VEL	LA	LP	TTCD	TTCL	TBCD	TBCL
0.956	0.956	0.926	0.938	0.951	0.939	0.946	0.967

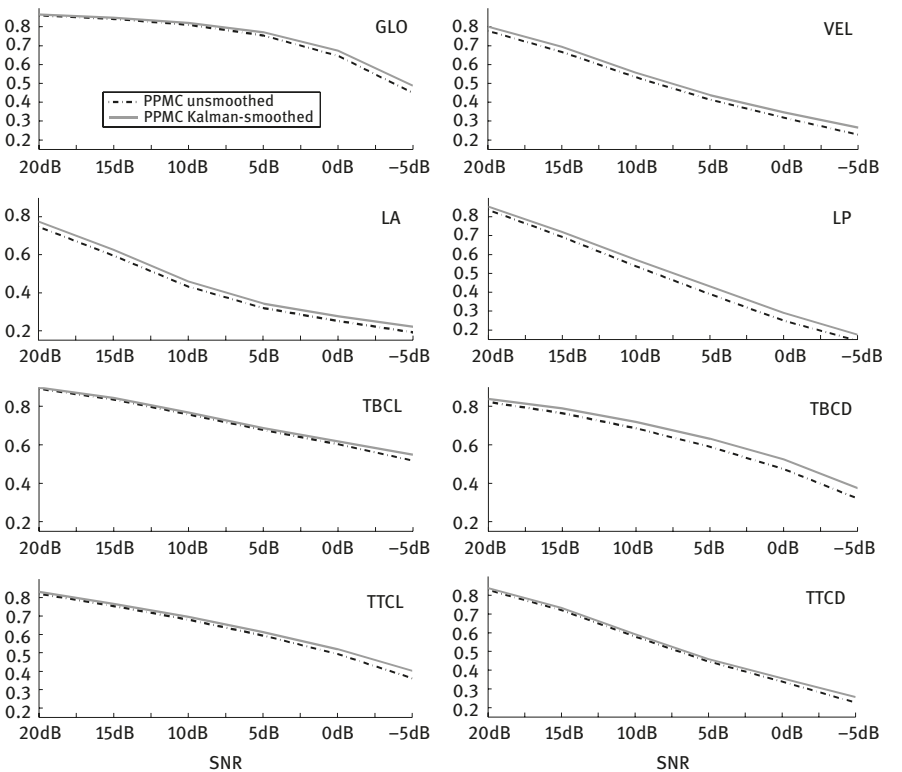

Figure 8.14: PPMC of estimated TVs at different SNRs for subway noise (Mitra et al. 2010a). The black dash-dotted line represents the PPMC from the system without smoothing, while the solid grey line represents the same after Kalman post-smoothing.

each with an average duration of 7 s. The dataset contains six additive noise versions with channel matched and mismatched conditions. Half of the training data was recorded using one microphone and the other half recorded using a different microphone (hence incorporating two different channel conditions), with

different types of added noise at different SNRs. The noise types are similar to the noisy conditions in the test data but the training set had relatively high SNRs. The test data includes 14 test sets from two different channel conditions and 6 different added noises (in addition to the clean condition). The SNR was randomly selected between 0 and 15 dB for different utterances. The six noise types used were: (1) car, (2) babble, (3) restaurant, (4) street, (5) airport and (6) train; a seventh no-noise "clean" condition was also used (set07). Thus, there were a total of 14 conditions: 2 channels × 7 noise types. The evaluation set comprised 5K words in two different channel conditions. The original audio data for test conditions 1–7 was recorded with a Sennheiser microphone, while test conditions 8–14 were recorded using a second microphone that was randomly selected from a set of 18 different microphones (more details are provided in (Hirsch, 2001)). The different noise types were digitally added to the clean audio data to simulate noisy conditions. These 14 test sets are typically grouped into 4 subsets: clean – matched-channel, noisy – matched-channel, clean with channel distortion and noisy with channel distortion, which are usually referred to as test sets A, B, C and D, respectively. A part of the clean training (893 out of 7,139 utterances) and the matched channel noisy training (2,676 utterances), which were not used in the multi-conditioned training set of Aurora-4, were used as the held-out cross-validation set that was used to track the cross-validation error DNN acoustic model training.

In Mitra et al. (2014a), the six-layered DNN-based TV estimator was used to generate estimated TVs from the Aurora-4 speech data. These estimated TVs were used in addition to MFCC features (more details in Mitra et.al. (Mitra 2014a, 2015)) to train a DNN acoustic model. For training the DNN acoustic model, initially a GMM-HMM (Gaussian mixture model-hidden Markov model) was used to align the Aurora-4 training data to produce senone (senones correspond to the leaves of a decision tree) labels for training the DNN system, where altogether 3,162 senones were used. The baseline DNN system was trained with mel-filterbank (MFB) energy features (with 40 channels); for the articulatory feature-based system, the eight-dimensional TV trajectory vector was appended with the 40-dimensional MFB energy vector. The input layer of the DNN system was formed using a context window of 15 frames (7 frames on either side of the current frame). The networks were discriminatively trained using an initial four iterations with a constant learning rate of 0.008, followed by learning rate halving based on cross-validation error decrease, a held-out cross-validation set was used for this purpose. Training stopped when no further significant reduction in cross-validation error was noted or when cross-validation error started to increase. Backpropagation was performed using stochastic-gradient descent with a mini-batch of 256 training examples. Table 8.3 shows the word error rates (WERs) from the MFB features and from the combination of MFB and TV (MFB + TV) features for

Table 8.3: WER for clean testing condition from DNN acoustic models trained with MFB features and MFB + TV features.

Feature	WER (%)
MFB40	9.8
MFB40 + TV	8.7

the clean testing condition. Note that 40 MFB energies were used in that experiment and the DNN model was trained with the multi-conditioned training data. The features were contextualized using 15 frames (7 frames on either side of the current frame).

The results in Table 8.3 show that the TV features are providing additional information helping to reduce the WER further from the baseline MFB features. This additional information could potentially be the uncovering of "hidden" gestures due to coarticulation so that errors at the phone level are reduced. Given the sophisticated language models used in ASR systems (a trigram language model distributed with the standard WSJ database was used in this experiment), many errors at the phone level are corrected by the language model. However, since this is not always possible, better phone recognition in these cases could underlie findings of lower WER. It is also worth noting that although the TVs are derived from the MFCCs, and the DNN recognition framework is providing a long temporal context (a context window of 15 frames covering ~170 ms of information), the addition of the TVs still improves accuracy possibly due to the constraints of the vocal tract imposed by the AP framework used in TADA.

Note that the training data size in Aurora-4 is not huge, so reevaluating the same modeling procedures on a larger dataset may provide more interesting results and insights regarding the efficacy of incorporating articulatory features for improving ASR performance.

8.2 SI using real data: from acoustics to TVs

In the previous section, the SI systems discussed were all trained using synthetic acoustic data and TVs generated by TADA. However, it is important to note that although TADA incorporates the primary theoretical constructs of AP, it is by no means a complete model of speech production, and the synthetic speech that it generates does not contain all of the richness and variability of natural speech. For example, prosodic information in the form of stress and intonation is not yet

included. Further, since the parameter settings used in TADA and the HLSyn synthesizer were kept constant (e.g., for F0 range, voice quality, vocal tract length and shape, and speaking rate), the training data that is produced is, essentially, data from a single "speaker". Given these constraints, it is all the more surprising (and gratifying) that adding the TVs to the features for recognition, where the system is tested on synthetic speech based on utterances from multiple real speakers, yields significant improvements. These results beg the question of whether using real data from actual speakers to train the SI system would result in greater improvements.

In this section, we explore some experiments that give insights into whether the use of real data can further improve the accuracy of TV estimation and, hence, the accuracy of ASR systems. We first describe the collection of articulatory and acoustic data produced at a normal rate and at a fast rate by actual speakers. The purpose of including a fast rate condition is to increase the occurrence and degree of coarticulation and lenition. These data will help us to determine if the SI system can indeed uncover "hidden" gestures. If so, then adding TVs to other acoustic features in an ASR system should boost its recognition accuracy.

8.2.1 Methodology

8.2.1.1 Data acquisition

To support investigation of the potential advantages of augmenting acoustic with associated kinematic information in our proposed approach to speech recognition, we have recorded a database of speech articulation using EMA. For the work discussed here, a female native speaker of American English in her mid-twenties with no speech or hearing deficiencies produced the 720 IEEE sentences (Rothauser et al. 1969) at "normal" and "fast" production rates, where normal was her preferred rate (approximately 2.9 syllables/s), and fast was produced approximately 20% more quickly.

A WAVE EMA system (Northern Digital) was used to observe the trajectories of sensors placed midsagittaly on her tongue (tongue dorsum, blade, and 1 cm posterior from apex), jaw (lower incisors), lips (upper and lower vermillion border, and left mouth corner), together with reference sensors placed on the upper incisors, nose and mastoid processes used to correct for head movement. The movement data were sampled at 100 Hz together with synchronized audio at 22,050 Hz. In post-processing, movement data were aligned to the speaker's occlusal plane and low-pass filtered at 20 Hz, providing the anterior/posterior, inferior/superior and lateral positions of each sensor relative to an origin centered on the upper incisor reference.

These data have been augmented by additional EMA data previously collected by Dr Joseph Perkell and the second author using a two-dimensional system (Perkell et al. 1992) in which a native male American English speaker produced the utterance "She had a perfect memory for details" at normal and fast rates as part of a series of additional tasks[1]; their kinematic data similarly included tongue, jaw and lip trajectories, but without the lateral dimension of movement, and were also aligned to the occlusal plane.

Finally, we have drawn on the University of Wisconsin XRMB corpus (Westbury 1994) which provides both acoustic and articulatory speech data. The articulatory data consisted of the horizontal and vertical trajectories of pellets placed on the lips and jaw and of four pellets placed on the tongue, also aligned with each speaker's occlusal plane (see Figure 8.15). We used this corpus to generate two datasets. The first consisted of naturally spoken utterances produced both as isolated sentences and short paragraphs. The acoustic and articulatory speech data were recorded from 57 American English speakers (32 females and 25 males), where each speaker completed 56 tasks, each of which can be either read speech containing a series of digits, TIMIT sentences, or even as large as reading of an entire paragraph from a book.

The second dataset consisted of XRMB utterances that were synthesized using TADA (Nam et al. 2004) and HLSyn as described in Nam et al. (2012). TVs for these data were also generated by TADA.

8.2.1.2 Converting EMA data to TVs

In the case of the XRMB data, the X–Y pellet displacement measures for 1,720 sentences across 46 different speakers were converted into 6 TV trajectories using a geometric transformation as outlined in (Nam et al. 2012) to define a corpus of "ground truth" TVs. The six TVs that were obtained from the pellets were – lip aperture (LA), lip protrusion (LP), tongue body constriction location (TBCL), tongue body constriction degree (TBCD), tongue tip constriction location (TTCL) and tongue tip constriction degree (TTCD).

A different procedure was used to convert the EMA data from the female speaker into TVs since a pharyngeal trace was not available. In this case, the sensors described above along with the palate trace of the female speaker were used to estimate constriction degree (TTCD, TBCD) TVs from the TT and TB EMA

1 This data was collected as part of a grant from NIDCD to the Speech Motor Control Group, RLE, MIT.

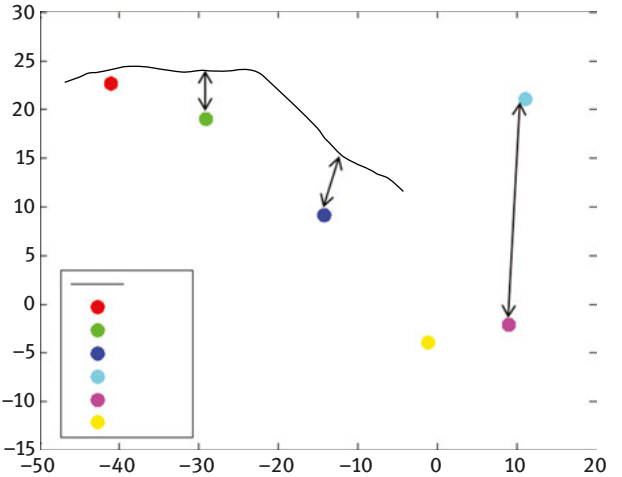

Figure 8.15: EMA pellets and TVs on the vocal tract.

positions by computing the minimum distance between the pellets and the palate. LA was computed as the distance between the upper lip (UL) and lower lip (LL) sensor positions. The TBCD, TTCD and LA TVs were computed by the following formulae:

$$LA = (UL_x - LL_x)^2 + (UL_y - LL_y)^2 + (UL_z - LL_z)^2$$
$$TBCD = \text{Min}\{\text{Distance(TB, palate)}\}$$
$$TTCD = \text{Min}\{\text{Distance(TT, palate)}\}$$

Figure 8.15 depicts the three TVs and the EMA pellets on the vocal tract.

8.2.1.3 Training of SI systems

In the previous section, we have stated that ANNs can be used to estimate TV trajectories (Mitra et al 2010) from the speech signal. Once trained, ANNs require low computational resources compared to other methods in terms of both memory requirements and execution speed. An ANN can have *M* inputs and *N* outputs; hence, a nonlinear complex mapping of *M* vectors into *N* different functions can be achieved. In such architecture, the same hidden layers are shared by all *N* outputs, endowing the ANN with the implicit capability to exploit any correlation that the *N* outputs may have among themselves. The feed-forward ANN used in our study to estimate the TVs from speech was trained with backpropagation using a scaled conjugate gradient algorithm.

Articulatory datasets

SI systems were trained using the different sets of TV data discussed in Sections 8.2.1.1 and 8.2.1.2. They are summarized in Table 8.4.

Feature extraction

The speech signal, downsampled to 8 kHz, was parameterized as MFCCs where 13 cepstral coefficients were extracted using a Hamming analysis window of 20 ms with a frame rate of 10 ms. The TVs and MFCCs were mean and variance normalized to have zero mean and a variance of 0.25. The MFCCs were then contextualized by concatenating every other feature frame within a 160 ms window on either side of each frame.

ANN Training

For the ANN-based TV estimator, the input dimension was 221 (= 13 MFCCs × 17 frames) and the output dimension was 6 (= 6 TVs). Eighty percent of the data was used for training, and 10% each was used for cross validation and testing. A two hidden layer neural network was trained in a greedy layer-wise manner. Networks with different hidden-layer neurons (100–500) were trained, and among them the best performing network was chosen for training the second hidden layer. The network was not trained to have further hidden layers as the performance improvement over the single layer network was marginal. The performance of the TV estimator was measured by computing the PPMC of the estimated TVs with the ground truth TVs on the test set.

Attempts to perform speaker normalization

The XRMB database consisted of speech and articulatory data from 46 different speakers. As a result, it was essential to perform speaker normalization in order to use all of the speakers' data to train a single SI system. As an attempt

Table 8.4: Description of different articulatory datasets used in this study.

Dataset	Description
XRMB synthetic speech	Utterances from XRMB database synthesized using TADA and HLSyn as described in (Nam et al. 2012). Note that the TVs for these data were also generated by TADA
XRMB natural speech	Complete XRMB database with pellet trajectories converted to TVs using the method outlined in (Nam et al. 2012)
EMA natural speech	EMA articulatory data described in Section 8.2.1.1 converted to TVs using the method outlined in Section 8.2.1.2

Table 8.5: Summary of different SI systems with their training information.

TV estimator name	Training dataset	Normalization scheme
XRMB_SYN	Synthetic XRMB database generated using TADA	Global mean and variance
XRMB_ALLNORM	Natural XRMB database converted to TVs	Global mean and variance
XRMB_SPKNORM	Natural XRMB database converted to TVs	Speaker-specific mean and variance
XRMB_FEMALES_SPKNORM	Females from natural XRMB database converted to TVs	Speaker-specific mean and variance
XRMB_MALES_SPKNORM	Males from natural XRMB database converted to TVs	Speaker-specific mean and variance
EMA_1SPKR	Single female speaker EMA data converted to TVs	Speaker-specific mean and variance

towards performing the normalization, the MFCCs and TVs were normalized based on speaker-specific means and variances. A TV estimator was trained based on this normalization. This estimator is referred to subsequently as "XRMB_SPKNORM." Males and females have different ranges for pitch and vocal tract lengths. Motivated by this fact, gender-specific TV estimators were trained on the XRMB database leading to "XRMB_FEMALES_SPKNORM" and "XRMB_MALES_SPKNORM" TV estimators. Table 8.5 summarizes the different SI systems trained.

8.2.2 Results

8.2.2.1 TV estimator training results

The trained TV estimators were tested on 10% of their respective datasets where the sentences were chosen randomly. The performance of the TV estimator was measured by the PPMC between the estimated and ground-truth TVs using the test set. The results for the different SI systems are given in Table 8.6.

Overall, the correlations increase as the variability in the training and test data is reduced. Variability is least for the SI system developed using synthetic data derived from TADA. Thus, the correlation is highest for the estimated and TADA-generated TVs. In the case of TVs derived and estimated from real data, the correlations are higher when the system is trained and tested on data from a single speaker, or when the data is from a single gender and the data for each speaker is normalized using the speaker-specific mean and variance.

Table 8.6: PPMC results of trained TV estimators on their respective test datasets.

	LA	TBCD	TTCD	LP	TBCL	TTCL
XRMB_SYN	0.89	0.91	0.93	0.92	0.94	0.89
XRMB_ALLNORM	0.57	0.56	0.73	0.71	0.70	0.59
XRMB_SPKNORM	0.67	0.64	0.77	0.57	0.80	0.64
XRMB_FEMALES_SPKNORM	0.72	0.67	0.79	0.62	0.83	0.66
XRMB_MALES_SPKNORM	0.70	0.65	0.79	0.58	0.83	0.72
EMATV_1SPKR	0.64	0.80	0.72	NA*	NA*	NA*

NA*: these TVs were not estimated since we are only looking at constriction degrees.

To determine if "hidden" speech gestures can be uncovered by any of our SI systems, we focus on three utterances:

1. She had a **perfect memory** for details.
2. The empty **flask stood** on the tin tray.
3. The beam dropped down on the **workman's** head.

The words in bold contain the consonant clusters that are the focus of study. Table 8.7 shows the correlations obtained by the different SI systems for these utterances. As before, the correlations in Table 8.7 show that less variability in the data results in higher correlations. The EMATV_1SPKR SI system has the highest correlations for sentences 2 and 3 since those utterances were produced by the same speaker (note that these utterances were not included in the training of this system). However, note that reasonably high correlations are also obtained for sentence 1 which was produced by a speaker that was not included in the XRMB database. In fact, we can see that the correlations improved significantly when the data for each speaker is normalized according to their means and variances. One significant finding is the poor correlations obtained for the SI system based on synthetic data. This result is not surprising given the synthetic data lacks the range of variability represented in real data.

Note that none of the SI systems produce TVs for TTCD that correlate well with the TTCD derived from actual data for the normal- or fast-rate production of "perfect memory." This is because the "perfect memory" utterance was from a speaker that was not present in the XRMB database. Moreover, the phonetic combination of /k/+/t/+/m/ was not present in any of the utterances in the XRMB database. However, looking at the individual speaker-dependent SI systems developed for each male speaker in the XRMB corpus, we found some speakers for whom the correlations were significantly better. We have shown the plots of the TVs estimated from two such speaker-dependent SI systems in the analysis discussed in Section 8.2.2.2.

Table 8.7: Correlations of the SI systems in Table 8.6 for the sentences in focus.

XRMB_ALLNORM	flask stood		perfect memory		workman's	
	fast	normal	fast	normal	fast	normal
LA	0.47	0.31	0.74	0.67	0.24	0.39
TBCD	0.53	0.59	0.19	0.75	0.80	0.77
TTCD	0.71	0.77	0.09	0.39	0.65	0.71
Average	0.57	0.56	0.34	0.60	0.56	0.62

XRMB_SPKNORM	flask stood		perfect memory		workman's	
	fast	normal	fast	normal	Fast	normal
LA	0.65	0.62	0.55	0.67	0.18	0.56
TBCD	0.30	0.32	0.70	0.73	0.71	0.72
TTCD	0.68	0.75	0.13	0.49	0.76	0.71
Average	0.54	0.56	**0.46**	**0.63**	0.55	0.66

XRMB_FEMALES_SPKNORM	flask stood		perfect memory		workman's	
	fast	normal	fast	normal	fast	normal
LA	0.50	0.63	0.57	0.63	0.19	0.57
TBCD	0.43	0.52	0.61	0.58	0.75	0.75
TTCD	0.58	0.71	−0.10	0.37	0.71	0.71
Average	0.50	0.62	0.36	0.53	0.55	0.68

XRMB_MALES_SPKNORM	flask stood		perfect memory		workman's	
	fast	normal	fast	normal	fast	normal
LA	0.72	0.68	0.62	0.52	0.41	0.51
TBCD	0.48	0.50	0.72	0.82	0.82	0.62
TTCD	0.76	0.72	0.01	0.43	0.70	0.57
Average	0.65	0.64	0.45	0.59	0.64	0.57

emaTV	flask stood		perfect memory		workman's	
	fast	normal	fast	normal	fast	normal
LA	0.85	0.77	0.35	0.61	0.68	0.77
TBCD	0.87	0.86	0.44	0.46	0.88	0.81
TTCD	0.85	0.83	−0.24	0.26	0.68	0.78
Average	0.86	0.82	0.18	0.44	0.75	0.79

(continued)

Table 8.7: continued

XRMB_SYN	flask stood		perfect memory		workman's	
	fast	normal	fast	normal	fast	normal
LA	0.18	0.43	0.31	−0.01	−0.27	0.46
TBCD	0.29	0.34	0.56	0.29	0.23	0.08
TTCD	0.18	0.43	0.04	0.08	−0.33	0.25
Average	0.22	0.40	0.30	0.12	−0.13	0.26

8.2.2.2 Analysis of reduction

In the analysis that follows, the TVs derived directly from point-source data are referred to as ACT_TV and those estimated by an SI system are referred to as EST_TV. Also, in the figures, the delimited gestures are shown by boxed regions determined by thresholding the velocity extrema associated with the ACT_TV trajectories using a 90% criterion applied to the local event range (mview software, Haskins Laboratories).

Analysis of "perfect memory"

Figure 8.16 shows spectrograms and the ACT_TVs and EST_TVs for sentence 1. As can be seen in the normal-rate production, the acoustics show silence for the /k/ closure followed by the /t/ burst, followed by a period of silence and then the /m/ murmur at the beginning of "memory." Both sets of TVs show a TB gesture for the /k/ that overlaps to some extent with the TT gesture for the /t/ and the lip gesture for the /m/. In contrast, there is no silence between the last vowel in "perfect" and the

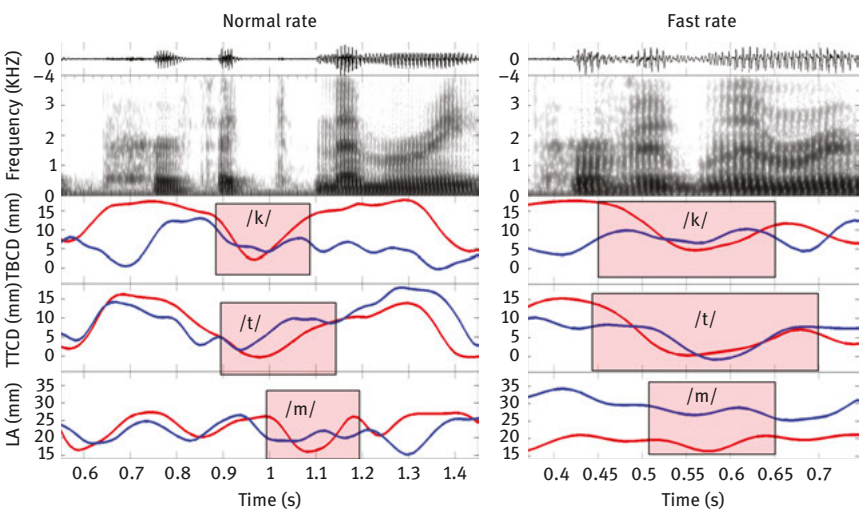

Figure 8.16: ACT_TVs (red) and EST_TVs (blue) for "perfect memory."

first vowel in "memory." Instead, this region appears as one sonorant consonant, that is, the /m/. However, the articulatory data tell a different story. As in the normal rate speech, we see gestures for the /k/, the /t/ and the /m/. The difference appears to be that there is considerably more overlap between the gestures. In particular, the /m/ gesture is fully overlapped with that of the other consonants. Thus, this fast-rate production of "perfect memory" contains what we refer to as "hidden gestures" for the /k/ and the /t/. Note that both these gestures are apparent in the estimated TVs, although the closure for the lip gesture is weaker than the actual gesture.

Analysis of "flask stood"

Figure 8.17 shows spectrograms and the TVs for the normal-rate and fast-rate productions of sentence 3. In the case of the normal-rate production, the consonant cluster /sk/ at the end of "flask" and the /st/ at the beginning of "stood" are clearly seen in the acoustics and both the ACT_TVs and EST_TVs show constrictions in the right regions. However, in the fast-rate production of this utterance, the acoustics suggest that the /k/ in "flask" was not produced. Instead, it appears as if the /s/ in "flask" and the /s/ in stood are combined (the duration of this /s/ is about 30 ms longer than the ones in the normal-rate production) and this /s/ is then followed by a the /t/ in "stood." This appears to be a case where the fast-rate production resulted in no gesture being made for the /k/. Although there is lowering of the TBCD gesture during the /t/, this lowering appears to be due to the

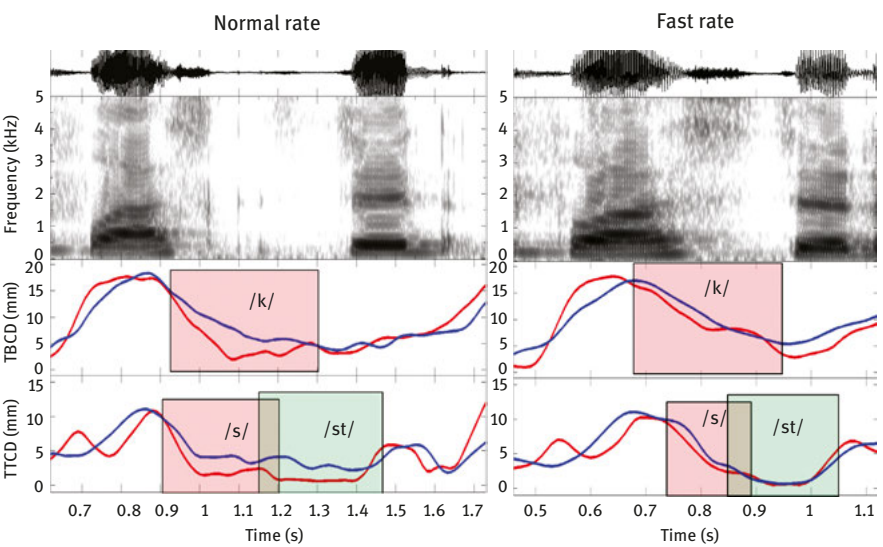

Figure 8.17: ACT_TVs (red) and EST_TVs (blue) for "flask stood."

/t/ closure and can be seen in situations where an /s/ or /t/ is produced without an adjacent velar consonant. It is possible that this apparent deletion of the /k/ gesture is due to the complexity of these cluster sequences, which include four consecutive consonants.

Analysis of "workman"

Figure 8.18 shows spectrograms and the TVs for the normal-rate and fast-rate productions of sentence 2. The ACT_TVs and EST_TVs are strongly correlated across the utterance. In particular, both show the /k/ constriction when it is produced as a stop in the normal-rate production and as a fricative in the fast-rate production. Note that the /k/ gesture in the fast-rate production of the utterance is weaker than it is in the normal-rate production of the same. This not surprising given the EST_TV gestures are derived from the acoustics. Finally, note that both sets of TVs for both productions show the closure of the lips for the /m/.

8.2.2.3 Reduction and ASR

We evaluated the performance of a state-of-the art recognition system with the six utterances in this study. The results are shown in Table 8.8. The phone recognizer was trained using SRI International's DECIPHER® Speech Recognition system, which uses a GMM-HMM, trained using the maximum likelihood criteria.

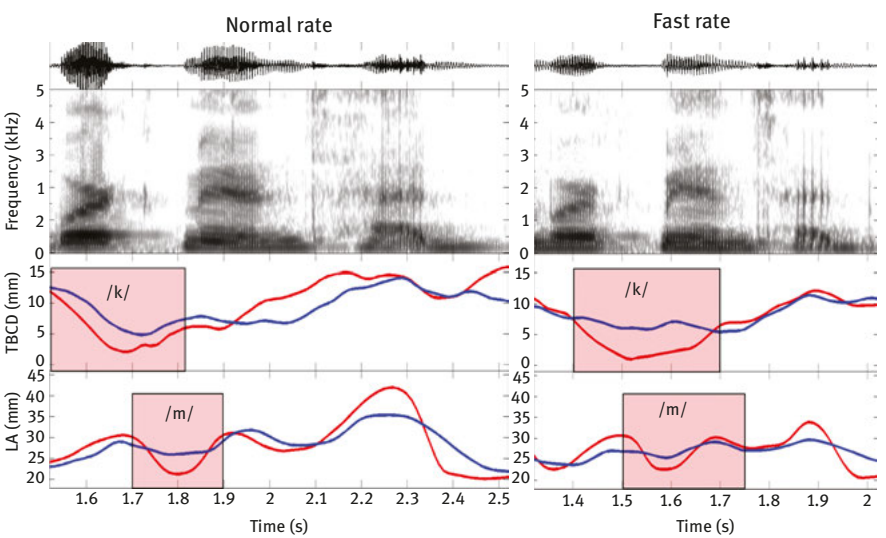

Figure 8.18: ACT_TVs (red) and EST_TVs (blue) for "workman's."

Table 8.8: Phones and word recognition results from SRI ASR system.

Ground truth	SRI phonetic recognizer	SRI word recognition
perfect memory (normal production	p er f ih k ah m er iy	she had a perfect memory for details
perfect memory (fast production	v er g eh r iy	share a part of the river details
workman's (normal production)	w er m eh n d	the beam jot down on the work manned had
workman's (fast production)	v er b ih n t	they've been cut down on the work been type
flask stood (normal production)	f l ae s k dh ey d	the empty flasks hidden under tinge tray
flask stood (fast production)	f l ae s t uh d	the empty flustered and that the tin tray

A 4-gram phone-based language model was used to decode the speech files. The system uses 41 distinct phone units and 4 other units defining nonspeech acoustic conditions. The setup uses 16 kHz audio files and produces 13 MFCCs with their deltas, double-deltas and triple-delta coefficients, resulting in a 52-dimensional feature set. The 52-dimensional feature vector is transformed to 39 dimensions using heteroscedastic linear discriminant analysis and the resulting features were fed to the phone recognition system. In addition to the recognized phones, Table 8.8 also contains the recognized sentence. Note that this recognizer, with natural language processing, is set up for conversational speech as opposed to read speech which may in part account for the poor word recognition rate for these sentences.

As can be seen, the phonetic recognition system does not transcribe the /t/ and /m/ in the normal production of "perfect memory," but it does get the /k/ and the second and third syllables of "memory" so that it is able to recognize the two words correctly given the language model. In the fast repetition, the recognizer misses too many of the sounds, including the /k/ in "perfect" and is therefore unable to recover the correct words. The phone recognizer also does not recognize the reduced /k/ in the fast repetition of "workman," and the /st/ cluster in "stood" is recognized as the voiced dental fricative /dh/. In the case of the latter, this result is not too surprising since it appears as if the speaker did not make a complete closure for the /t/. Instead, we see weak frication during what is normally silence when a complete closure is made.

The question is whether the articulatory data from the SI system can improve recognition. In the case of "perfect memory," the TVs contain gestures for all of the sounds not recognized by the phonetic recognition system. In the case of

"workman's," the recognizer is able to recognize the word "work" in both the slow and fast productions even though the /k/ is not recognized at the phone level. In the normal production of "flask stood," the TV for the TT shows both the constrictions for the /s/ and /t/ in "stood" where the degree of the constriction for the /t/ is considerably smaller than it is for the /s/. Thus, this information would suggest a fricative followed by a stop. These findings suggest that inclusion of articulatory information in speech recognition systems could potentially improve performance.

8.3 Discussion and conclusion

Working with naturally spoken data can result in SI systems that produce TVs that closely match TVs computed directly from real articulatory data. However, the variability in the data needs to be properly normalized or restricted. Thus, a future goal of this work is to devise an automatic method to determine which of possible several different SI systems will work for any given speaker, especially if that speaker's data has not been used as part of the training data for any of the SI systems. The results clearly show that an SI system trained with synthetic data does not generate TVs that are highly correlated with the ground truth, and that TVs generated from an SI system trained with natural data will be more highly correlated. Thus, one might expect that the results shown in Table 8.3 of the DNN-based recognizer would be significantly better had natural data been used to train the SI system.

Given the large contextual window used in the generation of the TVs, coarticulation is being properly modeled so that overlapping gestures are estimated even when it is not apparent from the acoustics that a sound was produced. Of course, the information for that sound being produced has to be present in the acoustics, but the information can sometimes be subtle and it may be distributed across the production of the coarticulated sounds. The results of the correlations shown in Tables 8.7 and 8.8 further suggest that the type of variability seen in the test set must also be represented in the training set. At least that is one interpretation for why the speaker-specific SI systems in Table 8.8 were able to give better correlations than other speaker-specific SI systems generated from the XRMB corpus, and better than the correlations of the SI systems in Table 8.7. This is not surprising since matched conditions for any technology gives better results than unmatched conditions.

Given the performance of the SI systems for the sentences analyzed in detail, it appears that adding TVs and/or gestures to recognition systems could greatly improve their phonetic recognition and, therefore, their word accuracy. In future work, we will explore the best way in which the estimated articulatory information should be incorporated.

Acknowledgments: We would like to thank Dr Joseph Perkell for permission to use his EMMA data, collected under an NIH NIDCD grant to the Speech Motor Control Group, RLE, MIT. This work was supported by NSF Grants 1436600, 1162525 and 0703859.

References

Atal, Bishnu S., Jih J. Chang, Max V. Mathews & John W. Tukey 1978. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. *Journal of the Acoustical Society of America* 63 (5). 1535–1555.

Bishop, Christopher M. 1994. Mixture density networks. Tech. Report NCRG/4288, Neural Computing Research Group, Dept. of Comp. Sc., Aston Univ., Birmingham, U.K.

Browman, Cathe P. & Louis Goldstein 1988. Some notes on syllable structure in articulatory phonology. *Phonetica* 45. 140–155.

Browman, Cathe P. & Louis Goldstein 1989. Articulatory gestures as phonological units. *Phonology* 6. 201–251.

Browman, Cathe P. & Louis Goldstein 1992. Articulatory phonology: An overview. *Phonetica* 49. 155–180.

Browman, Cathe P. & Louis Goldstein 2000. Competing constraints on intergestural coordination and self-organization of phonological structures. *Les Cahiers de l'ICP, Bulletin de la Communication Parlée* 5. 25–34.

Hanson, Helen M. & Kenneth N. Stevens 2002. A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn. *Journal of the Acousical. Society of America*. 112(3). 1158–1182.

Hirsch. G., Experimental Framework for the Performance Evaluation of Speech Recognition Front-Ends on a Large Vocabulary Task, ETSI STQ-Aurora DSR Working Group, June 4, 2001.

Hogden, John, David Nix & Patrick Valdez 1998. An Articulatorily Constrained, Maximum Likelihood Approach to Speech Recognition. Tech. Report, LA-UR--96-3945, Los Alamos National Laboratory, NM.

Jordan, Michael I. & David E. Rumelhart 1992. Forward models-Supervised learning with a distal teacher. *Cognitive Science* 16. 307–354.

McGowan, Richard S. 1994. Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: preliminary model tests. *Speech Communication* 14 (1). 19–48.

Mermelstein, Paul 1973. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America* 53 (4). 1070–1082.

Mitra, Vikramjit 2010. Articulatory Information for Robust Speech Recognition. Ph.D-Thesis, Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA.

Mitra, Vikramjit, Özbek, Yücel I., Nam, Hosung, Zhou, Xinhui, & Carol Espy-Wilson 2009. From Acoustics to Vocal Tract Time Functions. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan. 4497–4500.

Mitra, Vikramjit, Hosung Nam, Carol Espy-Wilson, Elliot Saltzman & Louis Goldstein 2010a. Articulatory information for noise robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing* 19 (7). 1913–1924.

Mitra, Vikramjit, Hosung Nam, Carol Espy-Wilson, Elliot Saltzman & Louis Goldstein. 2010b. Retrieving tract variables from acoustics: A comparison of different machine learning strategies. [Special Issue]. *IEEE Journal of Selected Topics on Signal Processing* 4 (6). 1027–1045.

Mitra, Vikramjit, Wen Wang, Andreas Stolcke, Hosung Nam Colleen Richey, Jiahong Yuan & Mark Liberman. 2013. Articulatory features for large vocabulary speech recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada. 7145–7149.

Mitra, Vikramjit, Sivaraman, Ganesh, Nam Hosung, C. Espy-Wilson and E. Saltzman. 2014a. Articulatory features from deep neural networks and their role in speech recognition. *Proceedings of ICASSP*, Florence, Italy. 3041–3045.

Mitra, Vikramjit, Elizabeth Shriberg, Mitchell McLaren, Andreas Kathol, Colleen Richey, Dimitra Vergyri & Martin Gracierana. 2014b. The SRI AVEC-2014 evaluation system. *4th International Audio/Visual Emotion Challenge and Workshop, ACM Multimedia*.

Mitra, Vikramjit, Ganesh Sivaraman, Hosung Nam, Carol EspyWilson & Elliot Saltzman 2015. Channel and noise robustness of articulatory features in a deep neural net based speech recognition system. *Journal of the Acoustical Society of America* 137. 2301. doi:10.1121/1.4920397.

Nam, Hosung & Elliot Saltzman 2003. A competitive, coupled oscillator model of syllable structure. *Paper presented at the 15th International Congress of Phonetic Sciences*, Barcelona, Spain. 2253–2256.

Nam, Hosung, Louis Goldstein, Elliot Saltzman & Dani Byrd 2004. TADA: An enhanced, portable task dynamics model in MATLAB. *Journal of Acoustical Society of America* 115 (5). 2430.

Nam, Hosung, Louis Goldstein & Elliot Saltzman 2009. Self-organization of syllable structure: a coupled oscillator model. In François Pellegrino, Egidio Marisco & Ioana Chitoran (eds.), *Approaches to phonological complexity*, 299–328. Berlin and New York: Mouton de Gruyter.

Nam, Hosung, Vikramjit Mitra, Mark Tiede, Mark Hasegawa-Johnson, Carol Espy-Wilson, Elliot Saltzman & Louis Goldstein 2012. A procedure for estimating gestural scores from speech acoustics. *Journal of the Acoustical Society of America* 132 (6). 3980–3989.

Ouni, Slim & Yves Laprie 1999. Design of hypercube codebooks for the acoustic-to-articulatory inversion respecting the non-linearities of the articulatory-to-acoustic mapping. *Proceedings of Eurospeech* 1. 141–144.

Papcun, George, Hochberg, Judith, Thomas, Timothy R., Laroche, François, Zachs, Jeff & Simon Levy 1992. Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data. *Journal of the Acoustical Society of America* 92 (2). 688–700.

Perkell, Joseph S., Marc H. Cohen, Mario A. Svirsky, Melanie L. Matthies, Inaki Garabieta & Michel T. T. Jackson 1992. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America* 92 (6). 3078–3096.

Qin, Chao & Miguel Á. Carreira-Perpiñán 2007. An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. *Proceedings of Interspeech*, Antwerp, Belgium. 74–77.

Rahim, Mazin G., Kleijn, W. Bastiaan, Schroeter, Juergen & Colin C. Goodyear 1991. Acoustic-to-articulatory parameter mapping using an assembly of neural networks. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toronto, Ontario, Canada. 485–488.

Rahim, Mazin G., Goodyear, Colin C., Kleijn, W. Bastiaan, Schroeter, Juergen & M. Mohan Sondhi 1993. On the use of neural networks in articulatory speech synthesis. *Journal of the Acoustical Society of America* 93 (2). 1109–1121.

Richmond, Korin. 2007. Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion. *Lecture Notes in Comp. Science* 4885. 263–272.

Richmond, Korin. 2001 Estimating articulatory parameters from the Speech Signal. PhD Thesis, University of Edinburgh.

Rothauser, E. H., W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek & M. Weinstock, 1969. IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, USA 17 (3). 225–246.

Ryalls, Jack & Susan J. Behrens 2000. *Introduction to speech science: From basic theories to clinical applications*. Allyn & Bacon.

Saltzman, Elliot, Hosung Nam, Jelena Krivokapic & Louis Goldstein 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. *Paper presented at the 4th Conference on Speech Prosody*, 175–184.

Saltzman, Elliot & Dani Byrd 2000. Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science* Vancouver, Canada 19. 499–526.

Saltzman, Elliot & Kevin G. Munhall 1989. A dynamical approach to gestural patterning in speech production. *Ecological Psychology* 1 (4). 333–382.

Westbury, John 1994. *X-ray microbeam speech production database user's handbook*. Madison: University of Wisconsin.

Francesco Cangemi and Oliver Niebuhr

9 Rethinking reduction and canonical forms

Abstract: We conclude the book's reflections on reduction and reduced forms by exploring the complementary concept of canonical forms, which has profoundly shaped research on sound segments and their realization. Canonical forms have been described as symbolic, linear, and minimalistically contrastive representations, as in the case of phonological transcriptions of words. They have been conceived as mental word templates that can be eroded step by step in speech production, and then have to be reconstructed in speech perception. As a consequence, in theories focusing on canonical forms, reduced forms have often been relegated to energy efficiency or mere performance accidents. Drawing insights from (a) the history of linguistics (with a focus on the reasons behind the long-standing success of canonical forms) and (b) the book's contributing chapters (with a focus on how the study of reduced forms can inform linguistic theory), we identify four directions into which reduction research must be extended in the future with empirically rather than canonically defined reference forms. These are reduction patterns and reference forms in the area of prosody, reinforcement or strengthening as the antithesis of speech reduction, factors for predicting degree of reduction and their phonetic results, and, with regard to the latter, the separate contribution of reduction to communicative function. These research directions will help us to reassess our understanding of the dichotomy between canonical and reduced forms.

Keywords: canonical forms, phoneme, alphabetic writing, non-lexical meaning, non-linear representations, hypo-hyper theory

9.1 Reduced and canonical forms: A good partnership?

The contributions to this book show that so-called reduction phenomena and reduced forms are of great importance for our understanding of how speech works. This

Note: The chapter is the product of the joint effort of the two authors, but FC is mainly responsible for Sections 9.2.1–2 and 9.3.3, and ON for Sections 9.1, 9.2.3 and 9.4.

Francesco Cangemi, IfL Phonetics, University of Cologne, Germany
Oliver Niebuhr, SDU Electrical Engineering, Mads Clausen Institute, University of Southern Denmark, Sonderborg, Denmark

https://doi.org/10.1515/9783110524178-009

applies to both research and applications in automatic speech recognition, natural text-to-speech synthesis, second language acquisition, psycholinguistics, phonetics, and phonology. Given this fact, it is surprising that reduced forms have received so little attention as yet. It is not long ago that research on reduction phenomena really got going, and it is only since recently that this line of research picked up speed.

Finding numbers in support of this statement is easy. Besides the fact that this book is the first that specifically addresses the concepts and processes of reduction, a title search from the proceedings of the International Congresses of Phonetic Sciences shows virtually no mention of these topics up to the end of the 1980s. Only a couple of titles are found for the congresses in 1991 and 1995, and then half a dozen titles for each of the following congresses – up to the latest (2015), in which the amount suddenly doubled. The numbers retrieved from the Google Scholar database are even more impressive. We found only 58 hits for "speech reduction" in the 25-year interval from 1965 to 1990. In contrast, the same interval from 1990 to present (2018) yielded almost 16 times as many hits, that is, 924. Only since January 2017, already 203 new papers dealing with "speech reduction" were added to Google Scholar. This corresponds to 350% of the total number of papers that Google Scholar found between 1965 and 1990. These numbers can hardly be explained by the generally increasing number of papers alone.

How could such an important aspect of speech remain "under the radar" for so long? It was stated in the introduction and reflected in virtually all preceding chapters that reduction is no simple and categorical all-or-none phenomenon, but rather a complex process that manifests itself to various degrees in various conditions. Revealing this complexity, although we are still far from having the full picture, has surely contributed to make reduction a research subject in its own right that poses new challenges and calls for new joint efforts at the intersection of phonetics, phonology, and speech technology.

The hyper- and hypospeech (H&H) theory of Lindblom (1990) and the influential paper of Nolan (1992) on the "descriptive role of segments" were major driving forces for this new way of thinking. The work of Mirjam Ernestus, Natasha Warner, and colleagues, which also includes the series of (Nijmegen) Speech Reduction Workshops,[1] represent valuable recent contributions to raise and maintain the awareness of the complexity of reduction. However, as strange as it may seem at first sight, it is probably not exaggerated to state that we owe the breakthrough of research on reduction to the progress in computer technology.

It was only some years ago that speech scientists still had to decide whether they wanted to make in-depth or large-scale analyses of their data (cf. Mattingly 1999).

1 http://www.u.arizona.edu/~nwarner/Workshop.html, see also Ernestus and Warner (2011).

Nowadays, computers have become so powerful that this technical barrier no longer exists. In-depth and large-scale analyses are not mutually exclusive aims anymore. As is stressed by Harrington (2010: 82): "As a result of the further development in computer technology [...] there are now large-scale acoustic databases, many of them phonetically labeled, as well as tools for their analyses", such as WebMAUS (Strunk et al. 2014), Prosomarker (Origlia and Alfano 2012), ProsodyPro (Xu 2013), and the analysis scripts of Barbosa et al. (2016). Researchers have begun to annotate and analyse the timing of face and body movements relative to linguistic elements in the acoustic domain (cf. Allwood et al. 2007), "based on speech corpora that are increasingly representative of natural, spontaneous speech" (Bird and Harrington 2001: 1). Note that it is less than 15 years ago that the outstanding speech scientist Peter Ladefoged stated in his book on Phonetic Data Analysis that cassette tapes and DAT tapes are still fairly widespread, although it is "likely that soon we will all be using some form of digital recording" (Ladefoged, 2003: 182). Ladefoged was obviously right. Yet, he underestimated how computers would change his whole discipline. An early prime example for the new opportunities offered by analyses of extensive annotated speech corpora is the study of Campbell and Mokthari (2013) on addressee-specific voice quality variation in Japanese, which was published in the same year as Ladefoged's book. About 15 years later, there is nothing "exotic" about such big-data studies anymore. In fact, six out of the seven contributed chapters in this book present data from corpus analyses, the most outstanding example being the analysis of the 900-minute Ernestus Corpus of Spontaneous Dutch (Ernestus 2000).

The reason why this development played (and still plays) a key role in impelling speech reduction research is that it provided us with one basic truth: Reduction is not an exception to the rule and hence not something that we can stay clear of in dealing with speech communication. Rather, reduction *is* the rule. Not even motherese and child-directed speech are spared from strong reductions (see Chapter 3 of van Dommelen). For example, frequency statistics derived from the Kiel Corpus of Spontaneous Speech (Wesener 1999) show that intervocalic /bdg/ are very often realized as approximants in German. In the case of /b/, like in *aber* (but), this applies to 72% of all tokens. The words *ist* (is) lacks its final /t/ in 89% of the cases; *einem* (indef. article) typically deviates more strongly from its full form and is to 80% realized as either [mm] or [m]. Liquids like /l/ are literally melting away in German spontaneous speech. About 50% of all *also* (so) and *mal* (discourse particle) tokens lack this sound, at least in the form of a separate segment (see the /l/ reduction in the chapter of Ernestus and Smith). Examples like these fit in well with the finding of Cutugno et al. and Adda-Decker and Lamel (Chapters 7 and 4) that so-called weak forms are particularly susceptible to reduction in German. They also fit in well with the fact that reduction is more frequent in spontaneous than in read speech (see Chapters 3, 4, and 2 of van Dommelen, Adda-Decker and Lamel, and Clopper and Turnbull).

This basic insight "that reduced forms [are] the most natural form of speech coding" (Cutugno et al., Chapter 7) obviously challenges a pervasive well-guarded concept of the speech sciences: canonical forms, that is, the symbolic, linear, and minimalistic representations of words. Therefore, we dedicate the final chapter of this book to this pervasive concept. Canonical forms and their "partnership" with phonemes and reduction offer us a privileged viewpoint to revisit, in an integrated way, some of the most important developments of linguistics in the twentieth century. On this basis, we outline with reference to important pages of the classic literature how and why canonical forms became so powerful that they were able to marginalize "troublemakers" like speech reduction for such a long time.

The roots of canonical forms can be retraced to the research questions and practices behind most work in phonology and phonetics up to the twentieth century (Section 9.2). We will see, however, that researchers from different fields and schools within the speech sciences have already anticipated some of the theoretical and empirical insights that started blooming in the recent investigations of reduced forms (Section 9.3). Following from this line of thought, we conclude that the study of reduced forms will be a catalyst in advancing research on speech communication, with reduction and canonical forms having a problematic, but not completely useless and overall improvable partnership (Section 9.4).

9.2 Why canonical forms?

In the early stages of modern linguistics, that is, from William Jones to Ferdinand de Saussure, the primacy of canonical forms can be explained on *methodological* grounds. Symbolic, linear and minimalistic representations (Section 9.2.3) were the main tool to address the questions raised in the scientific community, either in the form of written words in the early phase of diachronic studies in comparative philology or indoeuropeanistics (Section 9.2.1) or in the form of spoken words in the later phase of synchronic studies in the semiology of living languages (Section 9.2.2).

9.2.1 Indoeuropeanistics, diachrony, and written words

In what is customarily seen as the pivotal moment in the birth of comparative philology – and thus of modern linguistics (e.g. Cannon 1990; see Campbell 2007 for criticism) – Sir William Jones claimed in his speech at the Royal Asiatic Society of Bengala (1786) that "no philologer could examine [Sanskrit, Greek, and Latin], without believing them to have sprung from some common source, which, perhaps, no longer exists". This statement set the goals for generations of

indoeuropeanists in the century that followed, culminating in the groundwork by Bopp (1816) and the synthesis by Schleicher (1861–1862). Comparative philologists aimed at reconstructing a common hypothesized source language, by mapping similarities and differences between several current and extinct languages.

Crucially to our discussion, by including or focussing on extinct languages (as exemplified in Jones' quote above), indoeuropeanists had to restrict themselves to the use of *written* evidence, which is symbolic, linear, and underspecified – just as we defined canonical forms above. The use of canonical forms as the main tool of comparative philology can be illustrated with a passage from the first chapter of Saussure's *Cours*:

> a comparison of the paradigms of Latin *genus* (*genus, generis, genere, genera, generum,* etc.) and Greek (*génos, géneo, génei, génea, genéōn*, etc.) reveals nothing. But the picture changes as soon as we add the corresponding Sanskrit series (*ǵanas, ǵanasas, ǵanasi, ǵanasu, ǵanasām*, etc.). A glance reveals the similarity between the Greek forms and the Latin forms. If we accept tentatively the hypothesis that *ǵanas* represents the primitive state – and this step facilitates explanation – then we conclude that *s* must have fallen in Greek forms wherever it occurred between two vowels. Next we conclude that *s* became *r* in Latin under the same conditions. (1916: 2)

Finding a link between Sanskrit, Greek, and Latin forms thus relied, first, on isolating a unit (*s* in our example) and, second, on applying to this unit a variety of processes ("falling" or "becoming *r*"). But, while indoeuropeanists (and later on neogrammarians) put large emphasis on the study of such processes, they did not deal extensively with the concepts of units and of their isolation. After all, by working with written evidence using alphabetic scripts, constant temporally discrete units and their sequencing are almost self-evident notions (see also Section 9.3.3). This might help explain why work on the notion of phoneme, for example, only started towards the end of the nineteenth century, and with linguists (such as Kruszewski) whose interests lied more with the synchronic study of living languages than with diachronic reconstructions.

9.2.2 Semiology, synchrony, and spoken words

Canonical forms in the sense of linear sequences of discrete symbols remained the main tool also for linguists who started to focus on the synchronic aspects of living languages. In his *Cours*, Saussure redefined the goals of linguistics in the broader frame of semiology. Language is presented as "a system of signs that express ideas" (1916: 16). More precisely, linguistic signs are famously said to unite two psychological entities: "not a thing and a name, but a concept and a sound-image" (p. 66); the latter two being subsequently renamed *signifié*

(signified) and *signifiant* (signifier, p. 67). Both are psychological entities, constructs of the mind: the signified is not the object, and the signifier is not the sound either. But, what exactly are these ideas, or concepts, or *signifiés*?

In the following pages of his *Cours*, while discussing a variety of aspects about the "life" of linguistic signs, Saussure provides several examples of signs: Latin *arbor* (tree); French *sœur* (sister), and two pairs we will be focussing upon below: the signifiers of French *bœuf* and German *Ochs* (ox), and the historically related Latin *necāre* (to kill) and French *noyer* (to drown). For each of the four signs, the signified is a lexical (denotative) meaning. If the research goal is creating a system of linguistic signs, it is obvious to start with those signs for which the signified is clear-cut. Words meet this comfort criterion. In contrast, intonation and stress patterns do not; and this is not only because they convey pragmatic meanings (in Western Germanic languages) whose intricacies are by far less self-evident to the language user. In addition, it is also hard to pinpoint the syntactic and paradigmatic structure in which these stress and intonation patterns are organized (Baumann et al. 2016; Kügler et al. 2015; Ladd 2008; Torreira and Grice in press; Wagner et al. 2015). So, even if Saussure had had access to concepts such as pitch accents, edge tones, and descriptive categories such as partial topic and contrastive focus when starting the discussion on the arbitrary nature of the sign, we are confident he still would have stuck with the *bœuf*/*Ochs* example.

Crucial to our discussion, understanding words as the chief instantiation of signs makes them easier to handle not only on the side of the signified but also on the side of the signifier. Unlike the prosody of an utterance, whose representation is chronically insufficient in written language and is still largely non-consensual among researchers (Kügler et al. 2015), words seem to already come with a representation of their own. This representation is remarkably similar to its written form (with the exclusion of some "transduction" details), as exemplified in the following passage:

> The idea of "sister" is not linked by any inner relationship to the succession of sounds *s-ö-r* (*sœur*) which serves as its signifier in French; that it could be represented equally by just any other sequence is proved by differences among languages: the signified "ox" has as its signifier *b-ö-f* (*bœuf*) on one side of the border and *o-k-s* (*Ochs*) on the other. (p. 68)

These signifiers (*s-ö-r, b-ö-f, o-k-s*) inherited an important property from their written counterparts (*sœur, bœuf, Ochs*): they were conceptualized as consisting of linear sequences of discrete symbols. The influence exerted by written forms becomes evident when Saussure discusses language change. Rather than seeing language change as a consequence of either phonetic changes of the signifier or of changes in meaning of the signified, Saussure sees language change as the consequence of a shift in the relationship between signified and signifier. For example,

"Latin *necāre* (to kill) became *noyer* (drown) in French" (p. 75). But, what actually *is* the signifier, that is, the sound-image of Latin *necāre*? No doubt Saussure did have his mental representation of how *necāre* sounded. However, this imagined sound pattern had been deduced from work on written sources. That is, Saussure took for granted that grapheme sequences translate into sound sequences. This is only possible if it is beyond question that sequences of sounds are a valid concept and that these sequences always occur in the same complete (i.e. full) form.

The same passage on killing and drowning already indicates the difficulties of using written evidence for a reconstructing signs over centuries:

> If instead of comparing Classical Latin *necāre* with French *noyer*, we contrast the former term with *necare* of Vulgar Latin of the fourth or fifth century meaning "drown" the case is a little different; but here again, although there is *no appreciable change in the signifier*, there is a shift in the relationship between the idea and the sign. (p. 75, our emphasis)

The two Latin written forms only differ in the presence of a length diacritic on the vowel of the penultimate syllable. From this point of view, there is indeed no appreciable change in the signifier (at least when compared to the change in the signified). But, this point of view ignores that the loss of distinctive vowel length, and thus of the correlation between syllable weight and stress placement which operated in Classical Latin, had a massive impact on the new Latin system. Moreover, the Vulgar stress had probably nothing to do with the Classical one, both in terms of its function and its substance (cf. Burkard 2014). So, beyond what is contained in the mere chain of symbols, the signifier's "sound image" as well as its embedding in linguistic structure probably *did* change considerably over the first five centuries of the first millennium; the change from *necāre* to *necare* becomes all the more significant as we step away from the words' written representations as symbolic sequences of discrete units.

9.2.3 The phoneme

At about the beginning of the twentieth century, the consolidated use of written evidence suggested representing the signifier as a linear sequence of discrete entities, and the desire to ground semiology on intuitively accessible, tangible contrasts meant that the signified had to be a lexical meaning. In the 1930s, these two developments intertwined with the debate on the nature of the phoneme, and crystallized in the notion of distinctiveness. Phonemes were understood as discrete parts of a linear sequence, and as minimal units whose role was to distinguish between lexical meanings. By focussing on lexical distinctions, representations could be kept as minimalistic as possible, and

remained underspecified with respect to oppositions of meaning at the sentence or discourse level. This approach lies at the heart of the notion of canonical forms, which revolves around the concepts of linearity, discreteness, and (minimalistic and lexical) distinctivity.

One might wonder if what we have here is an example of the famous "cobra effect". The cobra effect means that a concept or an activity that is developed in dealing with a problem, for example, with the (implicit) intention to solve that problem, actually makes the problem worse (Siebert 2003). In a nutshell, although it is probably difficult to imagine for younger researchers, there was a time when there was no differentiation between phonetics and phonology, that is, they constituted one integrated discipline (Ohala 2004). The basic assumption in those days was that speech sounds were steady invariable units, and words and sentences were created by concatenating these units (Menzerath and de Lacerda 1933). Initial evidence of, for example, between-speaker variability and the discovery of variation due to, for example, co-articulation or the assimilation, lenition, and elision of entire speech sounds made linguistics separate these messy observations and measurements from the nicely ordered, invariable structure of language and created the need to find out how the former can be made consistent with the latter. It is obvious that the "former" is the object of phonetics, or, more specifically, experimental and instrumental phonetics, and the latter has continued as the object of the separate discipline of phonology. Not least because of the many different competences that are required to study communication, sharing the work in this way seems not a bad idea. One group of researchers takes care of the measurements and the physical nature of speech; and the other group of researchers uses these insights to develop and elaborate models and representations of sound systems.

However, what happened in practice was what Ohala (2004: 136) called the "estrangement of phonetics and phonology". That is, over time, the ties between phonetics and phonology were successively weakened or cut and that, at least in some cases, phonetic evidence was fit into phonological concepts instead of developing phonological concepts on the basis of phonetic evidence (Kohler 1995). With respect to the cobra effect, the point is that it was, amongst other things, the discovery of reduction that favoured the rise of the phoneme. However, instead of facilitating reduction research, for instance, in the form of variant-to-category mappings, the phonemic concept, together with the overarching framework of canonical forms and lexical meanings, rather blocked or hampered studies on reduction.

Nowadays, efforts are made to bring phonetics and phonology closer together again, and studies on reduction that, for example, blurred the dividing lines between segments and prosodies as well as between co-articulation and

phonological processes (e.g. Wood 1996) can be seen as major new ties in this development. Finally, note that it was again computer technology which significantly shaped all stages of this development. It contributed to the discovery of reduction, it loosened the ties between phonology and phonetics by pushing the latter constantly further away from a traditional humanistic discipline (Ohala 2004), and, as was outlined in Section 9.1, it finally contributed to leverage reduction research and bridge the gap between phonetics and phonology.

9.3 Why not canonical forms?

While canonical forms enjoyed a lasting success in the research and teaching practices of linguists – whose usefulness as a heuristic instrument we also acknowledge – the very building blocks behind this notion were already put under scrutiny in the first half of the twentieth century. In the following sections, we briefly review three threads of work in this spirit, focussing respectively on non-lexical meaning (Section 9.3.1), on non-linear phonological representations (Section 9.3.2), and on the impact of alphabetic writing on phonemic awareness (Section 9.3.3).

9.3.1 Reduction and non-lexical meaning

We suggested above that early developments of linguistics as semiology drew the attention of researchers to meaning at the lexical level. This research agenda relegated to the background some insights on the full scope of meaning which were already well established in European culture. In this respect, it is illuminating to quote a famous passage from Dostoevskij's *A Writer's Diary* (1873–1881) on an exchange between six drunkards, revolving around a single swear word (for a recent variation on the theme, see Simon et al. 2002):

> One Sunday night I happened to walk for some fifteen paces next to a group of six drunken young workmen, and I suddenly realized that all thoughts, feelings and even a whole chain of reasoning could be expressed by that one noun, which is moreover extremely short. One young fellow said it harshly and forcefully, to express his utter contempt for whatever it was they had all been talking about. Another answered with the same noun but in a quite different tone and sense – doubting that the negative attitude of the first one was warranted. A third suddenly became incensed against the first and roughly intruded on the conversation, excitedly shouting the same noun, this time as a curse and obscenity. Here the second fellow interfered again, angry at the third, the aggressor, and restraining him, in the sense of "Now why do you have to butt in, we were discussing things quietly and here you come

> and start swearing". And he told this whole thought in one word, the same venerable word, except that he also raised his hand and put it on the third fellow's shoulder. All at once a fourth, the youngest of the group, who had kept silent till then, probably having suddenly found a solution to the original difficulty which had started the argument, raised his hand in a transport of joy and shouted … "Eureka", do you think? "I have it"? No, not "Eureka" and not "I have it"; he repeated the same unprintable noun, one word, merely one word, but with ecstasy, in a shriek of delight – which was apparently too strong, because the sixth and the oldest, a glum-looking fellow, did not like it and cut the infantile joy of the other one short, addressing him in a sullen, exhortative bass and repeating … yes, still the same noun, forbidden in the presence of ladies but which this time clearly meant "What are you yelling yourself hoarse for?". So, without uttering a single other word, they repeated that one beloved word six times in a row, one after another, and understood one another completely.

One might wonder whether providing an account of such an exchange is a matter that linguistics needs to deal with. Indeed, the passage above has been cited several times in the history of psychology and linguistics (e.g. Albano Leoni 2009; Vygotskij 1934; *inter alia*) to demonstrate the multifaceted nature of language and communication. Especially at Vygotskij's times, the idea of further layers of meaning besides the lexical layer was largely controversial. Jakubinskij (1923) emphasized the importance of dialogue over monologue, and Spitzer (1921) claimed the right to study all forms of linguistic exchange when publishing letters from Italian war prisoners,[2] but these were no mainstream positions. Similar concerns about the importance of meaning beyond the lexical level drove Benveniste (1974). He suggested that language should be studied not only in a "semiotic mode", but also in what he refers to as a "semantic mode", in which the view on language is widened to include the situation, the context, and the activity of the language user. This line of thought contributed eventually to the study of conversational analysis and the phonetics of talk in interaction, where neither meaning is construed in terms of a lexical network nor sound structure is understood in terms of linear discrete units (Ogden 2012, see also Section 9.3.2).

The crux of the matter is that, if canonical forms are built around lexical meaning, then, by including dialogue, discourse, and interaction in the scope of linguistic meaning, one is also questioning the validity of canonical forms as truly viable tools for the study of language. There are a number of major papers in which this issue is more or less explicitly addressed. One of them is Lindblom's

2 "The reader will perhaps find unnecessary to publish all these clumsy meaningless texts, and think one might well write down and publish coffee table conversations or fish merchants' gossips. To this I reply in Italian: *Magari*! if only the greatest possible number of everyday conversation was published! From them, psychologists and linguists could have more to learn than from their beloved written sources" (our translation).

(1990) paper on the H&H theory. In citing and outlining this seminal work, researchers often pay little or no attention to the fact that Lindblom's aim was not to explain phonetic variation in general. Rather, his theory was to explain the phonetic variation and its conditioning factors that are "required for successful lexical access" (p. 405). The key assumption of the H&H theory is that the degree of reduction is exclusively determined by the speaker's strive for articulatory economy (which includes anticipating the interlocutor's top-down processes in speech perception) on the one hand, and the listener's need for sufficient discriminative power in the speech signal on the other. However, Lindblom himself stresses that this assumption of only two antagonistic forces that create the one-dimensional reduction continuum from hypo to hyper is a "deliberate simplification that is likely to be revised in the course of future work" (p. 419). Moreover, this statement about the one-dimensional simplification of H&H is made in the context of the fact that speech is "produced not only in the laboratory but also in its natural, ecological settings" (p. 418).

Lindblom's theory inspired many researchers and was refined several times, for example, by Aylett and Turk (2004; see Chapter 2 of Clopper and Turnbull). However, it has not been revised to date with respect to the simplification that Lindblom pointed out, see Niebuhr (2016) for an in-depth discussion of this fact. This is true although there is a growing body of evidence for (at least) a second dimension that drives the degree of reduction: communicative (i.e. non-lexical) meanings and functions. For example, we know for quite a long time from studies like those of de Jong (1995) and Harrington et al. (1995) that accentuation, or higher prominence levels in general, basically mean higher effort on the part of the speaker in terms of both prosodies and segments. Not so well known by now is that even more effort is put into those accents that signal new, unexpected or contrastive information (Chen et al. 2002; Dahan and Bernard 1996; Mücke and Grice 2014). In contrast, being ironic typically means investing less effort into speech production. This applies in particular to sarcastic utterances and again involves both the prosodic *and* the segmental levels (Byrant 2010; Niebuhr 2014).

Furthermore, reduction also plays a role in the syntagmatic structuring of the speech signal. For example, Local et al. (1986) and Docherty et al. (1997) showed for English that reduction variation in word-final plosives, formerly considered to be purely random, is actually systematic in that speakers reduce less at turn-final than at turn-internal phrase boundaries. Niebuhr et al. (2013) recently replicated this finding for word-final <#-en> syllables in German. Going beyond Local et al. and Docherty et al., they also conducted a perception experiment showing that listeners do in fact use the word-final segmental reduction levels to predict the end of the speaker's turn (see Graupe et al. 2014). Similarly, sounds of adjacent syllables show less strong assimilations of each other's features when there is a word boundary in

between, and again less when there is a phrase boundary in between (Kuzla 2009). Like Graupe et al. (2014), Kuzla et al. (2010) also attested the perceptual relevance of this relationship between the boundary level in the prosodic hierarchy and the degree of assimilation. In addition, regressive assimilations are much more frequent and strong than progressive assimilations, thus causing word-initial syllables to be less strongly reduced than word-final syllables, which can function as a cue to syntagmatic structure (cf. Sproat and Fujimura 1993; Vennemann 1972).

At the more social level of attitudes, speaker attributes, and pragmatic meanings, Plug (2005) analysed Dutch corpus data and found evidence for his assumption that disagreeing utterances are marked by significantly fewer and/or less strong segmental reductions. Schubotz et al. (2015) showed by means of a sub-sample of spontaneous conversations between American English speakers in the Ohio Buckeye Corpus that "discourse markers are realized with lenited segments when compared to their lexical counterparts" (p. 377) and that this reduction is stronger for younger than for older speakers. In Chapter 2, Clopper and Turnbull summarize an experiment whose result was that socio-indexical information, that is, the speaker's regional/dialectal background, is marked more strongly in the same contexts that lead to phonetic reduction. However, given Chapter 3 of van Dommelen, no similar link seems to exist between reduction and L2 speech. Ernestus and Smith note in their chapter that reduction correlates with socio-economic status and speaker gender.

We ourselves recently conducted a perception experiment in which we varied, in three steps from canonical through moderately reduced to extremely reduced, the degree of reduction of both segments and prosodies (pitch-accent ranges and stress-induced lengthening) in a constant test sentence. The test-sentence conditions were produced by a large number of speakers so that each listener heard each reduction condition from a different speaker. Moreover, the combinations of reduction conditions and speakers were balanced across the listener sample. Listeners judged these combinations with respect to 13 different speaker attributes. Results were analysed by means of a three-way ANOVA, based on the between-subjects factors Segmental Reduction, Prosodic Reduction, and Speaker Attribute. Significant main effects of Segmental Reduction and Prosodic Reduction, as well as significant interactions of these two factors with Speaker Attribute clearly showed that segmental *and* prosodic reductions both do affect, in attribute-specific directions and orders of magnitude, how a speaker is perceived. For example, being vain was associated with less segmental but more prosodic reduction. Furthermore, unreduced canonical speech made speakers sound least tired, clumsy, and scatty, but most educated and optimistic. Sounding maximally athletic, sincere, sociable, and composed required moderate degrees rather than no or high degrees of segmental and prosodic reduction (see Niebuhr 2017).

In summary, research on reduced (or, more generally, non-canonical) forms is often most fruitful and convincing when dealing not with lexical contrasts, that is, the traditional objects of canonical forms, but with interactional and pragmatic meanings. Particularly when it comes to attitudinal meanings and speaker attributes the reduced form can actually tell *more* than the full one. We illustrate and conclude this line of argument with Hawkins' example on the use of [ə̰ə̰ə̰], a massively reduced form for English *I don't know*, which

> could allow successful communication between relaxed family members. For example, it could be said by B when A asks B where the newspaper is, and B does not know, but does not feel that she needs to stop reading her book in order to help find it. Person A should understand from this that he should not expect help in looking for the newspaper, and should either stop talking to B, or introduce a more interesting topic. (Hawkins 2003)

9.3.2 Reduction and non-linear representations

Another path that leads to questioning canonical forms comes from research on non-linear representations of sound structure. Interestingly, the idea that words contain more than a sequence of phonemes can be glimpsed even in Trubeckoj's work:

> The signifier aspect of every word in the system of a language can be analyzed into phonemes, that is, it can be represented by a particular sequence of phonemes. Of course, the matter should not be oversimplified. The phonemes should not be considered as building blocks out of which individual words are assembled. Rather, each word is a phonic entity, a *Gestalt*, and is also recognized as such by the hearer, just as an acquaintance is recognized on the street by his entire appearance [...] As a *Gestalt*, each word always contains something more than the sum of its constituents (or phonemes), namely, the principle of unity that holds the phoneme sequence together and lends individuality to a word. Yet in contrast with the individual phonemes it is not possible to localize this principle of unity within the word entity. Consequently one can say that each word can be *completely analyzed* into phonemes, that it *consists of* phonemes. (1939: 35)

In this passage, while advocating the possibility of analysing words as linear sequences of phonemes, Trubeckoj suggests that words also have a holistic silhouette (along the lines of the *cement* of Kruszewski 1883[3] and the *Klanggesicht* of Bühler 1934; see Albano Leoni 2009). This "phonetic silhouette" is also claimed

3 "A sound complex cannot be considered a mechanical juxtaposition of a certain quantity of independent sounds. When combining with one another, sounds [...] accommodate themselves

to be important in perception, but since its features cannot be consistently associated with specific sound segments within a word, it is kept out of the scope of phonology. This very consequence is rejected by research on non-linear representations of sound structure, which seeks to provide a description of such non-local features and, in doing so, necessarily questions the adequacy of an approach based on exclusively linear representations.

It is beyond the scope of this chapter to provide a detailed account of non-linear approaches to phonology (e.g. see Chapter 8 of Espy-Wilson et al. for an outline of Articulatory Phonology and the Task-Dynamic Model of Speech Production). We will merely provide pointers to the relevance of this line of research for the notions of canonical and reduced forms. In particular, the notion of *prosodies* as developed by Firth (1948) is crucial insofar as it represents a bridge between Trubeckoj's neglected concerns on the one hand and the later literature on reduction on the other. Prosodies are defined as abstractions that concur to describe "word structure and its musical attributes", thus going beyond the linear and discrete representation provided by "the total phonological complex" (Firth 1948: 123). In this sense, Firth's prosodies relate to the phonetic silhouettes that hold words together, rather than to the small coloured bricks that account for differences in the signifier:

> Let us regard the syllable as a pulse or beat, and a word or piece as a sort of bar length or grouping of pulses which bear to each other definite interrelations of length, stress, tone, quality – including voice quality and nasality. The principle to be emphasized is the *interrelation of the syllables*, what I have previously referred to as the *syntagmatic relations*, as opposed to the *paradigmatic* or *differential relations* of sounds in vowel and consonant systems, and to the paradigmatic aspect of the theory of phonemes, and to the analytic method of regarding contextual characteristics of sounds as allophones of phonematic units. (1948: 128)

Work on such non-local syntagmatic interrelations has continued challenging a strict linear-based approach, as in the case of evidence from so-called short and long domain /r/-resonances in English (Heinrich et al. 2010; Kelly and Local 1986), which are responsible for pairs such as *miller* and *mirror* having pervasive acoustic differences beyond the intervocalic material.

Prosodies in this sense are seen by Kohler (1999) as what is left of a word (or group of words) when uttered in contexts favouring hypoarticulated speech. Such "articulatory residues may persist as non-linear, suprasegmental features of syllables, reflecting, e.g., nasality or labiality that is no longer tied to specific segmental units" (Kohler 1999: 89). That is, to use a paradoxical image, under the heat of

---

to one another. This accommodation is the *cement* which transforms several sounds into one integral complex." (Kruszewski 1883: 63)

spontaneous hypoarticulated communication, it is the phonetic silhouette that remains, not the building bricks that flesh out this silhouette. The fullest account of this approach is given in Niebuhr and Kohler (2011), who define articulatory prosodies as "distinctive suprasegmental vocal-tract and phonation features that identify words in spite of segmental reduction" (p. 320). Here, prosodies are seen as constituting the "phonetic essence" of words, and thus are given a sort of ontological primacy over segments. In their follow-up paper, Kohler and Niebuhr (2011) consider the example of German *Ihnen* (dative of courtesy pronoun), which is analysed in its canonical form as [iːnən], but which is often uttered by speakers (and recognized by listeners) as [iːn̩] or [nʲnʲ]. These reduced forms

> can be related to the same class (i.e. *Ihnen*) without an elaborate derivation from one canonical representation, because they both contain palatality and long alveolar nasality, as do other intermediate degrees of reduction. This means that all phonetic forms of this word must contain these features; they constitute the *phonetic essence* of *Ihnen*. This concept of phonetic essence may be assumed to apply to function words generally and possibly even to all lexical items. The phonetic essence of a lexical item manifests itself either in segmental units in the less reduced forms or as articulatory prosodies in more extreme reduction, where it appears to be sufficient for the listener to identify the word. (Kohler and Niebuhr 2011)

In this sense, research on reduction helps uncover the *fil rouge* that runs through Kruszewski's *cement*, Bühler's *Klanggesicht*, Trubeckoj's *principle of unity*, Firth's *prosodies*, and Niebuhr and Kohler's *phonetic essence*, which might also be glimpsed in Johnson's (2004) *islands of reliability* and Ernestus and Smith's *core properties* (Chapter 5). This connection is sometimes subterranean, sometimes explicitly acknowledged. And it ultimately joins ends with another thread of research which questions the viability of canonical forms: the issue of alphabetic writing and phonemic awareness.

9.3.3 Reduction and alphabetic writing

The implicit role of written alphabetic representation has surfaced several times in linguistic research throughout the course of the twentieth century. The positions of individual researchers differ in terms of the strength of the conclusions drawn, but all converge towards the need of questioning the viability of a representation of sound structure based on sequences of discrete units.

An early formulation of the problem, which joins ends with our discussion in Section 9.2.1 on the importance of written evidence in the early phases of modern linguistics, can be found again in Firth (1948):

> The development of comparative philology, and especially of phonology, also meant increased attention to transliteration and transcription in roman letters. Sir William Jones was not in any position to understand how all this might contribute to the tendency, both in historical and descriptive linguistics, to phonetic hypostatization of roman letters, and theories built on such hypostatization. In introducing my subject I began with sounds and the Roman alphabet which has determined a good deal of our phonetic thinking in western Europe. (pp. 125–126)

Alphabetic scripts, which represent words using sequences of symbols relating to units smaller than a syllable, are thus considered by Firth as the cause (and not the consequence!) of our tendency to think of words as composed of segments. The alternative view sees alphabets as a proof for the pre-existence of (psychologically or ontologically) meaningful segments. Fowler (2010: 58) discusses alphabetic writing systems, observing that "Their inventors must have had the impression that the spoken language had units to which the letters would correspond. Yet they had no alphabetic writing system to give them that impression". This notion of alphabet as an *invention*, while apparently intuitive at first sight, had actually been challenged by historical accounts of the development of writing systems, as in Gelb (1952), Sampson (1985), and, perhaps with slightly larger resonances in the linguistics research community, Faber (1992). This line of research suggests that alphabets are a *discovery*. That is, phonemic awareness is the consequence of the emergence of alphabetic systems out of the adaptation of writing systems across languages. In Faber's words, when discussing the derivation of the Greek alphabet from its Canaanite sources (in which vowels were not represented),

> it is not necessary to base an explanation for the structure of the Greek alphabet on the unattested existence of an unknown genius. The names of the Greek letters *alpha*, *beta*, etc., meaningless in Greek, have clear sources in a Canaanite acrophonic tradition, whereby each sound is associated with an object whose name begins with that sound. This fixed order of a traditional, invariant list is comparable to modern radio alphabets like *able*, *baker*, *Charlie*, etc. [...] Transmission of the Canaanite script using the acrophonic principle would have led to the *mis*interpretation of several Canaanite consonant symbols as representing vowels instead. The Canaanite words *ʔalpa* 'cow', *he* 'ʔ', *yoda* 'hand', and *ʕayna* 'eye', standing for /ʔ/, /h/, /y/, and /ʕ/, would have been perceived by speakers of a language in which, as in Greek, these sounds did not occur, as beginning in [a], [e], [i], and [ɑ], respectively. Thus, Phoenician [ʔalpa], with an initial [ʔ] became Greek [alpa], with no [ʔ]. (1992: 126)

Kohler (1995) takes an intermediate position in the debate about whether the invention of alphabetic systems is a consequence of phonemic awareness or vice versa. He claims that an alphabetic writing system has been developed only once in human language evolution, with all further systems being derived from this first system. In his opinion, the idea of representing complex sound patterns by means of sequences of discrete symbols was a direct consequence of the

three-consonant roots and their association with semantic fields in the lexicon of the Semitic language family. This morpho-phonological peculiarity favoured the emergence of mental models in which consonants and vowels were represented as separate elements, so that these elements, in turn, were finally represented by separate symbols in written language. Thus, Kohler sees phonemic awareness as a prerequisite for the development of alphabetic writing systems but stresses at the same time that phonemic awareness is not automatically created by inherent properties of linguistic structure. Rather, it needs very special conditions to occur.

We have already mentioned Firth's claim on how the Roman alphabet "has determined a good deal of our phonetic thinking in Western Europe". This fits in with the view expressed in O'Connor (1983: 441), according to which native speakers' analyses of a language are reflected (albeit not always systematically) in the structure of its orthography. Along these lines we might place evidence from Morais et al. (1979) on the fact that illiterates might have little awareness of segments, or none at all. Rather, there is increasing evidence that the syllable rather than the phone(me) is the basic unit of speech, see Greenberg (1996) for a summary with a focus on speech perception. Also famous phenomena like "phonemic restoration" are not inconsistent with this assumed primacy of the syllable. More detailed investigations confirm that restoration occurs but suggest that what listeners restore are syllabic rather than phonemic units (cf. Niebuhr 2011).

Ladefoged (1984) looks at the debate about the effects of alphabetic writing systems on spoken language representation from a meta-linguistic point of view. He states that the apparence of the Greek alphabetic system, "produced out of the spare symbols of a syllabary", set in motion a "startling conspiracy", tricking linguists into thinking that only because speech can be described in terms of segments, then language must also be structured in that way. Ladefoged elaborates on the classical example of Lindblom et al. (1984) on termite nests (see also Bybee 2001), which appear to the outside observer as having a structure revolving around pillars and arches, but are ultimately built by a simple behavioural pattern – the accumulation on grains of earth on spots on the ground containing pheromone secretions:

> Phonemes may be like arches in termite nests, visible to outside observers, but having no meaningful role in the activity of the individuals producing them. Speech *appears* to be composed of sequences of segments because of the interactions of the different systems of which it is composed. The complex gestures involved in producing syllables have diverse parts that look as if they are categorically distinct. We call these diverse parts vowels and consonants, but we must always remember that these are just names for readily distinguishable aspects of the stream of speech. Those of us who have been exposed to an alphabetic tradition may be influenced so that we are very conscious of the possibility of describing speech in terms of units of this kind. (Ladefoged 1984: 93–94)

Here we find echoes of Firth's views both on phonemic awareness as a product of alphabetical training[4] and on the polysystemic nature of sound structure (see also Hawkins 2003).

Port (2010) further elaborates on these ideas. He sees language as a "social institution that is shaped by generations of users" (much alike termite nests), which is only poorly described by segmental-based representations. These are "partially a side effect of our years of literacy education and extensive practice of literacy" (cf. Firth's "phonetic hypostatization of roman letters"). He thus rejects canonical representation of words as minimalistic strings of discrete units, and advocates richer representations. Such representations would be stored in the memory of language users not as "low-bitrate" vectors of phonemes, but as rich traces, containing detailed phonetic information which is used in our processing of linguistic variability (e.g. idiolectal and sociolectal). In doing so, Port brings together the reflexion on the "alphabetic fallacy" with research on episodic memory.

Crucial to our purposes, research on reduction has a lot to contribute to the debate on episodic memory. Johnson (2004) has convincingly shown that the traditional view of a mental lexicon built around individual word representations consisting of sequences of discrete units is inadequate to account for auditory recognition of connected speech. "Massive" reduction, as the one relating forms such as [dəvɪ̃ɪ̯ fʌdʒ] to *divinity fudge* in Stampe (1973) example, are shown to occur frequently in spontaneous speech. Speakers' knowledge of words is not adequately represented by entries such as *divinity* = /dəvɪnəti/ in two respects – in assuming that a word is represented with a single entry, and in assuming that this entry is composed by a sequence of discrete phonemes.

9.4 Reduction and canonical forms: Assessing the partnership

Research on reduction has continuously made us question theoretical assumptions and analytical practices which were, until then, commonplace in speech sciences. Among these established assumptions or practices, on which the present chapter focused, are canonical forms. So can we conclude from what we have briefly summarized in Sections 9.2 and 9.3 that reduction and canonical

4 "We ABC people, as some Chinese have described us, are used to the process of splitting up words into letters, consonants and vowels." (Firth 1948: 122)

forms do not have a good partnership? As is typical of scientific conclusions, the answer is not that simple.

Traditional canonical forms and their basic building blocks, the phonemes, are obviously useful for those who have to abstract away from reduction phenomena in order to, for example, develop and deal with alphabetic writing systems. Also, the chapter of van Dommelen suggests that canonical forms are a practical instrument for those who need a clear and easy entry to teaching and learning languages. Moreover, van Dommelen's chapter demonstrated in accord with the chapters of Cutugno et al., Adda-Decker and Lamel, and Clopper and Turnbull that for us, the researchers, canonical forms are handy points of reference that help us detect and describe the reduced and variable sound patterns that speakers produce. In other words, traditional canonical forms have a practical advantage for all of those who *talk about* speech communication. However, in view of the accumulating evidence that is provided by this book and summarized in Section 9.3, we dare to join the voices of those who claim that traditional canonical forms are likely not as relevant for those who actually *do* speech communication (e.g. see Kohler 2000).

This need not mean that, in understanding speech communication, we should throw over board the *basic idea* of canonical forms altogether. Evidence from perceptual restoration (which may not be simply phonemic restoration, see Niebuhr 2011), phoneme monitoring, the McGurk effect (Cox et al. 1999), segmental intonation (Niebuhr 2012), and many other findings clearly show that, compared to the acoustic signal, the hearer's speech perception can well be richer in phenomenological or structural respects. In fact, alternative concepts of proper, perceptually relevant reference forms that underlie or interconnect reduction phenomena are already waiting to be further elaborated. These alternatives are not based on linear, symbolic, and minimalistic representations. Rather, they put focus on *rich phonetic detail* in the form of "gross and subtle acoustic characteristics" (Ernestus and Smith, Chapter 5) that can be either sub-phonemic or supra-phonemic in that they "may be distributed across [...] sounds" (Espy-Wilson et al., Chapter 8). The keywords that reflect these alternative reference concepts in the chapters of this book include, for example, landmarks (Cole and Shattuck-Hufnagel, Chapter 6), articulatory prosodies, core properties, and phonetic essence (Ernestus and Smith, Chapter 5).

In terms of this next generation of alternative reference concepts, the chapters of this book also gave us an idea of which major questions will drive our investigation of speech communication with respect to reduction phenomena in the future.

First, one of the major questions will be how we define reduction in the future and whether it makes sense to continue using the term at all. For example, in terms of a decrease in duration of a particular linguistic unit, as, for example, in the

chapter of Adda-Decker and Lamel, or, similarly, in terms of "less acoustic-phonetic substance", as in the chapter of Clopper and Turnbull (cf. also Espy-Wilson et al.), the term reduction is quantifiable and hence still useful and intuitive. However, reduction seems to be more and more often equated simply with "abundant phonetic variability" (Chapter 3, van Dommelen) and phonetic variation in general. Speaking of reduction in these contexts is dispensable, as the equation makes no reference to a clear-cut, superordinated reference form anymore. The definition of reduction as "fewer phonetic cues to contrastive phonological units" by Cole and Shattuck-Hufnagel is at least not generally applicable, as empirical studies showed many times now that cues bundled in the form of sound segments can be "recoded" into articulatory prosodies when the sound segments themselves have disappeared, see Ernestus and Smith (Chapter 5). In Chapter 7, Cutugno et al. state that "every sound segment can undergo coarticulation" and then add to this statement that coarticulation is the most widespread form of reduction. Definitions like these are similar to equating reduction to variation, but they are terminologically still more problematic as they relate reduction to individual sound segments and in this way decouple the term from its original foundation, that is, the canonical form at the word level. That is, by stating that individual sound segments are reduced, phonemes become the new canonical forms; and, unless coarticulation affects distinctive features, it is simply impossible to say whether or not a sound segment is reduced (exceptions are cardinal vowels and reduction in the sense of a decrease in duration). For example, which of the allophonic variants of German /x/ ([ç], [x], or [χ]) is reduced against which reference form, and is the English light /l/ the reduced variant of the dark /l/ or vice versa?

We think that using the term reduction is still useful. However, in line with our distinction between talking about versus actually doing speech communication, reduction is useful in the sense of a phonetic parameter rather than as basic aspect of speech cognition and representation. The concept of canonical forms is also not completely obsolete. Yet, it should be further developed to more flexible forms that take into account empirical frequencies of word variants and include, wherever necessary, multi-word expressions rather than individual words (cf. Chapter 4 of Adda-Decker and Lamel). Moreover, it should allow for a counterpart of reduction, that is, *strengthening* (cf. Chapter 6 of Cole and Shattuck-Hufnagel). In fact, it seems that the idea of strengthening as a counterpart of reduction is already in use, but not as a mature, consistently applicable concept that is firmly grounded in a revised framework of canonical forms. Rather, inspired by the H&H terminology, people currently have to use terms such as "hyperspeech" and "hyperarticulated speech" whenever they have the impression that a certain speech pattern or way of speaking exceeds the articulatory or acoustic parameter ranges that can be expected under "normal" circumstances.

Second, closely linked with the first question about the use and usefulness of reduction and canonical forms, future studies will have to deal with the question of reduction in the domain of prosody. It seems to be widely accepted that lexical stress or prominence levels can undergo reduction, see Wagner et al. (2015) and the chapters of Ernestus and Smith, and Clopper and Turnbull. This is probably because stress is quantifiable in terms of duration (a common parameter in measuring reduction anyway) and moreover closely linked to words and lexical meaning. However, what about intonation? Clopper and Turnbull, for example, found lower intonation peaks for those pitch accents that were realized in predictable focus conditions. Are lower intonation peaks instances of reduction? (In the light of the effort code of Gussenhoven 2002, this would probably be the case.) Niebuhr and Hoekstra (2015) showed in a production study that Northern Frisian speakers produce intonation plateaux rather than higher intonation peaks under expressive conditions (including contrastive focus). Is this also a case of intonational reduction? Would it be possible to define an equivalent of canonical forms for intonational units like pitch accents and edge tones? And what about reductions and/or canonical forms of voice quality and loudness? All these important theoretical and empirical questions are currently widely unaddressed.

Third, Cutugno et al. state with regard to reduction that "Spontaneous speech is characterized by a great amount of unpredictable phenomena". Predictability plays an important role in reduction research, for example, in the related fields of psycholinguistic models of word recognition and automatic speech recognition. However, note that humans probably have to rely much less on bottom-up predictability than machines when it comes to word recognition in natural everyday conversation, see Chapters 4 and 8 of Adda-Decker and Lamel, and Espy-Wilson et al.

Cutugno et al. are certainly right with their above statement about the great amount of unpredictable phenomena, especially in view of the great range of within- and between-speaker variations in the pronunciation of words under constant conditions (see Chapters 5, 6, and 8 of Ernestus and Smith, Cole and Shattuck-Hufnagel, and Espy-Wilson et al.). Nonetheless, the amount of unpredictable reduction variation should also not be overestimated, as we have only just begun to uncover the factors that determine – and hence predict – speech reduction. As was summarized in Section 9.3, differences in the degree of reduction have communicative functions, and some of the supposedly random variation became explainable and predictable on this basis. This book is a further big step in advancing our insights into the determining factors of variation in reduction. For example, Ernestus and Smith showed that the rhythmic embedding of a word influences its degree of reduction. Moreover, Clopper and Turnbull add quite a bit of complexity to the already established reduction triggers by showing that these factors are not simply additive but interact in how they affect

individual reduction parameters, and that not only reduction itself but also its triggering factors are scalable. This line of research on identifying and understanding reduction triggers and their interaction should receive special attention in the future. Computer-based analyses of spontaneous dialogues and natural everyday speech recordings will contribute a great deal to address open questions. If lab speech has to be used, we also need a better understanding of how environment and task conditions affect speech production, and whether they can even be adjusted such that they facilitate rather than impede the production of reduced forms (see Chapter 3; Niebuhr 2015).

Last but not least, another major challenge of future research will be to revise the H&H theory of Lindblom (1990) and the alternative frameworks outlined by Clopper and Turnbull such that variation in the degree of reduction is no longer a simple one-dimensional trade-off between economy and comprehensibility. This book showed clearly that reduction is much more complex and involves a lot of additional, partly antagonistic forces; and any future framework that wants to explain phonetic variation has to find a way to address this complexity; see also Niebuhr (2016). Such a model would be a great achievement, also because its value would not be confined to the humanistic fields of the speech sciences. Rather, it would also have practical implications for improving speech technology and developing materials and strategies for (second) language teaching.

In general, as research on reduction is becoming increasingly intense and dynamic, it does not take much to foresee that the number of questions and challenges posed by reduction research will further grow in the future.

References

Albano Leoni, F. 2009. *Dei suoni e dei sensi. Il volto fonico delle parole*. Bologna: il Mulino.

Allwood, J., L. Cerrato, L. Dybkjaer, K. Jokinen, C. Navarretta & P. Paggio 2007. The MUMIN multimodal coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation* 41. 273–287.

Aylett, M. & A. E. Turk 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47. 31–56.

Barbosa, P. A., Z. A. Camargo & S. Madureira 2016. Scripts for the acoustic analysis of speech data. In S. Madureira (ed.), *Sonorities: Expressivity in speech, singing, and reciting*, 164–174. São Paulo: Pontíficia Universidade Católica de São Paulo.

Baumann, S., O. Niebuhr & B. Schroeter 2016. Acoustic cues to perceived prominence levels – evidence from German spontaneous speech. Proceedings of 8th International Conference of Speech *Prosody*, 1–5. Boston, USA.

Benveniste, E. 1974. *Problèmes de linguistique générale II*. Paris: Gallimard.

Bird, S. & J. Harrington 2001. Speech annotation and corpus tools. *Speech Communication* 33. 1–4.
Bopp, F. 1816. *Über das Conjugationssystem der Sanskritsprache in Vergleichung mit jenem der griechischen, lateinischen, persischen und germanischen Sprache*. Frankfurt: Andreäischen Buchhandlung.
Bühler, K. 1934. *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Jena: Fischer.
Burkard, T. 2014. Mythen und freie Erfindungen in der lateinischen Grammatik – das Nicht-Verstehen einer toten Sprache und seine Konsequenzen. In O. Niebuhr (ed.), *Formen des Nicht-Verstehens*, 45–76. Frankfurt: Peter Lang.
Bybee 2001. *Phonology and language use*. Cambridge: Cambridge University Press.
Byrant, G. A. 2010. Prosodic contrasts in ironic speech. *Discourse Processes* 47. 545–566.
Campbell, N. & P. Mokhtari 2013. Voice quality: The 4th prosodic dimension. *Proceedings of 15th International Congress of Phonetic Sciences*, 2417–2420. Barcelona, Spain.
Campbell, N. 2007. Why Sir William Jones got it all wrong, or Jones' role in how to establish language families. In J. Lakarra & J. Hualde (eds.), *Studies in Basque and historical linguistics in memory of R.L. Trask*, 245–264. Bilbao: Universidad del País Vasco/Euskal Herriko Unibertsitatea.
Cannon, G. 1990. *The life and mind of oriental Jones: Sir William Jones, the father of modern linguistics*. Cambridge: Cambridge University Press.
Chen, A., C. Gussenhoven & T. Rietveld 2002. Language-specific uses of the effort code. *Proceedings of 1st International Conference of Speech Prosody 2002*, 211–214. *Aix-en-Provence, France*.
Cox, E. A., L. W. Norrix & K. P. Green 1999. The contribution of visual information to on-line sentence processing: Evidence from phoneme monitoring. *Proceedings of Audio-Visual Speech Processing Conference*, 30–36. Santa Cruz, USA.
Dahan, D. & J.-M. Bernard 1996. Interspeaker variability in emphatic accent production in French. *Language and Speech* 39. 341–374.
de Jong, K. J. 1995. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America* 97. 491–504.
Docherty, G. J., J. Milroy, L. Milroy & D. Walshaw 1997. Descriptive adequacy in phonology: A variationist perspective. *Journal of Linguistics* 33. 275–310.
Dostoevskij, F. 1994 [1873–1881]. *A writer's diary*. Evanston: Northwestern University Press.
Ernestus, M. 2000. *Voice assimilation and segmental reduction in Dutch*. PhD dissertation, University of Utrecht, The Netherlands.
Ernestus, M. & N. Warner 2011. An introduction to reduced pronunciation variants. *Journal of Phonetics* 39. 253–260.
Faber, A. 1992. Phonemic segmentation as epiphenomenon. In P. Downing, S. Lima & M. Nooman (eds.), *The linguistics of literacy*, 111–134. Amsterdam-Philadelphia: Benjamins.
Firth, J. R. 1948. Sounds and prosodies. *Transactions of the Philological Society* 47 (1). 127–152.
Fowler, C. 2010. The reality of phonological forms: A reply to Port. *Language Sciences* 32. 56–59.
Gelb, I. 1952. *A study of writing*. Chicago: University of Chicago Press.
Graupe, E., K. Görs & O. Niebuhr 2014. Reduktion gesprochener Sprache – Bereicherung oder Behinderung der Kommunikation? In O. Niebuhr (ed.), *Formen des Nicht-Verstehens*, 155–184. Frankfurt: Peter Lang.
Greenberg, S. 1996 Understanding speech understanding – towards a unified theory of speech perception. *Proceedings of ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, 1–8. Keele, England.

Gussenhoven, C. 2002. Intonation and interpretation: phonetics and phonology. Proceedings of the First International Conference of Speech Prosody, 47–57, Aix-en-Provence, France.
Harrington, J., J. Fletcher & C. Roberts 1995. Coarticulation and the accented/unaccented distinction: Evidence from jaw movement data. *Journal of Phonetics* 23. 305–322.
Harrington, J. 2010. Acoustic Phonetics. In W. J. Hardcastle, J. Laver & F. E. Gibbon (eds.), *The handbook of phonetic sciences*, 81–129. Oxford: Wiley-Blackwell.
Hawkins, S. 2003. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics* 31 (3–4).373–405.
Heinrich, A., Y. Flory & S. Hawkins 2010. Influence of English resonances on intelligibility of speech in noise for native English and German listeners. *Speech Communication* 52. 1038–1055.
Jakubinskij, L. P. 1979 [1923]. On verbal dialogue. *Dispositio* 4 (11–12). 321–336.
Johnson, K. 2004. Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (eds.), *Spontaneous speech: Data and analysis*, 29–54. Tokyo: The International Institute for Japanese Language.
Jones, W. 1798 [1786]. The third anniversary discourse, delivered 2d February, 1786: on the Hindus. *Asiatick Researches* 1. 415–31.
Kelly J. & J. Local 1986. Long domain resonance patterns in English. *International conference on speech input/output: Techniques and applications*, 304–309. London: Institute of Electrical Engineers.
Kohler, K. J. 1995. Phonetics – a language science in its own right? *Proceedings of 13th International Congress of Phonetic Sciences*, 10–17. Stockholm, Sweden.
Kohler, K. J. 1999. Articulatory prosodies in German reduced speech. *Proceedings of 4th International Congress of Phonetic Sciences,* 89–92. San Francisco, USA.
Kohler, K. J. 2000. Investigating unscripted speech: Implications for phonetics and phonology. *Phonetica* 57. 85–94.
Kohler, K. & O. Niebuhr 2011. On the role of articulatory prosodies in German message decoding. *Phonetica* 68. 57–87.
Kruszewski, M. 1995 [1883]. *Outline of linguistic science*. In K. Koerner (ed.), *Writings in general linguistics*, 35–173. Amsterdam-Philadelphia: Benjamins.
Kügler, F, B. Smolibocki, D. Arnold, B. Braun, S. Baumann, M. Grice, S. Jannedy, J. Michalsky, O. Niebuhr, J. Peters, S. B. Ritter, C. T. Röhr, A. Schweitzer, K. Schweitzer & P. Wagner 2015. DIMA – Annotation guidelines for German intonation. *Proceedings of 18th International Congress of Phonetic Sciences,* 1–5. Glasgow, UK.
Kuzla, C. 2009. *Prosodic structure in speech production and perception*. Nijmegen: MPI Series in Psycholinguistics.
Kuzla, C., M. Ernestus & H. Mitterer 2010. Compensation for assimilatory devoicing and prosodic structure in German fricative perception. In C. Fougeron, B. Kühnert, M. D'Imperio & N. Vallée (eds.), *Laboratory phonology* 10. 731–757. Berlin: De Gruyter.
Ladd, D. R. 2008. *Intonational Phonology*. Cambridge: CUP.
Ladefoged, P. 1984. 'Out of chaos comes order': Physiological, biological, and structural patterns in phonetics. In M. P. R. Van den Broeke & A. Cohen (eds.), *Proceedings of the 10th International Congress of Phonetic Sciences*, 83–95. Dordrecht: Foris Publications.
Ladefoged, P. 2003. *Phonetic data analysis. An introduction to fieldwork and instrumental techniques*. Malden: Blackwell.
Lindblom, B., P. MacNeilage & M. Studdert-Kennedy 1984. Self-organizing processes and the explanation of language universals. In B. Butterworth, B. Comrie & O. Dahl (eds.), *Explanations for Language Universals*, 181–203. The Hague: Mouton.

Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H and H theory. In W. Hardcastle & A. Marchal (eds.), *Speech production and speech modelling*, 403–439. Dordrecht: Kluwer.
Local, J., J. Kelly & W. H. G. Wells 1986. Towards a phonology of conversation: Turn-taking in Tyneside English. *Journal of Linguistics* 22. 411–437.
Mattingly, I. G. 1999. A short history of acoustic phonetics in the U.S. *Proc. 14th International Congress of Phonetic Sciences, San Francisco, USA*, 1–6.
Menzerath, P. & A. de Lacerda 1933. *Koartikulation, Steuerung und Lautabgrenzung. Eine experimentelle Untersuchung*. Berlin and Bonn: Dümmler.
Morais, J., L. Cary, J. Alegria & P. Bertelson 1979. Does awareness of speech as a sequence of phones arise spontaneously? *Cognition* 7. 323–331.
Mücke, D. & M. Grice 2014. The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation? *Journal of Phonetics* 44. 47–61.
Niebuhr, O. 2011. On the domain of auditory restoration in speech. In B. Kokinov, A. Karmiloff-Smith & N. J. Nersessian (eds), *European perspectives on cognitive science*, 1–6. Sofia: New Bulgarian University Press.
Niebuhr, O. & K. Kohler 2011. Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics* 39. 319–329.
Niebuhr, O. 2012. At the edge of intonation – The interplay of utterance-final F0 movements and voiceless fricative sounds. *Phonetica* 69. 7–27.
Niebuhr, O., K. Görs & E. Graupe 2013. Speech reduction, intensity, and F0 shape are cues to turn-taking. *Proceedings of14th Annual SigDial Meeting on Discourse and Dialogue*, 1–9. Metz, France.
Niebuhr, O. 2014. "A little more ironic" – Voice quality and segmental reduction differences between sarcastic and neutral utterances. *Proceedings of 7th International Conference of Speech Prosody*. 1–5. Dublin, Irland.
Niebuhr, O. & J. Hoekstra 2015. Pointed and plateau-shaped pitch accents in North Frisian. *Laboratory Phonology* 8. 1–35.
Niebuhr, O. 2016. Rich Reduction: Sound-segment residuals and the encoding of communicative functions along the hypo-hyper scale. *Proceedings of 7th Tutorial & Research Workshop on Experimental Linguistics*, 11–24. St. Petersburg, Russia.
Niebuhr, O. 2017. Clear speech, mere speech? How segmental and prosodic speech reduction shape the impression that speakers create on listeners. *Proceedings of 18th Interspeech Conference*, Stockholm, Sweden.
Nolan, F. 1992. The descriptive role of segments: Evidence from assimilation. In G. Docherty & D. R. Ladd (eds.), *Laboratory phonology*, 2. 261–280. Cambridge: Cambridge University Press.
O'Connor, M. 1983. Writing systems, native speaker analysis, and the earliest stages of Northwest Semitic orthography. In C. Myers & M. O'Connor (eds.), *The word of the Lord shall go forth*, 439–465. Winona Lake: Eisenbrauns.
Ogden, R. 2012. Prosodies in conversation. In O. Niebuhr (ed.), *Understanding Prosody: The Role of Context, Function and Communication*, 201–218. Berlin and New York: de Gruyter.
Ohala, J. J. 2004. Phonetics and phonology then, and then, and now. In H. Quene & V. van Heuven (eds.), *On speech and language: Studies for Sieb G. Nooteboom*, 133–140. Utrecht: LOT Occasional Series.
Origlia, A. & I. Alfano 2012. Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabification. *Proceedings of 8th International Conference on Language Resources and Evaluation (LREC)*, 997–1002. Istanbul, Turkey.

Plug, L. 2005. From words to actions: The phonetics of eigenlijk in two communicative contexts. *Phonetica* 62. 131–145.
Port, R. 2010. Rich memory and distributed phonology. *Language Sciences* 32. 43–55.
Sampson, G. 1985. *Writing systems*. Stanford: Stanford University Press.
Saussure 1959 [1916]. *Course in general linguistics*. New York: Philosophical library.
Schleicher 1861–1862. *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*. Weimar: H. Böhlau.
Schubotz, L., N. Oostdijk & M. Ernestus 2015. Y'know vs. you know: What phonetic reduction can tell us about pragmatic function. In S. Lestrade, P. de Swart & L. Hogeweg (eds.), *Addenda – Artikelen foor Ad Foolen*, 361–380. Nijmegen: Radboud University Press.
Simon, D., E. Burns & C. Virgo 2002. Old cases [Television series episode]. In D. Simon & R. Colesberry (prod.), *The wire*. New York: Home Box Office.
Siebert, H. 2003. *Der Kobra-Effekt. Wie man Irrwege der Wirtschaftspolitik vermeidet*. München: Piper.
Spitzer, L. 1921. *Italienische Kriegsgefangenenbriefe*. Bonn: Hanstein.
Sproat, R. & O. Fujimura 1993. Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics* 21. 291–311.
Stampe, D. 1973. *A Dissertation on natural phonology*. PhD Diss. University of Chicago.
Strunk, J., F. Schiel & F. Seifart 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC')*, 3940–3947. Paris, France.
Torreira, F. & M. Grice (in press). Melodic constructions in Spanish: Metrical structure determines the association properties of intonational tones. *Journal of the International Phonetic Association*.
Trubeckoj 1939. *Grundzüge der Phonologie*. Prague: Travaux du Cercle Linguistique de Prague.
Vennemann, T. 1972. On the theory of syllabic phonology. *Linguistische Berichte* 18. 1–18.
Vygotskij 1962 [1934]. *Thought and language*. Cambridge, MA: M.I.T. Press.
Wagner, P., A. Origlia, C, Avesani, G. Christodoulides, F. Cutugno, M. D'Imperio, M. D. Escudero, B. Gili Fivela, A. Lacheret, B. Ludusan, H. Moniz, A. Ní Chasaide, O. Niebuhr, L. Rousier-Vercruyssen, A. C. Simon, J. Simko, F. Tesser & M. Vainio 2015. Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic promience. *Proceedings of 18th International Congress of Phonetic Sciences*, 1–5. Glasgow, UK.
Wesener, T. 1999. The phonetics of function words in German spontaneous speech. *AIPUK* 34. 327–377.
Wood, S. A. J. 1996. Assimilation or coarticulation? Evidence from the temporal co-ordination of tongue gestures for the palatalization of Bulgarian alveolar stops. *Journal of Phonetics* 24. 139–164.
Xu, Y. 2013. ProsodyPro – A tool for large-scale systematic prosody analysis. *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP)*, 7–10. Aix-en-Provence, France.

Editor Biographies

Francesco Cangemi (IfL Phonetics, University of Cologne) studied philology in Naples and linguistics in Zurich, and specialized in prosody in Aix-en-Provence within the Marie Curie Research Training Network "Sound to Sense." His dissertation on prosodic detail in Italian focuses on whether and how abstract canonical forms in intonation could be enriched in order to account for various sources of phonetic variability. His current research focusses on the plasticity of phonological categories, and more specifically on how language users can represent highly individual phonetic behavior into their general phonological knowledge.

Meghan Clayards (Department of Linguistics and School of Communication Sciences and Disorder, McGill University, Canada) completed her PhD research in Brain and Cognitive Sciences at the University of Rochester (NY) on the role of phonetic variability in speech perception and learning. She did postdoctoral training at the University of York (UK) as part of the Marie Curie Research Training Network "Sound to Sense" examining how language background influences the production and perception of contextually conditioned assimilation across word boundaries. She joined McGill in 2010 as an Assistant Professor. She has examined how hyperarticulation and prominence affect the production of segmental contrasts within and across individuals and is interested in the production–perception relationship and individual differences more generally.

Oliver Niebuhr (Mads Clausen Institute of the University of Southern Denmark) earned his doctorate in Phonetics and Digital Speech Processing from Kiel University and worked then as a postdoc researcher at linguistic and psychological institutes in Aix-en-Provence and York as part of the interdisciplinary Marie Curie Research Training Network "Sound to Sense." In 2009, he was appointed Junior Professor of Spoken Language Analysis and returned to Kiel University, where he also became director of the Kiel research center "Speech & Emotion." Since 2015, he is the head of the Innovation Research Cluster Alsion and Associate Professor of Communication and Innovation at the Mads Clausen Institute of the University of Southern Denmark. His main research interests include the interplay of segments and prosodies, the multiparametric complexities of prosodic constructions and speech reduction and their subtle but powerful communicative functions, also with respect to human–robot–interaction, persuasive technology, and speech in business contexts.

Barbara Schuppler (Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria) pursued her PhD research at Radboud Universiteit Nijmegen (The Netherlands) and NTNU Trondheim (Norway) within the Marie Curie Research Training Network "Sound to Sense." The central topic of her thesis was the analysis of conditions for acoustic reduction in large conversational speech corpora using ASR technology. Currently, she is working on a FWF-funded Elise-Richter Grant entitled "Cross-layer prosody models for conversational speech." Her research continues to be interdisciplinary; it includes the development of automatic tools for the study of pronunciation variation, the study of reduction and phonetic detail in conversational speech and the integration of linguistic knowledge into ASR technology.

Margaret Zellers (Institute for Scandinavian Studies, Frisian Studies and General Linguistics, Kiel University, Kiel, Germany) conducted her doctoral research at the University of Cambridge (UK), investigating the role of prosodic variation in signaling topic structure in spoken discourse as part of the Marie Curie Research Training Network "Sound to Sense." Her early postdoctoral research, conducted in York (UK) and Stockholm (Sweden), focused on the development of new methodologies for studying segmental and prosodic variation linked to syntactic and discourse structure in conversational speech. Since 2017, she is Junior Professor for Phonetics and Phonology at Kiel University. Her research interests include prosodic and gestural variation related to turn-taking in conversation, as well as factors influencing the perception of prosody in spontaneous speech.

Index

https://doi.org/10.1515/9783110524178-010