DE GRUYTER

ACOUSTIC ANALYSIS OF PATHOLOGIES

FROM INFANCY TO YOUNG ADULTHOOD

Edited by Amy Neustein, Hemant A. Patil

SPEECH TECHNOLOGY AND TEXT MINING IN MEDICINE AND HEALTHCARE

Acoustic Analysis of Pathologies

Speech Technology and Text Mining in Medicine and Health Care

Edited by
Amy Neustein

Volume 7

Acoustic Analysis of Pathologies

From Infancy to Young Adulthood

Edited by
Amy Neustein, Hemant A. Patil

DE GRUYTER

Editors
Dr. Amy Neustein
Linguistic Technology Systems
1530 Palisade Avenue, Suite 28R
Fort Lee NJ 07024
USA
amy.neustein@verizon.net

Prof. Hemant A. Patil
ISCA Distinguished Lecturer 2020–2022
APSIPA Distinguished Lecturer 2018–2019
Room No. 4103, Faculty Block-4
Near Indroda Circle
DA-IICT Gandhinagar
382 007, Gujarat State, India
hemant_patil@daiict.ac.in

ISBN 978-1-5015-1962-8
e-ISBN (PDF) 978-1-5015-1313-8
e-ISBN (EPUB) 978-1-5015-1316-9
ISSN 2329-5198

Library of Congress Control Number: 2020931441

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Cover image: MEHAU KULYK/SCIENCE//PHOTO LIBRARY/Agentur Focus
Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Acknowledgments

This coedited book volume is furthered by the stellar contributions of authors across the globe, including three IEEE fellows, Professors Douglas O'Shaughnessy, Richard C. Rose and Professor Björn W. Schuller, the latter of whom wrote the foreword. We are honored to have a coauthored contribution from Professor Mark Hasegawa-Johnson [Fellow of Acoustical Society of America, and Fellow of International Speech Communication Association (ISCA)].

In addition, we thank the authorities of DA-IICT Gandhinagar and the members of Speech Research Lab at DA-IICT, whose steadfast encouragement has sustained us throughout this lengthy project. In particular, Prof. Patil thanks his students, Dr. Anshu Chittora, Dr. Hardik B. Sailor and Ms. Neeharika Buddha, with whom he had a great pleasure to coauthor several publications related to the leitmotif of the book – infant cry analysis and classification. Special thanks to Series Editor Dr. Amy Neustein, whose patience, trust and confidence in the research capabilities of both the contributing authors and her coeditor helped bring this challenging book project to fruition. Dr. Amy also invited Prof. Patil to write a chapter on Infant Cry Analysis in her book *Advances in Speech Recognition* (2010). This book chapter became the springboard for the two book chapters in this volume as well as the supervised doctoral thesis of Dr. Anshu Chittora. Finally, we acknowledge the generous support, cooperation and patience from the editorial team and production staff at De Gruyter, in particular Leonardo Milla, who guided us properly and made this book possible, and Project Manager Mervin Ebenezer who carefully attended to every detail in production with much efficiency and good spirit.

Dr. Amy Neustein, Series Editor, *Speech Technology and Text Mining in Medicine and Healthcare* (De Gruyter); Editor-in-Chief, *International Journal of Speech Technology* (Springer Nature); Series Editor, *SpringerBriefs in Speech Technology*; Series Editor, *Signals and Communication Technology* (Springer)

Prof. Hemant A. Patil, DA-IICT Gandhinagar, Gujarat, India.
ISCA Distinguished Lecturer 2020–2022
APSIPA Distinguished Lecturer (DL) 2018–2019

https://doi.org/10.1515/9781501513138-202

Computers hearing children's cries and pathologies – a foreword

When the children cry, and computers listen, huge potential for improved health and well-being opens up. This comes not only as parents and caretakers may not be around 24/7 without a pause. Much more, according to Greg Irving and colleagues (2017) investigating "international variations in primary care physician consultation time [in] a systematic review of 67 countries" across 111 publications, general practitioners (GP) have only shockingly limited average consultation time per patient reaching from just 48 seconds in Bangladesh to 22.5 minutes in Sweden. Imagining a reliable, robust, and re-explainable detection, analysis and potential interpretation of vocalizations and spoken language of infants to young adults by machines first provides the possibility to collect such observations from considerably longer time windows with undivided attention as only computers can provide. Furthermore, subsequent analysis of these by computers can be based on precise signal processing and a data experience a single GP could not experience throughout a human life time given available "big" data. Computers could further be free of bias and objective. But even if actual diagnoses and decisions are not made by computers, GPs can be significantly supported by suited pre-selection of informative examples of vocalizations and speech worth paying attention to. The general idea of such bringing health sensing to our everyday environment by means of mobile computing devices is, for example, met by the field of "mobile health," or mHealth as coined by my colleague Robert S. H. Istepanian more than a decade ago.

In this fine collection of five chapters provided by a dozen authors from leading institutions across Asia, Europe and Northern America, the focus is put entirely on the microphone as sensor for the computational acoustic analysis of pathologies, and the target group of this analysis are infants to young adults. The book can be roughly divided into two main parts of interest: early infant vocalizations and pathological speech of young individuals. The focus in the first two chapters is put onto robustly recognizing and classifying infant cries. In relation to this, Chapter 3 investigates toddlers' vocalizations, who are on the autism disorder spectrum. The remaining two chapters address effects of dysarthria on computational modeling. In Chapter 4, pronunciation accuracy is estimated by computing technology. Finally, the last chapter investigates the effect of learning not only of humans, but also computers to improve the recognition of dysarthric speech.

https://doi.org/10.1515/9781501513138-203

In more detail, Chapter 1 discusses techniques for and challenges in infant cry analysis and classification used so far. It further shows novel experimental results showing that spectrographic analysis provides good performance in some pathological cases.

Chapter 2 deals with the recent and popular topic of unsupervised representation learning for acoustic modeling. Here, again with the purpose of modeling infant cries, but to distinguish normal from pathological ones, and in particular by auditory filter-banks. To this end serve convolutional restricted Boltzmann machines. The observed learnt banks perform better but are found to be very distinct from expertly shaped ones.

Chapter 3 discusses the use of wavelet-based and speech modulation spectral features for autism diagnosis based on cries, laughs and other sounds made by toddlers as young as one and a half years of age. These are found complementary to other traditional features. Further, by means of support vector machines, an impressive number of above 80% accuracy for the diagnosis of autism spectrum disorder at this age are observed on the data. Interestingly, the authors find that these vocalizations are better suited than speech-like vocalizations to this end.

Chapter 4 addresses the problem of recognizing mispronunciations of children with selected neuromuscular disorders on the phoneme level. The considered dysarthric speech can be challenging to recognize both for humans and current automatic speech recognition, as it is overall affected including impact on spectral and prosodic characteristics, coarticulation and concerning pronunciation rules. The authors show that representing the arising variabilities can help improve automatic recognition.

Chapter 5 is directly related to this topic: it touches upon improving dysarthric speech recognition of humans and computers by "familiarization." In the case of machine listening, this leads to adaptive machine learning. The authors further note that such adaptation benefits from an initial model already incorporating information on a target group speaker's pattern.

Overall, one can only congratulate the two editors, of which one contributed himself to the chapters, having put together these significant and most recent contributions to the field by eminent authors as well as for their guidance of a thorough and richly detailed iterative review process. As target audience for this collection, speech scientists, clinicians and pathologists will find interest in the first place. One can easily see this book as a further milestone on the highway toward earlier diagnosis and richer feedback, potentially available to more of those concerned, ultimately leading to prediction and

prevention with most patients in the loop over diagnosis and treatment available only to a few.

Professor Björn W. Schuller
(Doctor of TUM, Germany,
IEEE Fellow, and IEEE Computer Science Golden Core Member)
Imperial College London, UK, and University of Augsburg, Germany

Contents

List of contributors

Stefany Bedoya
Nexalogy
Montreal, Quebec
Canada
stefanybed@gmail.com

Jessica Brian
Holland-Bloorview Kids Rehab
Toronto, Ontario
Canada
jbrian@hollandbloorview.ca

Nirit Brosh Katz
Afula Child Development Center
Clalit Health Services
Givat Ada, Israel
nbrosh@netvision.net.il

Anshu Chittora
Healthark Wellness Solutions LLP
Memnagar, Ahmedabad
Gujarat
India
anshu.chittora@gmail.com

Tiago H. Falk
Associate Professor
INRS-EMTDirector, MuSAE Lab800
Rue de la Gauchetière Ouest
suite 6900 (NW wing)
Montreal, QC, Canada
falk@emt.inrs.ca

Mark Hasegawa-Johnson
Professor
Beckman Institute and Department of Electrical and Computer Engineering
The University of Illinois at Urbana-Champaign
IL, USA
jhasegaw@illinois.edu

Heejin Kim
Research Assistant Professor
Dept. Linguistics
The University of Illinois at Urbana-Champaign
IL, USA
hkim17@illinois.edu

Douglas O'Shaughnessy
INRS-EMT
University of Quebec
Montreal
Canada
dougo@emt.inrs.ca

Hemant A. Patil
Speech Research Lab
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT)
Gandhinagar
Gujarat
India
hemant_patil@daiict.ac.in

Richard Rose
Research Scientist
Google
New York City
USA
rickrose@google.com

Hardik B. Sailor
Speech and Hearing Research Group
The University of Sheffield
UK
sailor.hardik2000@gmail.com

https://doi.org/10.1515/9781501513138-205

Shou-Chun Yin
Nuance Communications Canada Inc.
Montreal, Quebec
Canada
sss123ca@yahoo.com

Kirtana Sunil Phatnani
Speech Research Lab
Dhirubhai Ambani Institute of Information
and Communication Technology (DA-IICT)
Gandhinagar
Gujarat
India
kirtana_phatnani@daiict.ac.in

Editors' introduction

Acoustic Analysis of Pathologies from Infancy to Young Adulthood covers a rather broad pediatric population, starting with infant cry analysis to toddler squeals/shouts/verbalizations to the dysarthric speech of college students. Among some of the innovative approaches to signal processing examined in this book is the study of *unsupervised auditory filterbank learning* using convolutional restricted Boltzmann machine (ConvRBM). The experimental results show that the proposed features perform better than the standard handcrafted feature sets such as mel-frequency cepstral coefficients (MFCC), using various statistically meaningful performance measures. Given that the corpus of literature on infant cry analysis commenced in the late 1960s, these new findings add significantly to methodologies used to analyze an infant's cry for signs of pathology, as well as to the corpus of knowledge on signal processing in general.

The contributors to this volume use novel methods in analyzing acoustic sounds of infants and children. To wit, in studying toddlers with autism spectrum disorder (ASD), they employ specific research methods that incorporate a much broader set of features in analyzing toddler acoustic productions. In so doing, they show how their approach improves the accuracy of diagnosing toddlers with ASD.

In this example of the study of toddlers suffering from ASD, the authors demonstrate the novel use of wavelet-based and speech modulation spectral features for ASD diagnosis based not only on speech-like verbalizations but also on toddlers' cries, laughs, squeals, shouts and other nonverbal sounds. They show that the proposed features are complementary to existing ones and, on a cohort of forty-three 18-month old toddlers, they showed how a support vector machine classifier was capable of correctly discriminating the ASD group from the typically developing toddlers with accuracies above 80%, thus outperforming existing methods. In short, they show that with these new features, vocalizations such as cries, squeals, whines, shouts and so forth, prove to be more discriminative than babble and speech-like vocalizations.

This is of importance in practice because by broadening the acoustic cues of autism to incorporate nonverbal productions, clinicians will be better able to diagnose ASD earlier on in the toddler, even before they are able to talk. Equally important, the contributors show greater precision in analyzing toddler acoustic signals by not grouping them together into one category regardless of age. Instead, they base their acoustic analysis on toddlers of roughly the same age, thus removing the potential bias from natural age-related acoustic changes and variability. Given the rapid development of children of that age, a difference in 3 or 4 months can prove significant.

https://doi.org/10.1515/9781501513138-206

Rounding out the discussion of acoustic pathologies in infants and toddlers are the studies on dysarthric children and young adults, which show the same rigor and precision as the volume's studies on infants and toddlers. For example, contributors show how the use of an interactive system for verifying phoneme-level mispronunciations in children's speech (by labeling utterances at the phonemic level according to the accuracy of pronunciation) will enable speech therapists to significantly reduce the investment of time and effort in providing therapy to children with dysarthric speech. This is extremely important in regions and communities where travel and financial constraints do not allow children easy access to highly skilled speech therapists. The authors show that by using this system based on acoustic modeling for automated verification of pronunciation accuracy, much of the work of assessing a patient is eliminated. This allows the speech therapist to use their time more efficiently for assimilating the evaluations obtained from the human–computer interaction and to provide performance assessment and prescribe additional therapy. As such, they show how cutting-edge work in computer-aided speech therapy can bring the much-needed therapeutic resources to dysarthric children who otherwise would be left without treatment.

In studying young dysarthric adults, particularly college-age students, the contributors examine novel approaches to machine adaptation to dysarthric speech. In so doing, they present new methods that improve the ability of a listener to understand dysarthric speech. They consider both human and machine listeners in this process. As such, they look at how one can improve the ability of the human listener in understanding dysarthric speech by invoking *familiarization*, which is a listener training method where listeners receive brief, yet structured, exposure to dysarthria. Similarly, they look at how one can improve the machine listener's understanding of dysarthric speech by using adaptive machine learning methods. They show how the efficacy of such methods can be improved by starting with an initial model that already incorporates some information about the person's speech patterns. Both approaches are speaker-centric, using methods to optimize adaptation to the speaker rather than force the dysarthric speaker to try to normalize their speech by undergoing strenuous rehab that may prove ineffective.

The contributors to this anthology are drawn from prominent universities and research labs in the United States, Canada and India, as well as from the commercial sector, such as Google AI. They instill passion and interest in finding the best acoustic methods to diagnose and detect medical problems in the pediatric population as well as to provide early intervention in treating a wide range of pathologies stemming from illnesses, developmental disorders and injuries. By bringing such eminent researchers under one rubric, this volume

offers an insightful analysis of acoustic methods for diagnosing and treating pathologies at each chronological stage of the pediatric population. We endeavor to make this anthology an informative resource for signal processing experts and speech scientists. We, likewise, strive to make this volume a useful resource to pediatricians, psychologists and speech therapists so that it becomes a game changer in the delivery of medical care and therapeutic support, particularly in the under-resourced regions where access to diagnostic analysis and therapeutic intervention is noticeably scarce.

Anshu Chittora and Hemant A. Patil

1 Understanding infant cry analysis for pathology classification

Abstract: Infants are difficult to understand as they cannot communicate their requirements. This motivates us to decode their language in meaningful interpretations so that adults can understand the requirements of their children. In this chapter, the cry analysis techniques used so far are discussed and some experiments in this direction are reported. Spectrographic analysis is also shown for its good performance in some pathological cases of infant cries. Along with this, what makes the infant cry analysis task a difficult task is also elaborated. Some results on infant cry analysis and classification work are reported in the later sections of this chapter.

Keywords: Cry modes, spectrographic analysis, higher order spectral analysis (HOSA), bispectrum, higher order singular value decomposition (HOSVD)

1.1 Introduction

The word infant is derived from the Latin word "Infan" which means speechless. Since infants cannot communicate using a language used by adults, they use cry as their communication language. Any language has linguistic and paralinguistic content in it, but the infant crying has only paralinguistic content. Crying is generated from a set of complicated and sophisticated physiological activities that involves coordination among the brain, respiratory and motor control, and the vocal system. It is considered that the crying helps in development of infant's physiology by increasing the pulmonary (lung) capacity [1].

In order to understand the actual reasons of the crying of an infant and to measure the well-being of an infant, development of a tool is needed which can

Note: Anshu Chittora is now at Healthark Wellness Solutions LLP, Ahmedabad, Gujarat, India. The work was done while the author was at DA-IICT Gandhinagar, India, and it does not contain any Healthark Wellness Solutions LLP proprietary information.

Anshu Chittora, Healthark Wellness Solutions LLP, Ahmedabad
Hemant A. Patil, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar

https://doi.org/10.1515/9781501513138-001

support the parents to take necessary actions and ensure the healthy development of an infant. This requires analysis of the infant cries as it is the only tool used by the infants to convey their emotional and physical states. Infant cry analysis is not limited to pediatrics, it requires inputs of neurologists, engineering and linguistics also. Infant cry analysis can unfold the correlations among these fields.

A team of Scandinavian researchers started the research in infant cry analysis by applying signal processing methods such as spectrographic analysis and acoustic analysis of the cry signals. For the initial twenty years, spectrographic analysis remained the primary tool for cry analysis. Later on, researchers tried automatic and semiautomatic methods using computer-based algorithms. Using these methods, researchers tried to classify and analyse different cry types such as hunger, pain and pleasure cries and few attempts have also been made towards pathological cry analysis. Compared to other fields of the speech signal analysis, the domain of infant cry analysis is comparatively less explored and researched.

Apart from the signal processing point of view, cry has also been studied from the perception point of view. It has been noticed that the mother of the infant can recognize her infant's cry. However, recent research shows that recognition of an infant's cry by the caregiver is dependent on the time spent by the caregiver with the infant irrespective of the caregiver's gender [2]. How parents perceive the cry impacts their parenting, if a parent can identify the correct reason of the crying then they can sooth the infant immediately. The identification of the reason of crying of an infant is recognised by the parents through the changes in the cry acoustics. If parents misunderstand the baby's crying pattern, then they find it difficult to calm the baby and it leads to child abuse [3]. It is also observed that the way the parents respond to the infant cries also impacts the changes in the neuro behaviour mechanism [4]. Parents perception of the deviations in the crying pattern are reflected in their parenting and misunderstanding these calls may result in compromised infant care and parenting effectiveness [5]. In case, an infant is found with abnormal cry patterns and characteristics then it is recommended to refer the infant for neurological examination. In the early 3 months of the infant's life, crying is a signal of vigour which helps in establishing parent-child contact [6].

1.2 Methods available for infant cry analysis

Since the inception of the idea to analyze the infant cry signals to understand the physiology of crying, many researchers have used different methods. Brief information of these methods is as follows:

1.2.1 Auditory analysis

It is generally noticed that a caretaker or mother can identify the reason of crying of her infant just by listening to it in daily routine. Wasz Hockert et. al. have observed different patterns of infant vocalizations for different reasons of crying, such as hunger, pain, birth and pleasure [7]. In some pathological cases, the pain cries of the infants are different from the pain cries of the healthy infants [8]. Auditory analysis of the infant cries is dependent on the manual training and perception, the accuracy in judgement is always questioned.

1.2.2 Time-domain analysis

This method of cry analysis makes use of the conventional chart recorders and oscillographs to record the cry signal parameters. The durational features have proved useful in the classification of the severely sick infants from the normal ones, for example, an infant suffering from brain haemorrhage needs higher stimulation to generate the cry signal of the same intensity compared to the healthy infants. Similarly, latency period is also higher in the pathological infant cries, which varies from 2.6–5.2 seconds compared to 1.2–1.6 seconds in healthy normal infants. It must be noted that the latency time depends on the wakeful/sleeping state of the infants. The duration of the cries is also a parameter of interest because it also changes with the physical fitness of the infants. In sick infants, the duration of the cries is smaller than the normal infants of the same age [9]. Duration of the cries can be a good measure of the developmental changes in the infants. As infant grows, the cry duration also increases.

Though this method is easy and conveys a lot of information, yet, it suffers from the drawbacks of human error in reading, instrument inertia and paper and pen speed for taking records. Generally, the features used in time-domain analysis are duration of the cry, latency period, and second pause and subsequent pauses. Latency Period is defined as the duration between the pain stimuli applied to the infant and the onset of infant cry sound. The time between the onset of the infant cry and the end of the signal is defined as the duration of the cry and it consists of the total vocalization occurring during a single expiration or inspiration. The time interval between the end of the first cry signal and the following inspiration is called second pause.

1.2.3 Frequency-domain analysis

In this method, bandpass filters are used to find the strength of the signal in different frequency ranges and give the information about the relative magnitude and pitch, formants and frequency related parameters of the signal. These parameters are of limited value when used alone but can be combined with other features to extract more information about a particular cry type.

1.2.4 Spectrographic analysis

As the name suggests, this method uses the spectrograms of the cries to analyze the cries. The spectrogram of a signal is a pictorial representation of the distribution of energy in both time and frequency domain. Most of the work in the field of infant cry analysis is mainly based on the spectrographic analysis of the cries. Using this method, duration and frequency domain features are calculated from the spectrogram of a cry. Using these features, different cry types are analyzed and significant differences based on these features are identified in order to classify various cry types. Some of the durational features extracted from the spectrogram of a signal are duration of a cry, latency period, pause length between cry units. The frequency-domain features are statistical parameters derived from the pitch contour, glottal roll, melody type, biphonation etc. Spectrographic analysis has shown excellent results in classification of pathologies also. Some the characteristics found in the spectrograms of the pathological cries are as follows [9]:

Cri-du-chat: It has been found in the research that the cries of the infants who suffer from this disease show a flat pattern of the harmonic contour and the pitch values lies in the range of 600–1,000 Hz [9].

Down' s syndrome: The vocalization of the cries was long and the mean minimum and maximum pitch were observed around 270 Hz and 510 Hz.

Congenital hypothyroidism: Spectrographic analysis of the cries of the infants who were suffering from congenital hypothyroidism showed a maximum pitch of 470 Hz and a minimum pitch of 270 Hz.

Infants with cleft palate: In their spectrographic analysis, no significant differences were found between their cries and the normal infant cries. Biphonation was absent in their spectrograms. Mean maximum and minimum pitches were 710 Hz and 360 Hz, respectively.

Neonatal hyperbilirubinemia: In these cries, biphonation and furcation were observed in the spectrograms. Moreover, the maximum and minimum

pitches are higher than the normal cries which are 2120 Hz and 960 Hz respectively.
Hypoglycemia: Biphonation was observed in the spectrogram and the fundamental frequency was also higher (1,590 Hz) compared to the normal healthy infants.

The above results show that the cry characteristics changes with the health condition of an infant. In diseases affecting the central nervous system, the frequency domain features changes. Biphonation and glide are mostly seen in the pathological cries. Though this method of cry analysis gives better insights in the cry analysis and characteristics of pathological cry, however, it also suffers from the disadvantage that identification of cry modes is based on the experience of a person and subjective.

1.2.5 Recent trends in infant cry analysis

With the advancement of computers and development of the programming skillset, researchers are using algorithm to analyze the infant cry data which make the results more accurate and reduces the processing time. The major work in this domain is related to the cry identification, cry classification, pathological cry classification and developmental studies of the infants. The identification of an infant from his or her cry is reported in [10, 11] where use of MFCC is proposed for this work. The work done by Xie et. al. is a landmark in infant cry research, they defined the ten cry modes from the spectrogram and used them for the automatic classification of various diseases [12–14]. Other important publications in this area are related to using different machine learning methods for the cry types classification, normal versus asphyxia pathology classification and normal versus pathological cries classification [15–29].

1.3 Challenges in infant cry analysis

This section discusses the problems faced in infant cry analysis when the standard signal processing methods are applied on infant cry signals. the signal processing methods discussed here are short-time Fourier transform (STFT) analysis, linear prediction (LP) analysis, cepstral analysis and Teager energy operator (TEO) based analysis. For the analysis, the adult samples are taken from the TIMIT database and the infant cry samples are taken from the collected corpus

which is described in detail in [30]. All the samples (of adults and infants both) used in the analysis have been down-sampled to 12 kHz to maintain the similarity in the analysis.

1.3.1 Short-time fourier transform (STFT) analysis

In speech signal processing applications, short-time Fourier transform (STFT) is frequently used. The STFT of a frame $s[n]$ is given as [31]:

$$X(m,\omega) = \sum_{n=-\infty}^{\infty} s[n]w[n-m]e^{-j\omega n} = \langle s(n), w_{n,m}\, e^{j\omega n} \rangle, \tag{1.1}$$

where $s[n]$ is the signal, $w[n]$ is the window, ω is the frequency and $<, >$ is the *inner product* operator of $s[n]$ with time-frequency atoms $\{w_{n,m}e^{j\omega n}\}$, where $w_{n,m} = w[n-m]$. Thus, STFT shows the frequency and time in the same plot.

A comparison of STFT of the voiced speech segments of the vowel /*aa*/ for a male, female, child sound and infant cry signals (voiced) are shown in Figures 1.1–1.4. The samples were recorded at a sampling frequency of 12 kHz. The speech signals are segmented into frames of duration 50 ms (30 ms in Figure 1.4) with an overlapping rectangular window of 10 ms. From Figures 1.1 and 1.2, it can be observed that the male voice has clear harmonic structure compared to the female voice. However, in females and infants, the harmonics are not much clear and have negligible amplitudes after the fourth harmonic (as shown in Figure 1.3).

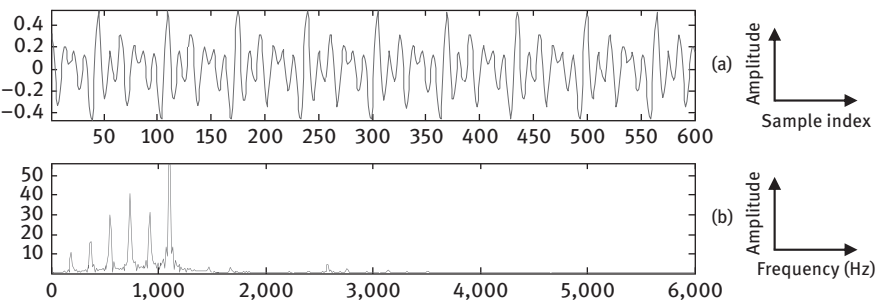

Figure 1.1: STFT representation of a speech signal (male voice): (a) time domain signal and (b) STFT of (a).

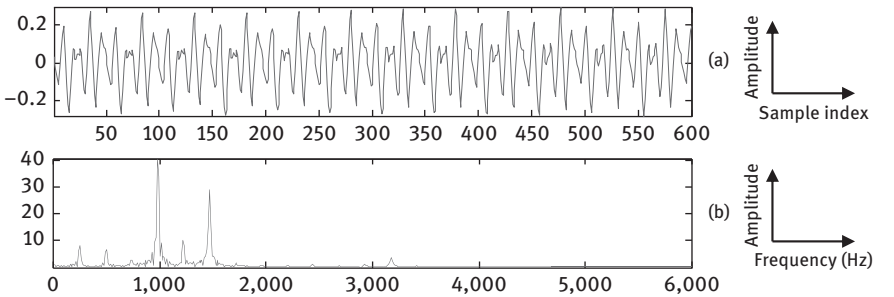

Figure 1.2: STFT representation of a speech signal (female voice): (a) time-domain signal and (b) STFT of (a).

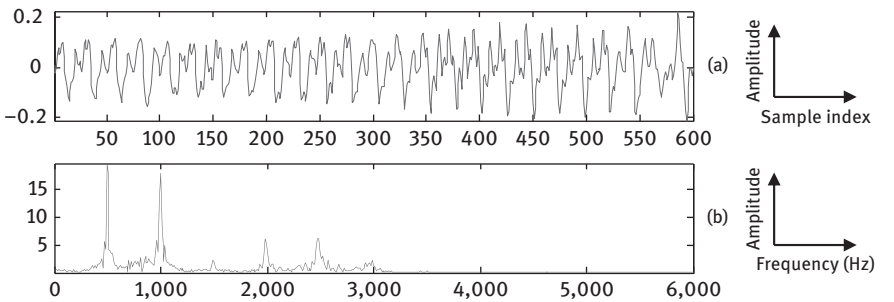

Figure 1.3: STFT representation of a speech signal (infant cry signal): (a) time-domain signal and (b) STFT of (a).

In the male and female speech, F_O ranges around 125 Hz and around 200 Hz spectral range, respectively. These values are around 250–400 Hz in children while in infants, it is around 500 Hz range and can raise up to 1 kHz in some cases (pre-mature infants). These variations in F_O are due to the different sizes and masses of the vocal source (vocal folds). The size of the larynx in men is about 40% taller and longer than women. The vocal fold length in male speaker is 60% longer than the female speaker, i.e. mass of vocal folds is significantly higher for male speakers than the female and hence, it takes longer time interval (i.e. more T_O and hence lesser F_O) to complete the glottal cycle for male speakers, making F_O lower than the female speakers. For the same reason, F_O in children and infants is comparatively much higher. F_O is related to the vocal fold length (L) by the equation [32]:

$$F_0 = \frac{1}{2L}\sqrt{\frac{\sigma}{\rho_,}} \tag{1.2}$$

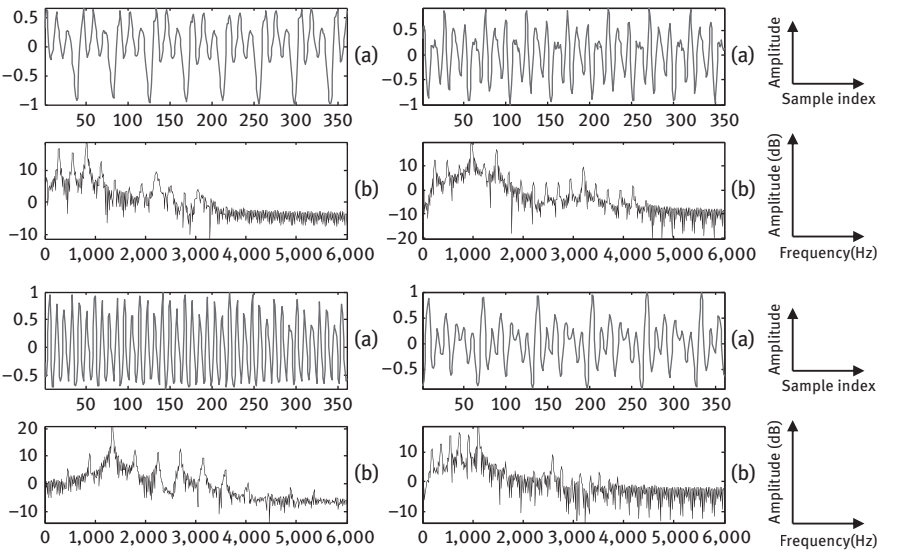

Figure 1.4: A comparison of STFT of male, female, children's speech and infant cry signal (Panel I- Panel IV) respectively. In all subfigures, (a) time-domain signal and (b) STFT of (a).

where σ is the longitudinal stress, and ρ is the tissue density in vocal folds.

The high values of the fundamental frequency in infants causes interference in the formant frequencies as both lies around similar values (1 kHz). Sometimes, it causes difficulty in the analysis of cries using STFT.

1.3.2 Linear prediction (LP) of the speech

The LP analysis of a male speech sample for the vowel /*aa*/ is shown in Figure 1.5. The length of the speech signal frame is 50 ms and the order of LP is taken as $p = 12$. The LP error signal is plotted in Figure 1.5(b) which is calculated by subtracting the original speech segment s(n) from its estimated approximation s^(n) as shown in eq. 1.3.

$$e(n) = s(n) - \hat{s}(n) \tag{1.3}$$

The log- magnitude spectrum of the LP residual and of the STFT spectrum is plotted in Figure 1.5(c). The STFT spectrum is shown by a light solid line while the LP spectrum is shown by dark solid line. It can be observed that when the LP order is

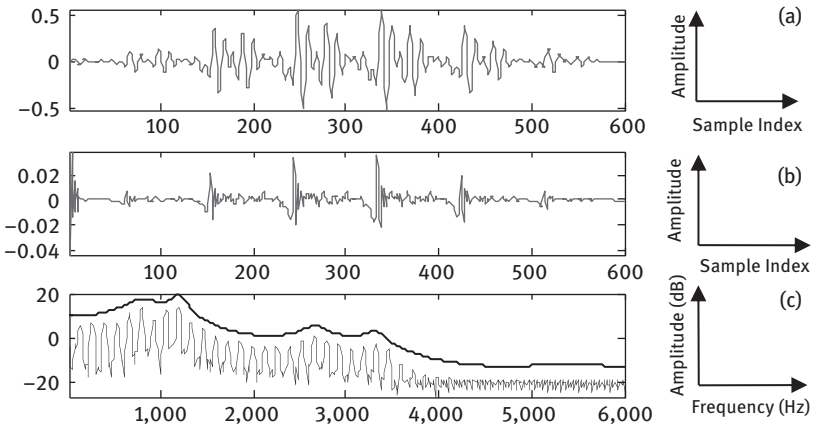

Figure 1.5: LP analysis of a male speech for vowel /aa/ (a) time-domain signal (b) LP residual and (c) LP spectrum for p = 12 and STFT, here, STFT is shown by light colour and LP spectrum by dark line.

low, the resonance peaks matches with each other whereas increasing the LP order will lead to matching with the STFT peaks. A similar analysis is also done for the male, female, children and infant voice samples in Figures 1.6–1.9.

The spectral peaks of the STFT spectrum for an infant cry signal matches with the source harmonics due to the sampling of the vocal tract. For comparatively higher values of LP order, i.e. p, the LP model tries to match all the peaks of the spectrum for infants than those of male, female and children subjects.

Increasing the order of LP analysis results in the matching of the LP spectrum and the STFT spectrum and similarly, the first formant frequency approaches to the fundamental frequency of the vocal folds vibration as evident from Figures 1.6–1.9. In case of infants, for small values of the p, harmonics of fundamental frequency (F_0) are detected instead of formants. Hence, it is difficult to find the formant frequencies of infants using LP analysis.

The reason behind the difficulty in estimation of the formant frequency in infants is the small size of the vocal tract and its narrow cross-section. The light weight of the vocal folds and sometimes under developed vocal folds and larynx adds to complexity in the estimation of fundamental frequency and also make F_0 higher, approaching to the first formant frequency. Thus, the problem of spectral resolution is much more complex in case of infants compared to female voices [33–35]. The number of formants covered in [0, $Fs/2$] is smaller (almost half) than male, female voices because the formant frequencies are higher

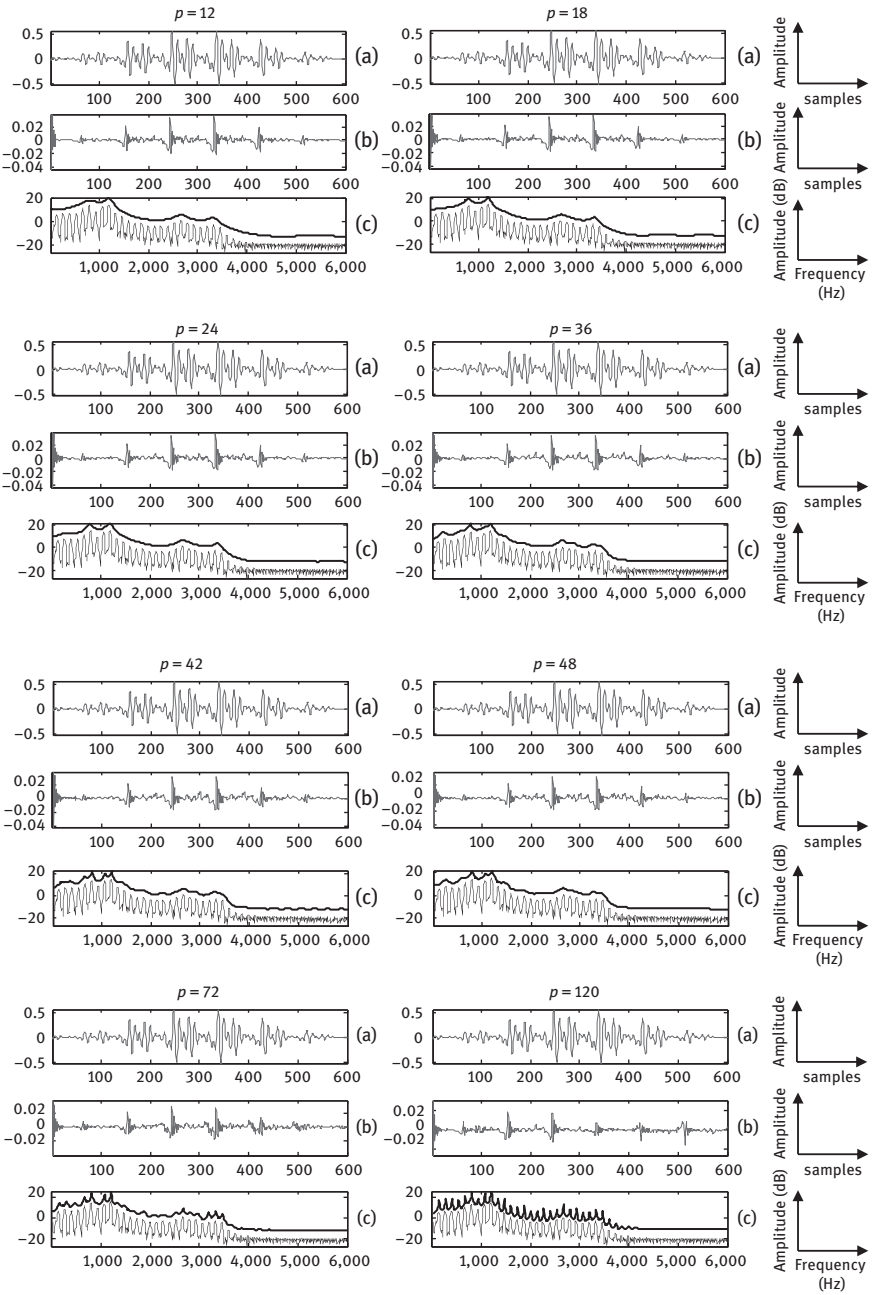

Figure 1.6: Changes in LP spectrum with LP order p in male speech. (a) time-domain signal (b) LP error signal and (c) LP and short-time Fourier spectra (in dB).

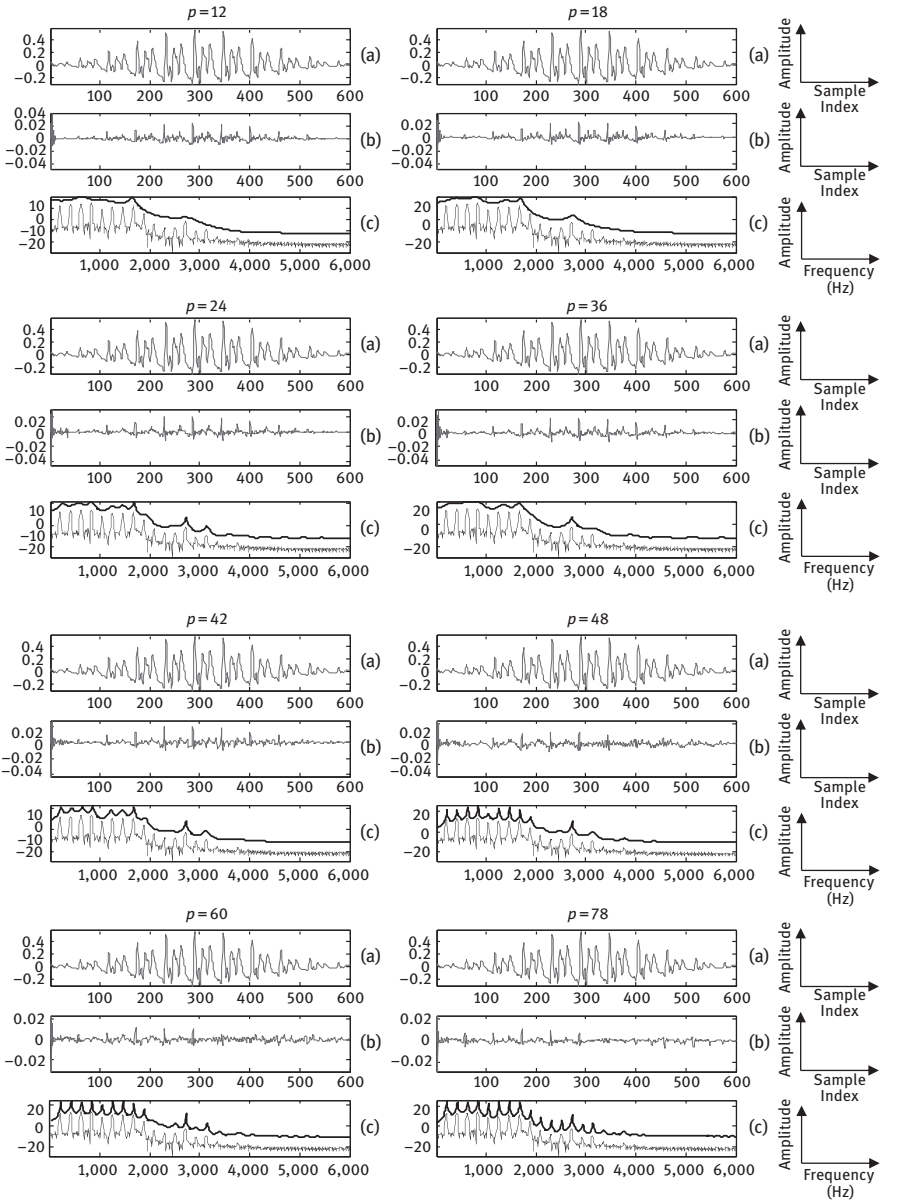

Figure 1.7: Changes in LP spectrum with LP order p in female speech. (a) time-domain signal (b) LP error signal and (c) LP and short-time Fourier spectra (in dB).

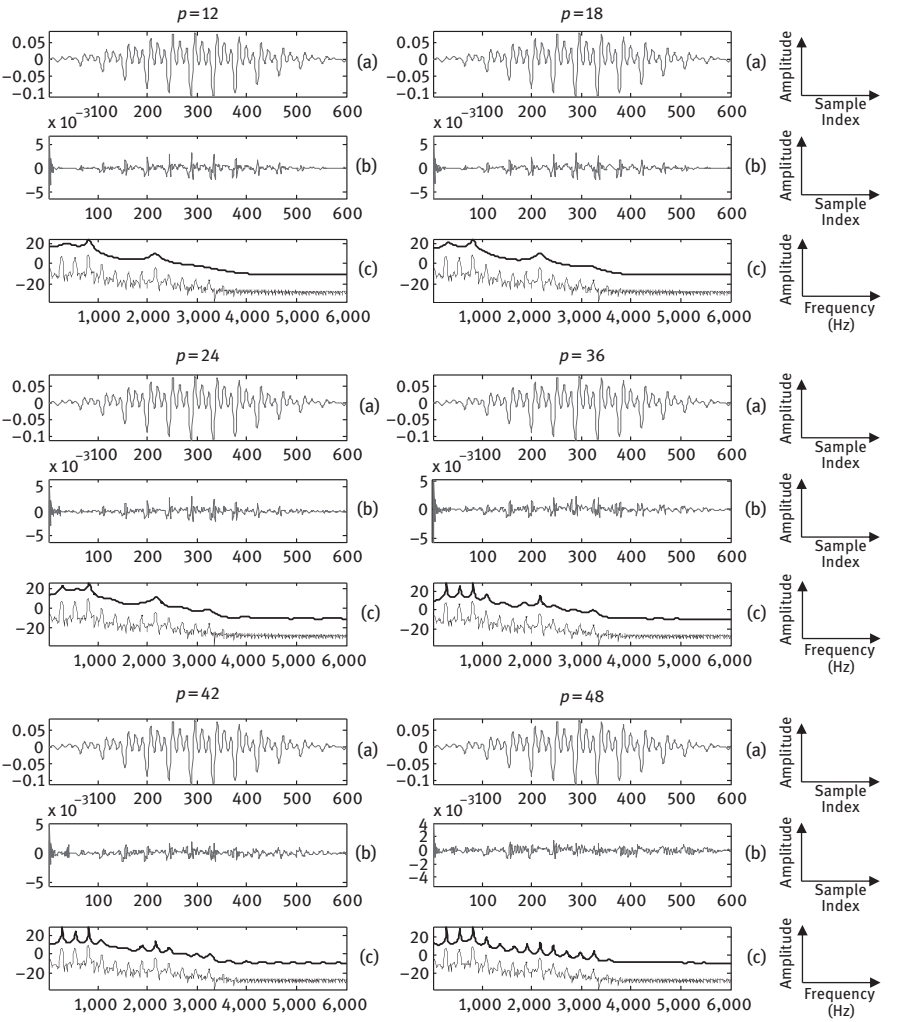

Figure 1.8: Changes in LP spectrum with LP order p in child's speech. (a) time-domain signal (b) LP error signal and (c) LP and short-time Fourier spectra (in dB).

in infants due to small length of the vocal tract. It indicates the need for using higher sampling frequency in case of infant cry analysis to capture the system-related information.

From Figure 1.10, it can be observed that the l^2 norm of residual exhibits a sharp decay initially (upto LP order *10–15*) indicating that the all-pole LP model

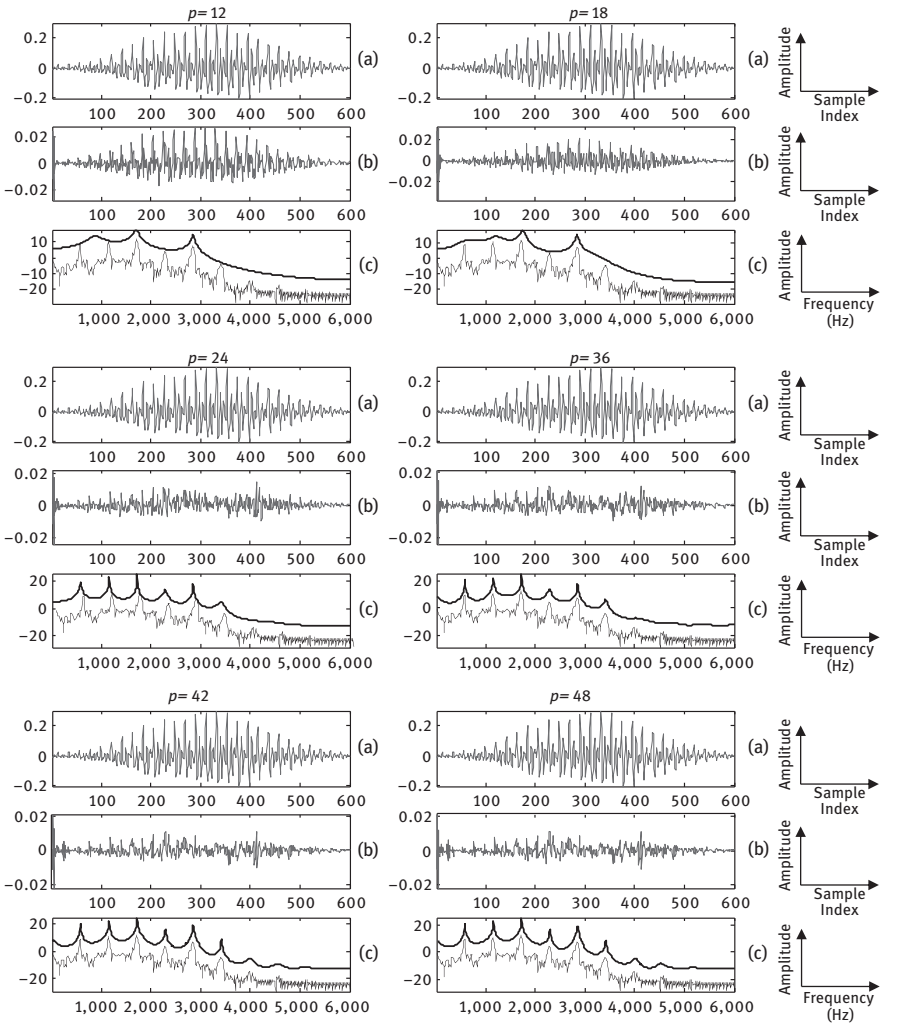

Figure 1.9: Changes in LP spectrum with LP order p in infant cry signal. (a) time-domain signal (b) LP error signal and (c) LP and short-time Fourier spectra (in dB).

first tries to match the dominant peaks in the spectrum (which corresponds to the formants). Afterwards, there is a little gradual decay in l^2 norm and then, it remains almost constant indicating no more optimization of LPCs is possible. This in turn means that speech samples are also related to each other with a dependency which is *nonlinear* in nature. The gradual decay in the l^2 norm

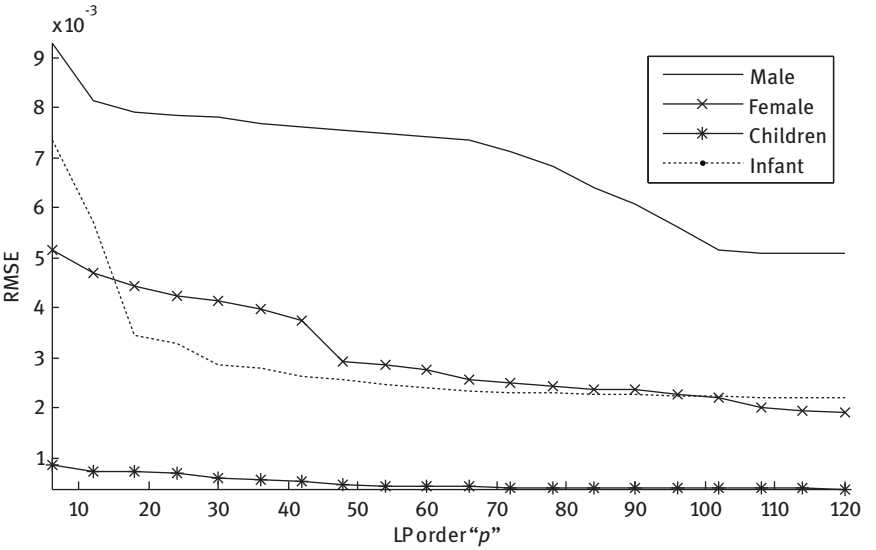

Figure 1.10: Trend in RMSE of LP error and LP order (p).

indicates that the LP spectrum tries to match the other dominant peaks (except formants) in STFT. This spectral matching happens at different LP orders for different speakers (male-to-infant). Due to sampling of vocal tract spectrum by the very distantly-spaced source harmonics, the spectral peaks in the STFT of infant cry are of almost similar height and hence, LP model tries to match all the peaks simultaneously for comparatively larger values of LP order than that of children, female, and male speakers.

It is clear from the Figure 1.11 that for the same sampling frequency (F_S), the number of formants covered in the range [0, $F_s/2$], is almost half in infants compared to the adults (because of the fact that vocal tract length is almost half in infants compared to the adult speakers due to eq. (1.2)). This draws an important observation that in case of adult speech analysis, sampling frequency (F_S) as low as *12 kHz* (or even *8 kHz*) is sufficient. However, for infant cry, to extract system-related information, the sampling frequency should be kept especially high in order to capture more formants.

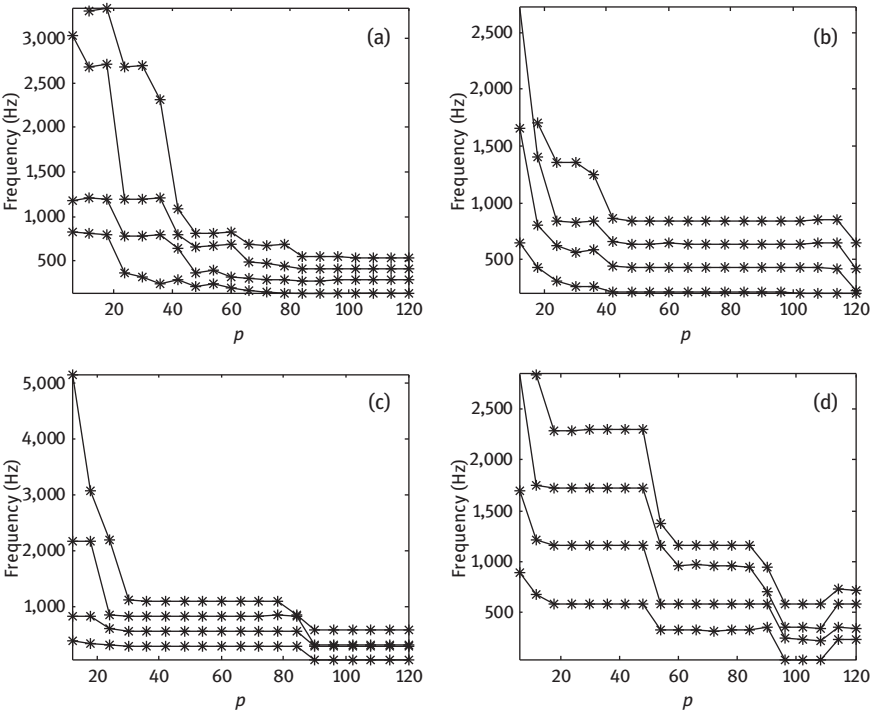

Figure 1.11: Transition of formants into F_0 harmonics with increasing order of linear predictor (LP). (a) male speech, (b) female speech, (c) children speech and (d) infant cry signal.

1.3.3 Cepstral analysis on infant cry signal

In Figure 1.12, the cepstrum analysis is shown for a short speech segment of 30 ms of a female subject's voice of the vowel /*aa*/. In all the samples, the sampling frequency is kept at 12 kHz. In Figure 1.12, the subfigure (a) shows the time domain signal and in subfigure (b) its cepstrum is plotted. A lifter shown by thick line in subfigure (b) is used to separate the initial portion of the signal. The liftered signal is then processed with Fourier transform to get the vocal tract response in frequency domain. The peaks of the vocal tract response (shown in Figure 1.12(c)) correspond to cepstrally smoothed vocal tract frequency response superimposed on STFT, i.e., formants. In the Figures 1.13–1.16, the subfigure (c) shows a thick line that corresponds to the cepstrally smoothed vocal tract response and the thin line corresponds to the STFT of short-time signal shown in each subfigure (a).

A similar analysis is done on male, female, child and infant voices and corresponding outcomes are shown in Figures 1.13–1.16. In all these analyses, a speech

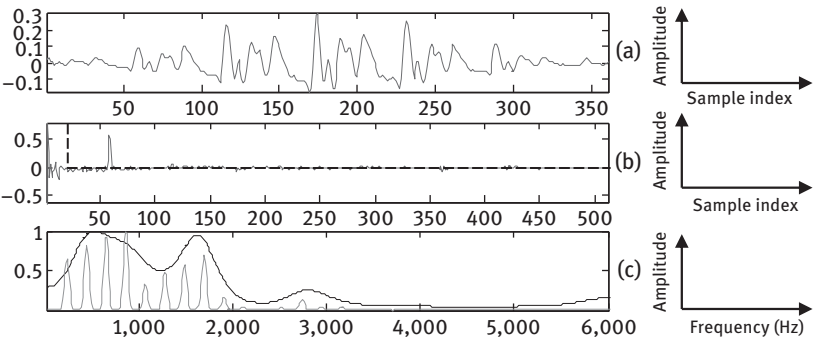

Figure 1.12: Cepstrum analysis of a speech segment of female speaker (a) time-domain signal, (b) real cepstrum of (a) and (c) estimated cepstrally smoothed vocal tract frequency response with lifter size 20 samples superimposed on STFT.

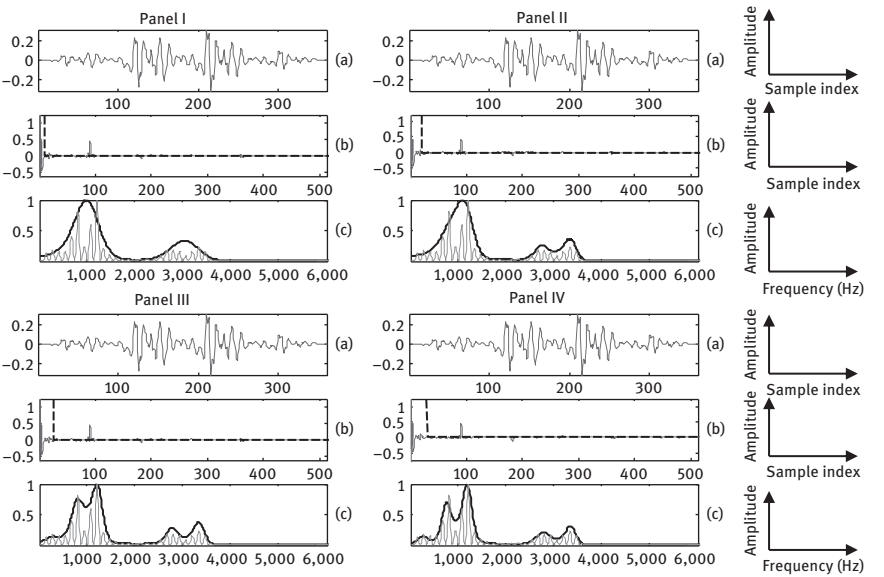

Figure 1.13: Cepstrum analysis for male speech for different lifter sizes. (a) speech signal, (b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response obtained by liftering (b) and superimposed on STFT. Panel I: lifter size = 10 samples, Panel II: lifter size = 20 samples, Panel III: lifter size = 25 samples and Panel IV: lifter size = 30 samples.

segment considered is of 30 ms duration. Comparison of Cepstral analysis with LP analysis for the same voices shown in the previous section indicates that Cepstral analysis performs much better in estimation of the formant frequencies in all

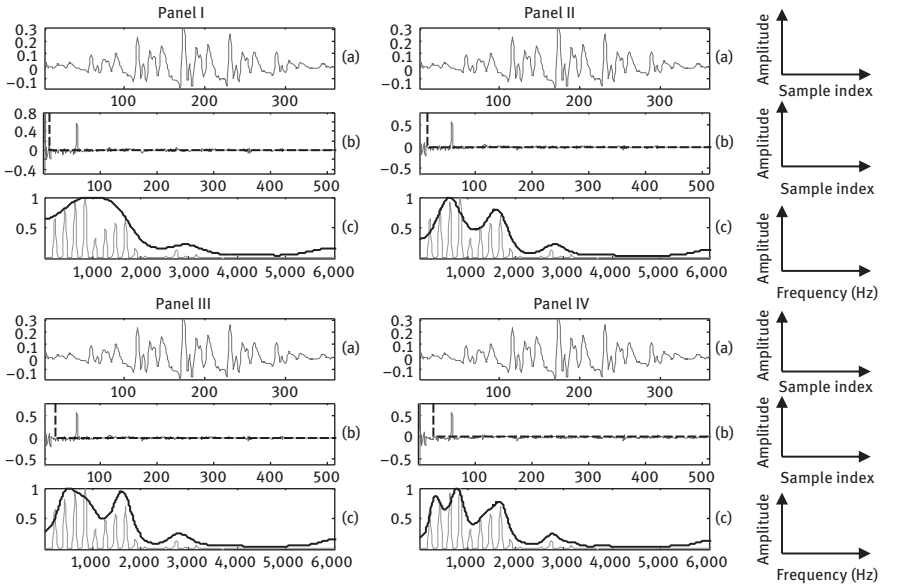

Figure 1.14: Cepstrum analysis for female speech for different lifter sizes. (a) speech signal, (b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response obtained by liftering (b) and superimposed on STFT. Panel I: lifter size = 10 samples, Panel II: lifter size = 20 samples, Panel III: lifter size = 25 samples and Panel IV: lifter size = 30 samples.

speakers; especially in case of high pitch speakers, it outperforms. To minimize the interference of the speech source in the system frequency response, the lifter size should be kept small enough to capture the system-related information, i.e., formants and ignoring the source information, F_O.

From the analysis, it can be observed that for the infants the lifter size of 10–15 samples (~1–1.25 ms) is sufficient to estimate the impulse response of the vocal tract. Increasing the lifter size to high values (comparable to adult speech lifters, 30 samples 2.5 ms) causes interference of the source harmonics with the vocal tract impulse response. The reason behind this interference is the high frequency of vocal fold vibrations (F_O, pitch period of 1–2 ms) in infants which causes high-frequency harmonics in the same frequency range where the vocal tract response lies. Hence, the separation of the source and system response becomes difficult in case of infants. With the same small lifter size, it is not possible to capture the system information in adults. Moreover, a fixed lifter size can work for different speakers of varying ages in adults while in infants because of higher variations in pitch among different infants and their ages, the lifter size may vary. Thus, algorithm development is needed for infant cry analysis for considering these variations in pitches.

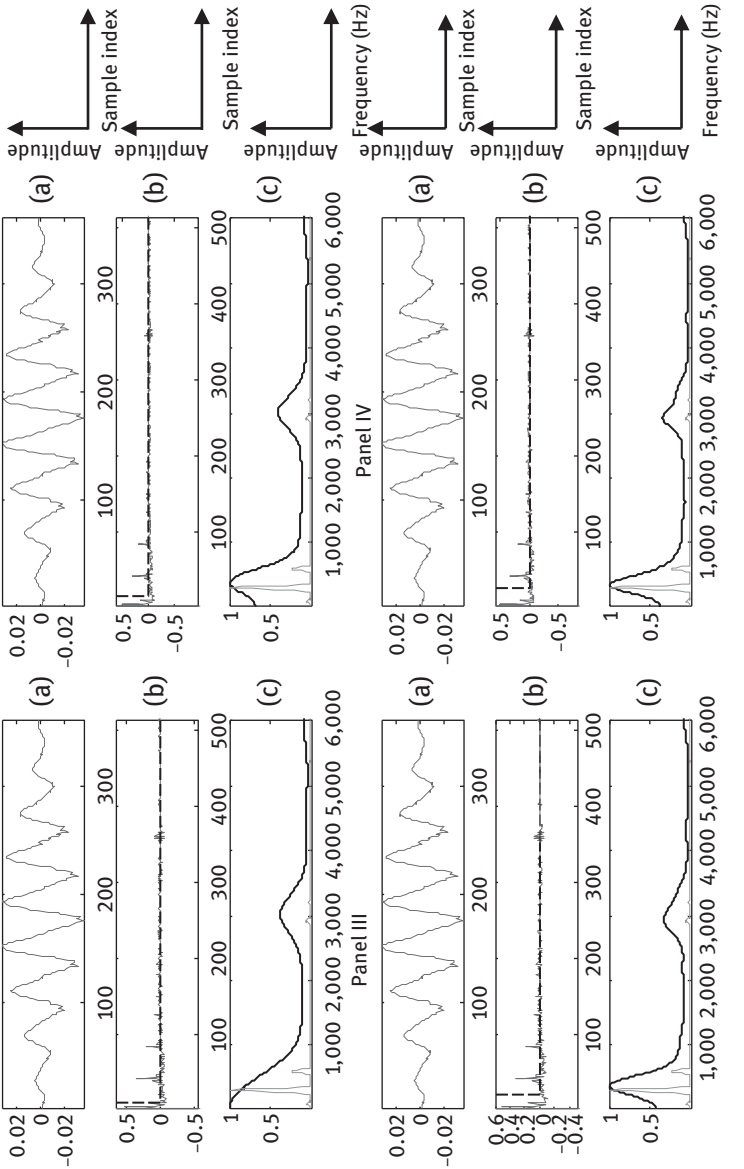

Figure 1.15: Cepstrum analysis for child's speech for different lifter sizes. (a) speech signal, (b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response obtained by liftering (b) and superimposed on STFT. Panel I: lifter size = 10 samples, Panel II: lifter size = 20 samples, Panel III: lifter size = 25 samples and Panel IV: lifter size = 30 samples.

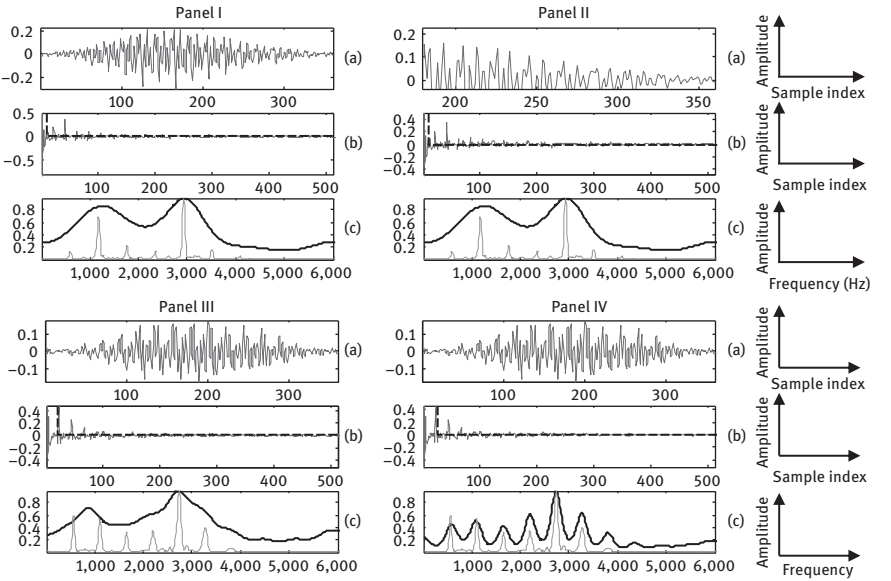

Figure 1.16: Cepstrum analysis for infant cry signal for different lifter sizes. (a) speech signal, (b) cepstrum of (a) and (c) cepstrally smoothed vocal tract frequency response obtained by liftering (b) and superimposed on STFT. Panel I: lifter size = 10 samples, Panel II: lifter size = 20 samples, Panel III: lifter size = 25 samples and Panel IV: lifter size = 30 samples.

1.3.4 TEO analysis of the infant cry signal

The speech production system is considered as an LTI (Linear Time-Invariant) system in conventional speech signal processing applications. However, the results reported by Teager indicates that speech production is a non-linear phenomenon. According to Teager, the excitation source of speech production contains the vortices which are distributed across the vocal tract and interact non-linearly with the pulsatile or aperiodic airflow. These interactions of the vortices and airflow are the actual source of speech production which is non-linear in nature. Thus, Teagers suggested a nonlinear model of speech production using energy of the airflow. This model suggests an energy tracking operator known as Teager Energy Operator (TEO) [36, 37]. For a short speech segment $x(n)$, TEO is given by

$$\psi\{x(n)\} = x^2(n) - x(n-1).x(n+1), \tag{1.4}$$

where $\psi\{.\}$ denotes the TEO operator. The dependence of the TEO on the past, present and future samples of the signals can be observed in eq. (1.4). Thus,

TEO has high resolution and gives a running estimate of the energy. The estimate of energy can have positive and negative values as well. In this Section, TEO analysis of the speech signals of the male, female, children and infant's cry is presented and compared with respect to others. All the samples in the analysis are down sampled to 12 kHz and then low pass filtered to 5 kHz. The speech signal is then segmented into the frames of 50 ms with 10 ms overlap. In all the following analyses, the LP residual $e(n)$ (i.e., eq. (1.3) and the TEO profile of the same signal are plotted to compare them.

The observations from Figures 1.17–1.21 are as follows:

1. It can be observed from Figure 1.17 that the TEO profile of a signal is not always positive though it correlates to the energy of a signal. The polarity of the TEO signal depends upon the following conditions
   a) If $x^2(n) > x(n-1).x(n+1)$ then TEO is positive, and
   b) If $x^2(n) < x(n-1).x(n+1)$ then TEO is negative
2. During the silence portion of the signal, the TEO is zero (as shown in Panel II Figure 1.17(b)) which indicates the ability of the TEO signal to identify the silence regions in the speech signal or infant cry signals. Thus, TEO can detect glottal activity from the no glottal activity area.
3. Glottal closure instances (GCI) are the instances where the maximum change in energy occurs due to the sudden closure of the vocal folds at GCIs. This event is an effect of sudden decrease in pressure at epochs. At these instances, the peaks are observed in the LP residuals well as in the TEO profile of the signals. The TEO signal match with the peaks of LP residuals. This indicates that like LP analysis, TEO also has the capability to capture the excitation source information. Thus, TEO analysis can also be utilized for the pitch extraction of the speech signals (As shown in Figure 1.18).

It can be observed from the TEO energy profiles of male, female, child speech, and infant cry signals shown in Figure 1.18–1.21 that the TEO energy profile is not smooth, it has many bumps within two consecutive GCI locations. If the signal under consideration is a damped sinusoid, then its TEO profile will be linearly decaying [38]. Presence of bumps in the TEO profile indicates deviation of the system from the linearity. This is an indicator of the non-linearity present in the speech production mechanism [31]. Hence, any sound produced by human beings is caused by the nonlinear interactions of the excitation source with vocal tract system (in particular with the first format F_1) [31].

From Figures 1.18–1.21, it is clear that the variations in Teager energy are different for different speakers (gender). Thus, the location and manner of nonlinearity are different in different subjects.

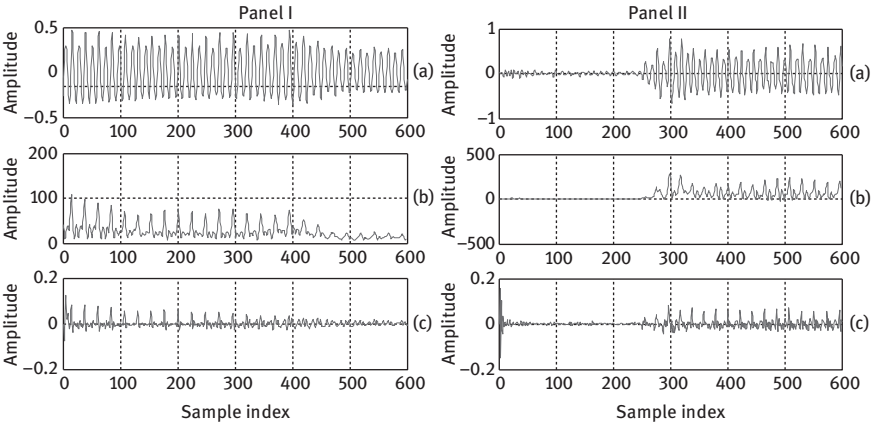

Figure 1.17: TEO analysis of infant cry signal for voiced signal (Panel I) and unvoiced signal (Panel II). (a) Time-domain signal of an infant's cry and (b) corresponding TEO profile and (c) LP residual of (a).

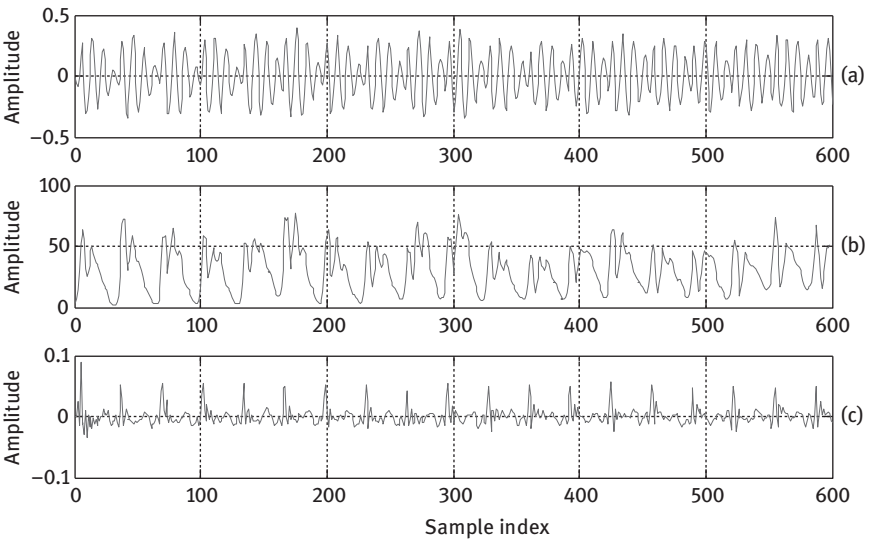

Figure 1.18: TEO analysis of voiced infant cry (normal) signal, (a) time-domain signal (b) TEO profile and (c) LP residual of (a).

From the analysis of TEO profiles of the infant, the presence of bumps indicates the presence of non-linearity in speech production similar to the adults. This non-linear behavior of the production mechanism can be explained by the presence of airflow vortices during speech production.

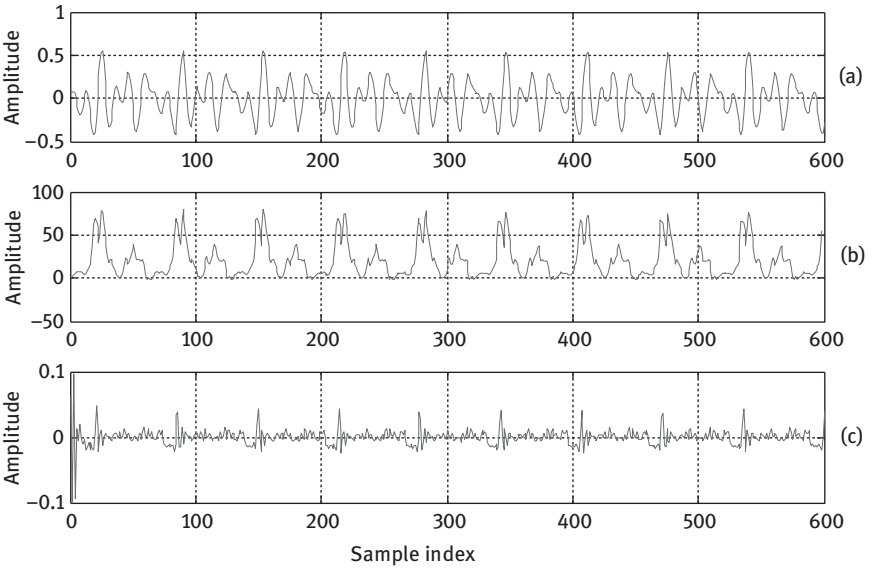

Figure 1.19: TEO analysis of voiced male speech signal, (a) time-domain signal (b) TEO profile and (c) LP residual of (a).

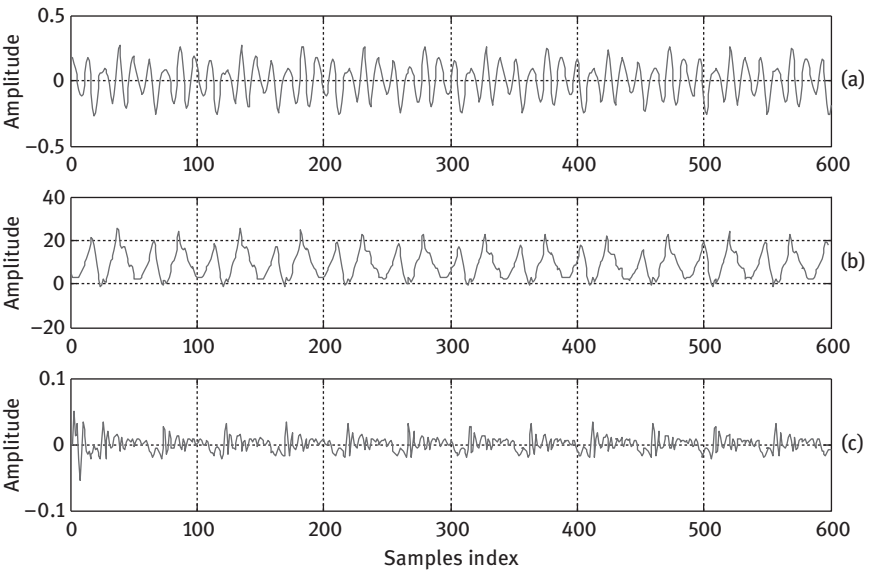

Figure 1.20: TEO analysis of voiced female speech signal, (a) time-domain signal (b) TEO profile and (c) LP residual of (a).

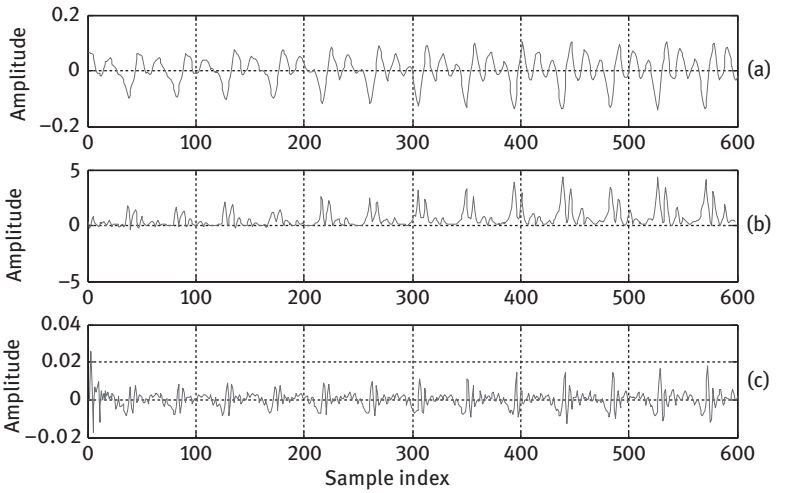

Figure 1.21: TEO analysis of voiced children's speech signal, (a) time-domain signal (b) TEO profile and (c) LP residual of (a).

1.4 Analysis of infant cries using spectrogram

If the signal under consideration is a stationary signal, we can work with Fourier analysis [12, 39, 40] but the cry signal is a non-stationary hence, Short-Time-Fourier Transform of the signal is used for the analysis which represents signal energy in the time-frequency domain. In STFT analysis, we divide the signal in smaller (comparatively) stationary segments using an analysis window, and then Fourier analysis is performed on the smaller segment of the signal.

A spectrogram is the representation of variation of signal energy along time and frequencies. Spectrograms are generally used in the fields of radar, sonar, music, and speech processing. In the analysis of speech signals, spectrogram is used for the identification of voiced, unvoiced, and plosive sounds. Spectrograms are used to study the voice excitation source and vocal tract system.

To analyze the signal in frequency-domain Continuous Time Fourier transform (CTFT) is used. CTFT of a signal $s(t)$ is given by

$$S(\omega) = F\{s(t)\} = \int_{-\infty}^{\infty} s(t)e^{-j\omega t}dt, \tag{1.5}$$

$$= \int_{-\infty}^{+\infty} s(t)\cos(\omega t)dt - j\int_{-\infty}^{+\infty} s(t)\sin(\omega t)dt \tag{1.6}$$

The CTFT has infinite time-dimensional sine and cosine basis functions and therefore, it shows poor resolution in time. Hence, instead of working with an infinite-dimensional basis function, it is truncated to localize events in non-stationary signals or highly time-varying signals. This gives motivation to the introduction of short-time Fourier transform (STFT). The windowed Fourier transform (WFT) was introduced to measure the frequency variations of the sound in 1946 by Dennis Gabor [41]. Spectrogram of a signal represents squared magnitude of the STFT of a signal. On the X and Y- axis time and frequency are represented, respectively. For a signal $s(n)$, the spectrogram is given by:

$$S(n,\omega) = \left| \sum_{n=-\infty}^{\infty} s[n]w[n-m]e^{-j\omega n} \right|^2, \tag{1.7}$$

$$= |< s(n), w_{m,\omega}(n) >|^2, \tag{1.8}$$

where $s(n)$ is the sampled signal of $s(t)$, $w(n)$ is the analysis window, $w_{m,\omega}(n) = w[n-m]e^{j\omega n}$ and $< s(n), w_{m,\omega}(n) >$ indicates the inner product of signal $s(n)$ with time-frequency atoms, i.e., $w_{m,\omega}(n) = w[n-m]e^{j\omega n}$.

Depending on the size of the analysis window, spectrograms are of two types: (1) narrowband spectrogram and (2) wideband spectrogram. In the case of wideband spectrograms, the analysis window is taken as less than a pitch period (<3 ms) whereas in the case of narrowband spectrograms the window width is taken as 2–3 pitch periods. Narrowband spectrograms are used to define the vocalizations in birds, animal and human beings. The wideband spectrograms are used to analyze the excitation source harmonics (formants) and its variations over time.

Wideband spectrograms are generally used to analyze the signals and it is computed by estimating the spectrum of a short segment of the signal. The short time segment or the window of the signal enables to capture the rapid variations in the amplitude of the signal. During the voiced portion of the speech, the vocal folds flap together and cause a rapid increase in the amplitude which is reflected as vertical lines in the wideband spectrograms. In a narrowband spectrogram, a longer time window is used to capture the rapid increase in amplitude that occurs at the time of vocal fold closure. Narrowband spectrograms have good frequency resolution. However, wideband spectrograms have good temporal resolution. Wideband spectrograms are generally used in speech signal processing-related applications, such as word segmentation, phoneme segmentation, voicing, unvoicing, and plosive detection. In Figure 1.22, wideband spectrograms of a male, female, child speech, and infant cry of same duration are shown for comparison.

STFT obeys the Heisenberg's uncertainty principle. The principle says that for a particle, more precisely the momentum is known, less precisely position is known and vice-a-versa. The same principle can be applied in the signal

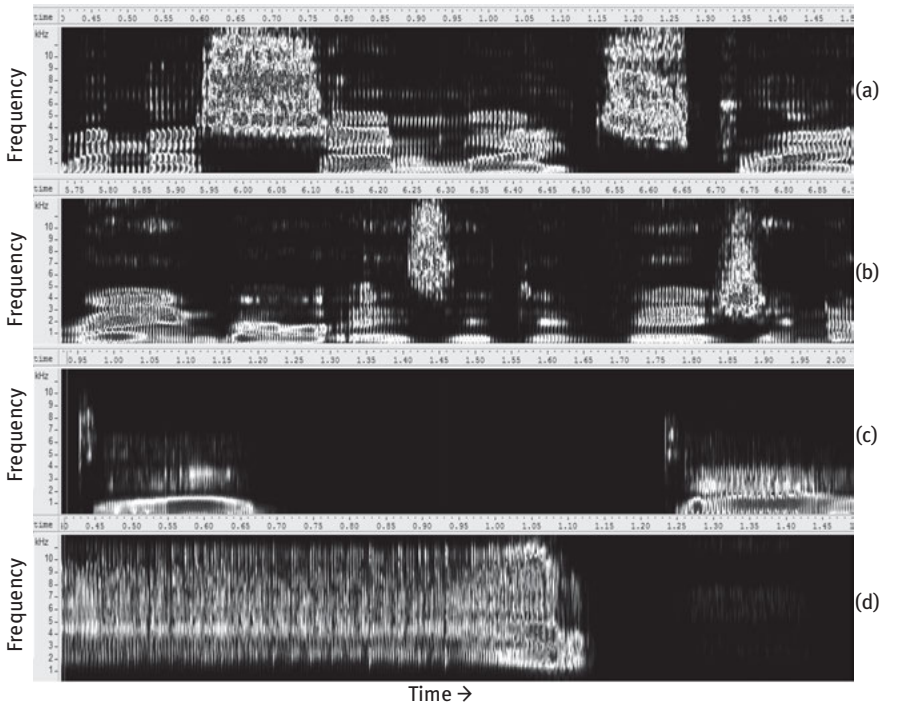

Figure 1.22: Wideband spectrogram of (a) normal male speech, (b) normal female speech, (c) normal child speech and (d) infant cry.

processing framework for time-frequency representation of the signal. As has been stated earlier that the length of the window is directly proportional to the frequency resolution of the spectrum, and it is inversely proportional to the temporal resolution of the signal in the time-frequency analysis. The uncertainty principle in the signal processing framework states that one cannot know what spectral components exist at what instance of time. However, one can know the time intervals in which a certain band of frequencies exists. This is known as time-frequency resolution [42]. Hence, if the signal length and window length are of same duration, we get good frequency resolution and temporal information is lost. Reducing the length of the window function improves the temporal information and reduces frequency resolution. Thus, there exists a trade-off between the window length and spectro-temporal resolution in STFT. For the spectrographic analysis, the selection of window size is also very critical. The uncertainty principle can be written as

$$\sigma_t^2.\sigma_\omega^2 \geq \frac{1}{4}, \tag{1.9}$$

where σ_t^2 and σ_ω^2 are the spread in time and frequency-domain with zero mean, respectively. In particular, $\sigma_t^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} t^2|f(t)|^2 dt$, $\sigma_\omega^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \omega^2|F(\omega)|^2 d\omega$ and $||f(t)||^2 = 1$.

In a wideband spectrogram, vertical striations correspond to the local energy fluctuations. The rate of vocal fold vibrations is called fundamental frequency (F_O) of speech sound. It is known that the F_O increases in the order defined as male, female, child, and infant. From Figure 1.22(a) as F_O is low in male voice, the vertical striations are clear in the spectrogram. However, as we move from male to child's voice, these are not at all clear and in infant's cry, these are very closely spaced, and do not impart any significant information. From the wideband spectrograms of male and female voices, vocal tract resonances are clearly visible. However, in case of child and infant cries, formants are *not* visible in the spectrogram. This is primarily due to the sampling of the vocal tract spectrum by the largely-spaced pitch (F_O) or excitation source harmonics. Thus, formant structure is there in the spectrum, however, we cannot see it in machines due to signal processing artifacts of sampling. This is the reason that wideband spectrograms are not so useful in infant cry analysis.

In Figure 1.23, narrowband and wideband spectrograms are shown for the same infant cry signal. In both the spectrograms, excitation source harmonics are dominating, making it difficult to identify vocal tract resonances in the wideband spectrogram which is primarily due to the serious interaction of pitch (F_O) source harmonics with vocal tract spectrum. However, the excitation source and vocal tract harmonics are mixed together in the wideband spectrogram. On the other hand, in narrowband spectrogram, the excitation source harmonics are clearly visible and hence, these can be used to define various infant cry modes (*e.g.*, study reported in [12]). Therefore, in remaining portion of this chapter, narrowband spectrograms are used for infant cry analysis.

1.4.1 Infant cry modes from narrowband spectrogram

From the spectrographic patterns of the infant cries, various cry modes have been identified by many researchers. The ten distinct cry modes used here for the analysis of infant cries are as follows [14, 43]:

Rising: In this mode, the F_O increases with time and is clearly observable.

Flat: If the pattern has constant and clearly visible F_O, it is termed as flat melody.

Falling: In this mode, the F_O contour has a falling trend with time and is clearly visible

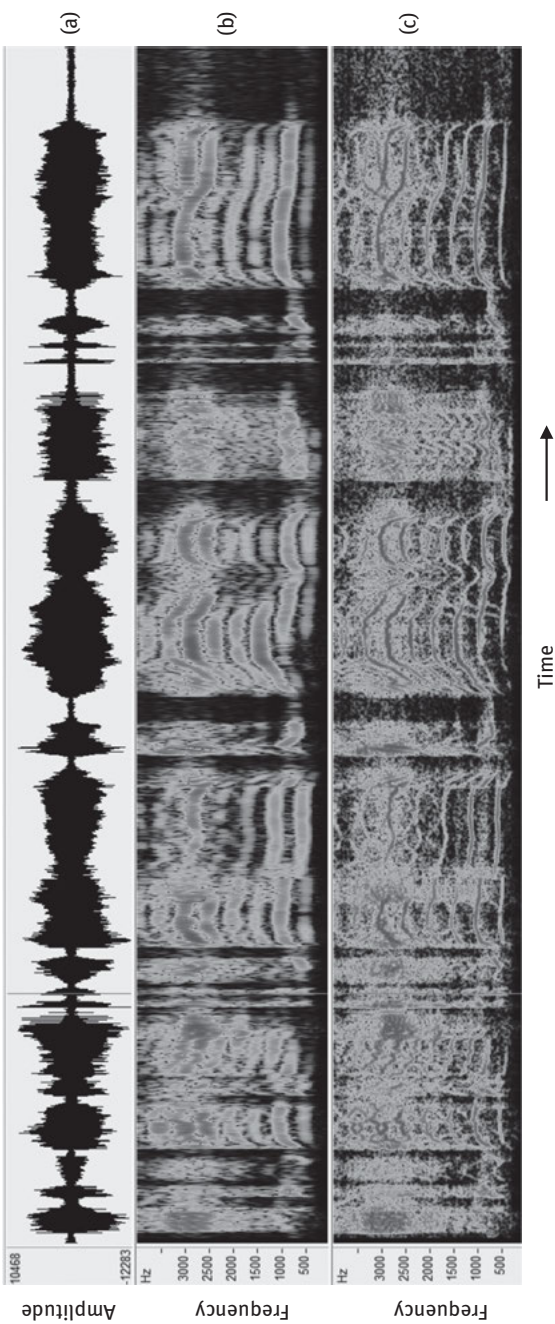

Figure 1.23: (a) Time-domain waveform of the cry signal (b) wideband spectrogram and (c) narrowband spectrogram.

Double harmonic breaks: When simultaneous parallel lines are observed between the harmonics of F_0, they are called double harmonic breaks. The lines in between the harmonics of the F_0, are harmonics of a frequency other than F_0. Generally, it presents the pathological condition of an infant, sometimes, they are also observed in newborn cries.
Glottal roll or glide: it is observed at the end of the cry and has a vibratory pattern of the harmonics of F_0. During this mode, the energy of the harmonics decreases slowly.
Weak vibration or vibrato: It is similar to the glottal roll except that this may occur in the middle of the infant cry also instead of at the end of the cry. The energy in weak vibrations is much smaller.
Hyperphonation: It is defined as the regions where F_0 exceeds 1 kHz. The presence of hyperphonation is related to the presence of pathology (i.e., neural disorders).
Inspiratory phonation: It occurs due to the sound made by the infant during inhalation. This occurs before the phonation region of the cry sound. Typically, of much smaller duration.
Dysphonation: This is the noise concentration found in the infant cries. This is characterized by the irregular or unstructured distribution of energy and typically the energy in this region is very high, and heavy *turbulence* is created in this region. If dysphonation dominates in the spectrogram, it may indicate the presence of pathology. However, newborn infant cries also have high dysphonation regions.
Vibrations: These are similar to weak vibrations, however, occurs with high energy.

All these cry modes are shown in Figure 1.24. In the next Section, spectrographic studies are reported carried out for infants suffering with laryngomalacia, deafness, and normal cries.

1.4.2 Spectrographic analysis and observations

Normal infant cry

The cry used in the analysis is collected during the vaccination of a 3-month old infant. The cry signal and its narrowband spectrogram are shown in Figures 1.25 and 1.26. In the spectrogram, flat, rising, falling and glide melody patterns are observed. Pitch harmonics are also seen along with hyperphonation and double harmonic breaks.

Figure 1.24: Infant cry modes derived from narrowband spectrogram (a) flat (b) rising (c) falling (d) double harmonic break (e) glottal roll (f) vibrations (g) hyperphonation (h) inspiratory phonation (i) dysphonation and (j) weak vibration.

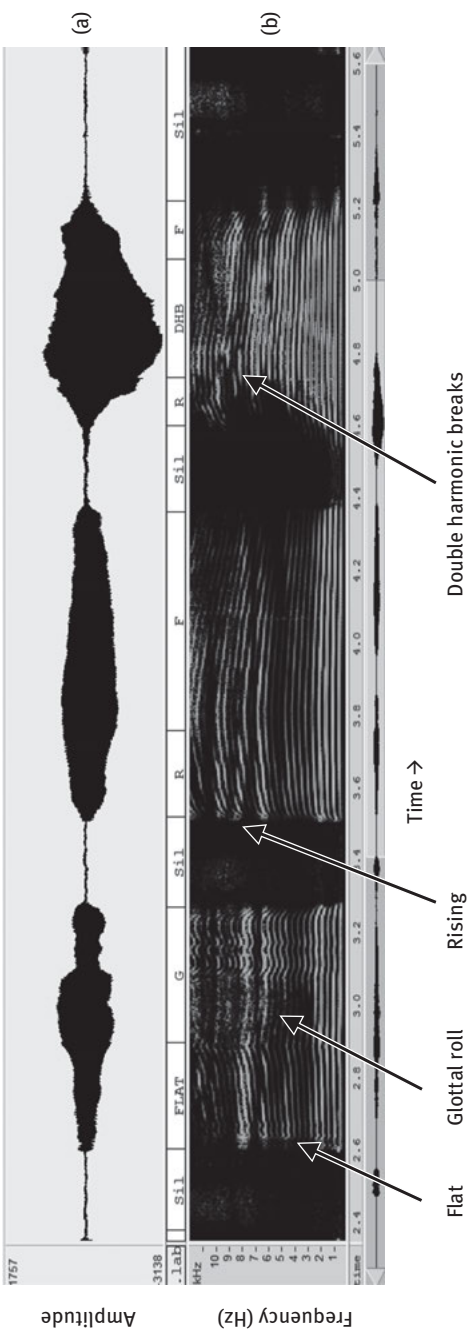

Figure 1.25: (a) Time-domain waveform, (b) corresponding spectrogram and cry modes present in the spectrogram of a normal infant's cry.

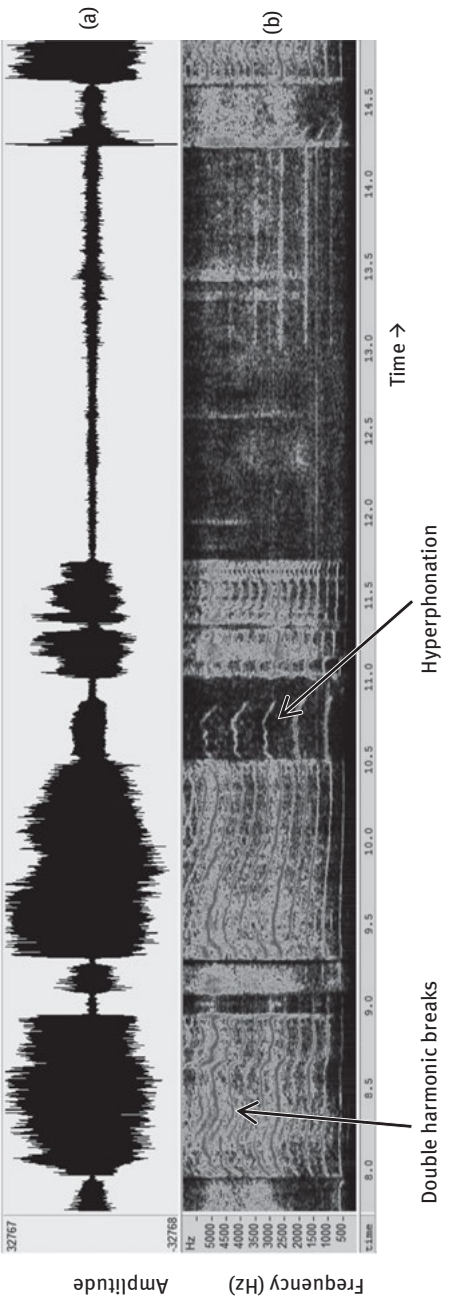

Figure 1.26: (a) Time-domain waveform, (b) corresponding spectrogram and cry modes present in the spectrogram of a normal infant's cry.

Neonatal cry

This is another case of normal infant crying. The cry is recorded while the neonate was given an injection (pain cry). The age of the neonate was 5 days. The spectrogram of the neonatal cry is shown in Figure 1.27 and it can be seen that the duration of the cryunits is very small and lacks in energy due to poor physical strength and poor regulation of the rib cage movement by his brain. The spectrogram has double harmonic breaks and hyperphonation is also present. Frequency inhalation is visible in the spectrogram.

Cry in infants with larynx not developed (Laryngomalacia)

It is a condition in which larynx are not fully developed in the infant and is a kind of laryngeal abnormality [44]. Infants suffering from this abnormality produce a noisy breathing sound which is seen as unvoiced region in the spectrogram.

In the spectrogram of the cry signal of an infant suffering from laryngomalacia (Figure 1.28), dysphonation, hyperphonation and inhalation patterns are seen. Spectral resolution is poor causing unstructured energy distribution in these cries due to turbulence. However, double harmonic breaks, glottal roll and glide are not seen in the spectrogram. Similar observations were made in [45].

Deaf infant's cry

The causes of hearing loss in infants are infection, genetics and mother with diabetes condition during pregnancy. Newborns with hearing loss do not respond to the sound and sound levels [46].
The cries of deaf infants are very short and cryunits are followed by long silences as shown in Figure 1.29. In the spectrogram of these cries, dysphonation is seen. Melody pattern is generally rising with a sharp fall and a presence of weak vibrations. Inspiratory phonations are absent in these cries while source harmonics are seen only in small sections of the spectrogram.

1.4.3 Summary of observations from spectrographic analysis

A summary of the presence or absence of various cry modes is given in Table 1.1.

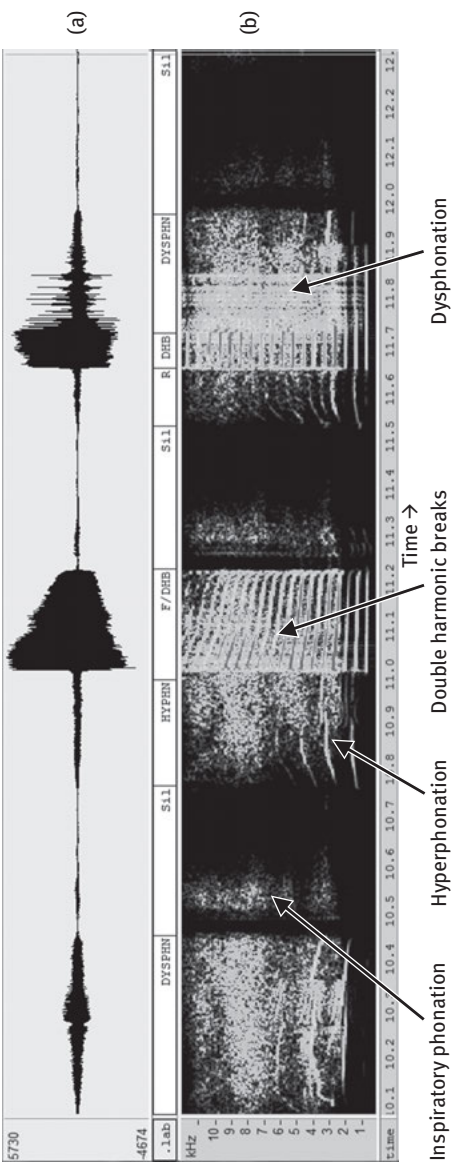

Figure 1.27: (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of a neonate's cry.

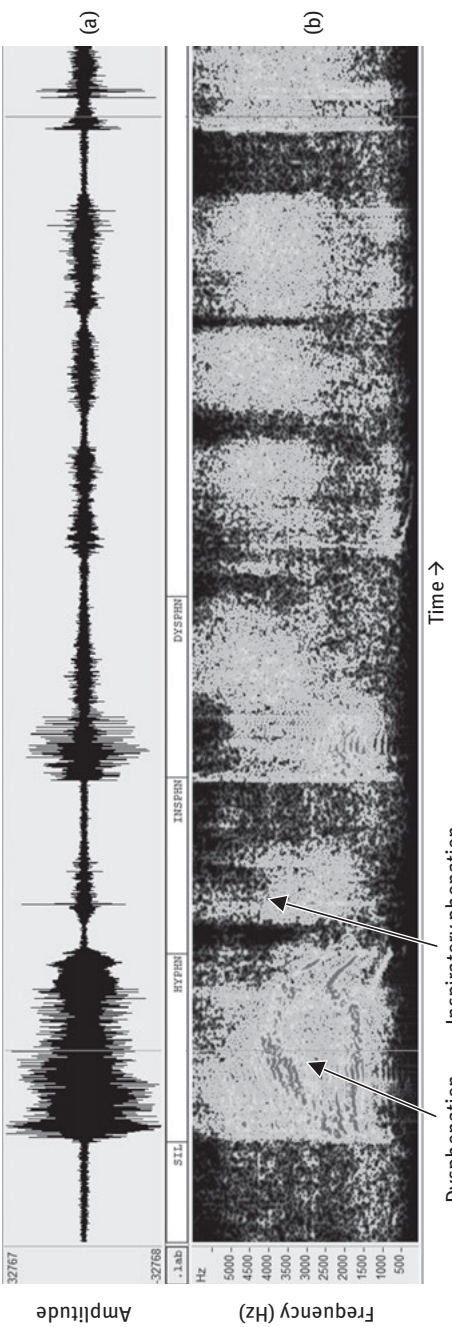

Figure 1.28: (a) Time-domain infant cry waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of an infant suffering from laryngomalacia.

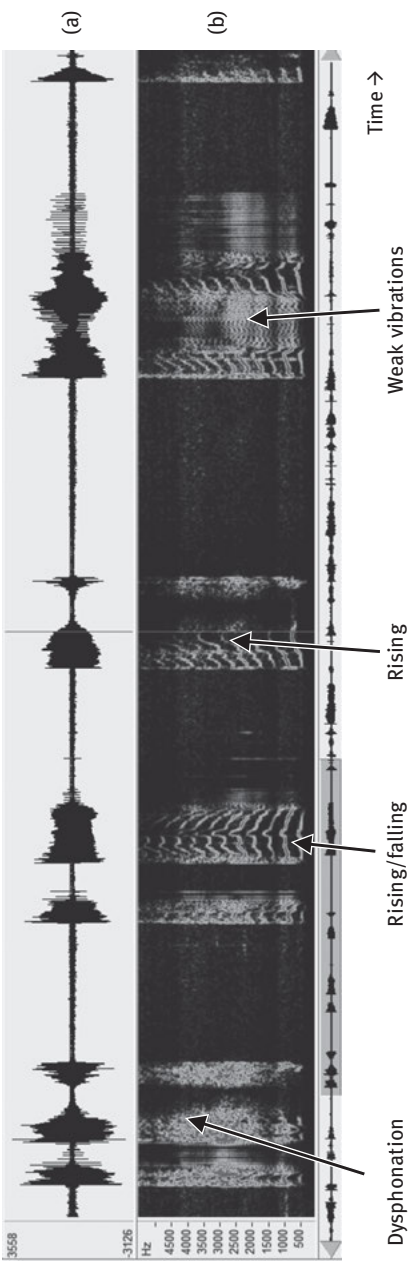

Figure 1.29: (a) Time-domain waveform, (b) corresponding narrowband spectrogram and cry modes present in the narrowband spectrogram of a deaf infant's cry.

Table 1.1: Presence of different infant cry modes in the spectrogram of a cry.

Infant Cry Modes Pathology	Flat	Rising	Falling	Double harmonic breaks	Glottal roll	Weak vibration	Hyper phonation	Inspiratory phonation	Dysphonation	vibrations
Normal	P	P	P	P	P	N	N	N	N	P
Neonatal	P	P	P	P	N	N	N	P	P	P
Laryngomalacia	N	P	P	N	N	N	P	P	P	N
Asthma	P	P	P	P	N	N	N	P	N	N
Heart disease	N	P	P	N	P	N	N	N	P	N
Down syndrome	P	N	N	N	N	N	N	N	N	N
Malnutrition	P	P	P	N	N	N	N	P	N	N
HIE	P	P	P	P	N	N	N	P	N	N
Hydrocephalus	P	P	P	P	P	N	N	N	N	N
Meningitis	P	P	P	N	P	N	N	P	P	P
RDS	P	P	N	P	N	N	P	P	N	N
Deaf	N	P	P	N	N	P	N	N	P	N
Asphyxia	N	P	P	N	N	P	N	N	N	N
Brain hemorrhage	P	N	N	P	N	N	N	P	P	N

P = Present, N = Absent

From the spectrographic analysis of infant cries of normal and pathological infants, it is observed that dysphonation is generally associated with the pathological condition of an infant. Inspiratory phonation and dysphonation are also seen in the pathological cries, however, hyperphonation is also observed in the newborn cries. Weak vibrations are also present in the pathological cries. However, any mode cannot be associated with a particular disease and the presence of any of these modes can not assure the pathological condition of an infant. Thus, a robust tool is needed to identify the pathological state of an infant.

The spectrographic analysis can be used as a primary tool for the screening of the pathological condition of an infant which advices the caretaker to go for a detailed examination of the infant. The shortcomings of the spectrographic analysis are listed below:

1. Poor dynamic range and spectral resolution of the spectrogram,
2. Prior experience is required in spectrogram reading (in addition, it is subjective and depends upon cognitive factors), and
3. Analyzing a large dataset with spectrograms is a tedious and time-consuming work.

Spectrograms employ a fixed duration window function $w(n)$, which limits the joint time-frequency atoms, i.e., $w_{m,\omega}(n) = w(n-m)e^{j\omega n}$.

1.5 Analysis of normal and pathological infant cries

1.5.1 Pre-processing and feature extraction

In this analysis, the data were collected at the sampling frequency of 44.1 kHz. From the infant cry samples, data is divided into several cryunits. In all, we have 229 normal cryunits and 145 pathological cryunits. The cry signal is passed through a fourth-order lowpass filter with a cutoff frequency of 3 kHz. Then, for each cryunit, voiced region is selected using energy measure (in particular, l^2 norm-based algorithm). From the voiced portion of the cryunit, F_O contour is extracted using a sliding window of 30 ms with overlap duration of 15 ms. For each of the cry sound frame, mean fundamental frequency (F_O) is estimated using autocorrelation method. After finding the F_O contour and the length of each cryunit (i.e., duration), a feature vector (1×4) is formed. The feature vector is given by V = [min $F_{O,}$ max $F_{O,}$ mean $F_{O,}$ duration]. For the estimation of F_O, autocorrelation method is used.

1.5.2 Autocorrelation method

The method used for fundamental frequency (F_O) estimation is the autocorrelation-based method [47]. In this method, autocorrelation of the short segment of the signal is found. The autocorrelation of the signal $s(n)$ is given by [47]

$$R(l) = \sum_{n=0}^{N-1-l} s(n)s(n+l), \tag{1.10}$$

where l is a lag element. The autocorrelation function is a non-invertible transformation of the signal, which represents the structure of the waveform. Hence, for the pitch period (P) detection of a voiced segment of speech, if the signal $s(n)$ is periodic with period P, i.e., $s(n) = s(n+P)$ then its autocorrelation function will also be periodic with the same pitch period P, i.e., $R(l) = R(l+P)$. Using this property of the autocorrelation function, peaks of the autocorrelation function of the cry signal are found. The difference in these peaks corresponds to the pitch of the signal. The fundamental frequency can be found from pitch period using $F_0 = F_s(samples\,per\ \text{sec})/pitch\,period(samples)$. This method is illustrated in Figure 1.30.

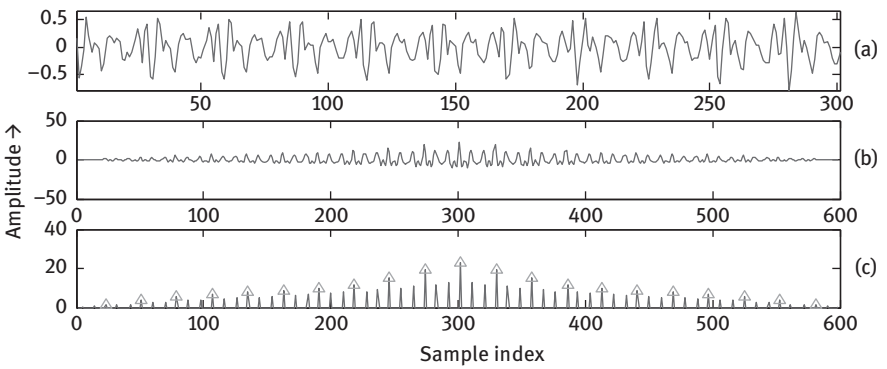

Figure 1.30: Estimation of pitch period using autocorrelation method. (a) time-domain infant cry signal, (b) its autocorrelation function and (c) peaks corresponding to pitch period. In all the subfigures, X-axis is sample index and Y-axis corresponds to the amplitude of the signal.

1.5.3 Experimental results

One-way ANOVA analysis [48] is conducted and the significance of proposed features is observed for the following cases:

a. Normal *vs.* pathological cry

b. Normal pain *vs*. pathological pain cry
c. Normal pain *vs*. normal hunger cry
d. Pathological pain *vs*. pathological hunger cry

From Table 1.2, it can be observed that for normal pain cries, maximum F_O is higher than the normal hunger cry. The mean F_O of pathological cries is significantly higher than the normal infant cries and the same trend is observed in the pain cry analysis of normal and pathological cries. Minimum F_O of pathological pain cry is higher than the hunger cry of similar type. Hunger cry is longer than pain cry in both normal as well as pathological infant cries. Minimum F_O of pathological cries is higher than normal cries. Minimum F_O of pathological cry is higher than normal pain cry. However, minimum F_O of hunger cry in both the cases are almost the same. Hunger cry of pathological infant has higher mean F_O than the normal infant's hunger cry.

Table 1.2: Statistics of features for all cry types.

Cry type ↓ / Features →	Minimum F_O (Hz)	Maximum F_O (Hz)	Mean F_O (Hz)	Cry length (s)
Normal pain (229)	206.89	765.14	415.07	1.48
Normal hunger (52)	216.53	737.02	416.38	2.14
Normal all (279)	206.82	779.24	405.44	1.60
Pathology all (145)	214.66	761.23	440.92	1.74
Pathological pain (52)	233.32	738.21	483.21	1.40
Pathological hunger (140)	216.87	763.54	437.27	1.87

* Numbers in brackets indicate the number of cryunits

Observations from Table 1.3, suggest that minimum, maximum and mean fundamental frequencies are good features for cry classification of normal *vs*. pathological cries. In both cases, minimum F_O and maximum F_O are good features to identify hunger and pain cries as well. Duration does not seem to be a good feature for identifying the normal and pathological cries. However, it is a good indicator of identifying pain and hunger cries in both normal and pathological infants. Mean F_O feature does not serve as a feature for distinguishing normal pain and normal hunger cry (as the number of samples and number of classes in the analysis are same, it resulted in same values of degree of freedom 'Df' in the ANOVA analysis).

Table 1.3: ANOVA analysis of different cry types.

		Df	Sd	F	P			Df	sd	F	P
Normal *vs.* pathological cry	**Min F_0**	452	14.34	32.07	2.64e-08	**Normal pain *vs.* normal hunger cry**	**Min F_0**	277	12.12	25.93	6.53e–07
	Max F_0	452	22.44	69.28	1.02e-15		**Max F_0**	277	37.40	23.18	2.42e–06
	Mean F_0	452	20.72	315.31	6.66e-54		**Mean F_0**	277	27.61	0.09	0.76
	Duration	452	1.246	1.45	0.22		**Duration**	277	1.23	11.97	0.0006
Normal pain *vs.* pathological pain cry	**Min F_0**	279	13.28	172.71	4.94e-31	**Pathological pain *vs.* pathological hunger**	**Min F_0**	190	24.08	17.08	5.35e-05
	Max F_0	279	36.82	23.69	1.89e-06		**Max F_0**	190	32.62	23.25	2.89e-06
	Mean F_0	279	31.69	197.35	2.87e-34		**Mean F_0**	190	34.10	68.39	2.28e-14
	Duration	279	1.17	0.18	0.66		**Duration**	190	1.17	5.90	0.016

Df: Degree of freedom, sd: standard deviation, F: *F*-ratio, P = *p*-value (probability)

Looking at the boxplots shown in Figure 1.31 of durational features for all the four cases, it can be observed that the means of the two classes within each case are not much different. However, the duration of hunger cries is found to be longer than the pain cries in both normal and pathological infant cries. Furthermore, normal pain cries are generally longer than pathological cries. When minimum F_0 plots are analyzed as shown in Figure 1.32, it is observed that this feature is not very much different in the two classes in all four cases. Only pathological pain cries have higher minimum fundamental frequency than the normal pain cries.

The mean fundamental frequency (F_0) of pathological cries is significantly higher than the normal infant cries and the same trend is observed in the pain cry analysis of normal and pathological cries as shown in Figure 1.33. There is no significant difference in the maximum F_0 of all four cases. Normally, pain cries have higher maximum F_0 (Figure 1.34). The significant difference in duration and F_0 of normal and pathological pain cries is the reason that these features can be used to classify normal *vs.* pathological cries.

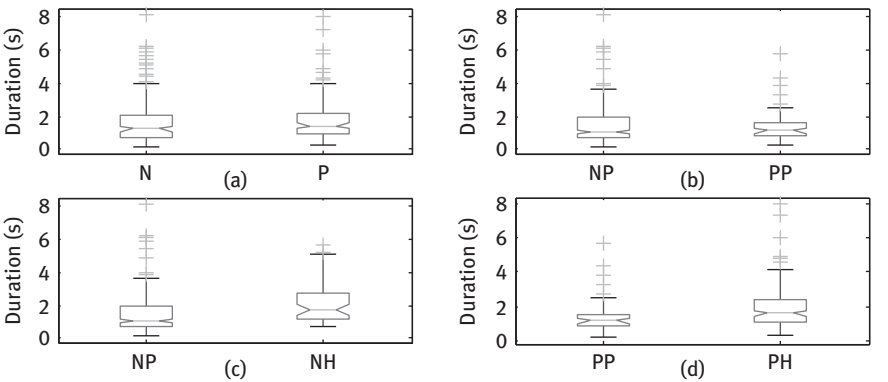

Figure 1.31: Box plots of duration features for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all subplots, Y-axis is time in sec. and X-axis represents the class type. Notations: N: Normal, P: Pathological, NP: Normal Pain, NH: Normal Hunger, PP: pathological pain, PH: Pathological hunger infant cry.

Features derived from F_0 contour for a cryunit play an important role in characterization of cry type and identify the state of the infant. For hunger cries, generally the duration is longer and for pain cries, F_0 is higher. These features are important acoustic cues for a parent or a caretaker of an infant for recognizing the need of infant (hunger) and for identifying the urgency of cry call (in case of pain cry). For infants who suffer from some pathology, it has been observed by their

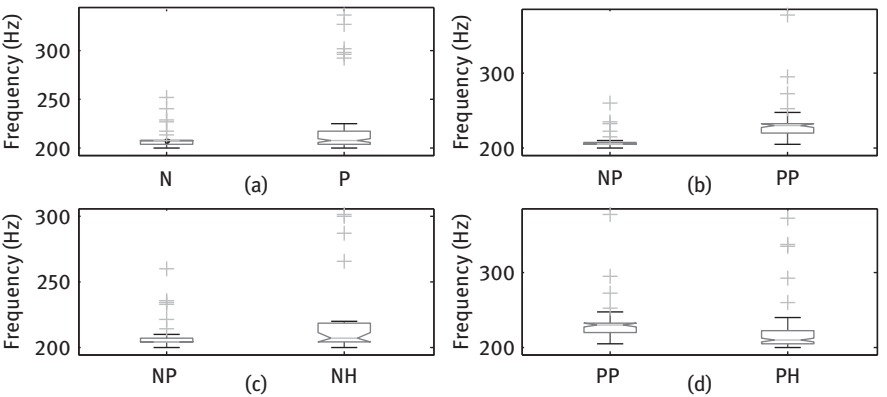

Figure 1.32: Box plots of minimum fundamental frequencies (F_0) features for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all F_0 plots, Y- axis is frequency in Hz and X-axis represents the class type.

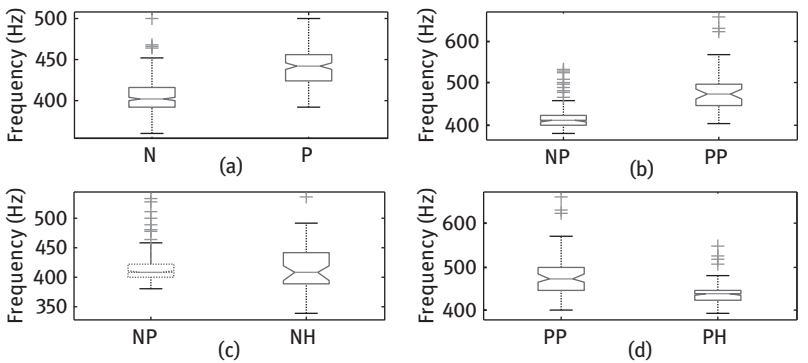

Figure 1.33: Box plots of mean fundamental frequencies (F_0) features for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all F_0 plots, Y- axis is frequency in Hz and X-axis represents the class type.

parents that their infant's cries are either shorter than normal infants or they have comparatively either higher or lower pitch than normal infants. The same is observed in our experiments as well. Change in F_0 of pathological cries can be attributed to instability in neural control of the larynx and lower vocal tract [49]. These parameters cannot be used alone for infant cry classification. However,

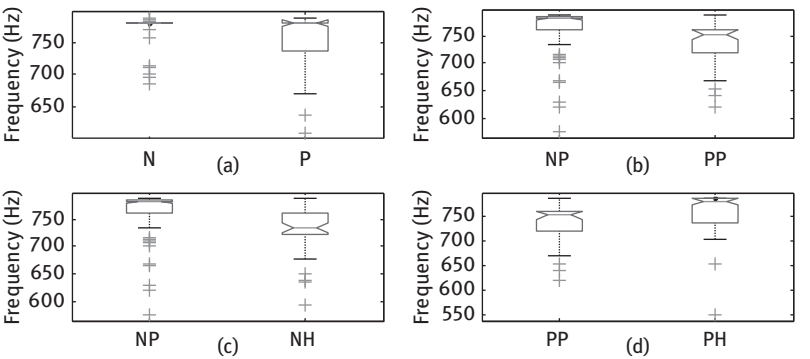

Figure 1.34: Box plots of maximum fundamental frequencies (F_0) features for (a) normal and pathological (b) normal pain and pathological pain cries, (c) normal pain and normal hunger cries and (d) pathological pain and pathological hunger cries. In all F_0 plots, Y- axis is frequency in Hz and X-axis represents the class type.

these features can be used along with some suitable features for improving the classification results.

1.6 Sudden infant death syndrome (SIDS)

Sudden infant death syndrome (SIDS) is the condition of an infant's death where the reason for death remains unanswered even after thorough medical examination and autopsy. However, sudden unexpected death in infancy (SUDI) or sudden unexpected infant deaths (SUID) refer to deaths in infancy, where the reason may be explained or unexplained. The distinction between SIDS and SUID is generally very difficult. Most of the SIDS deaths occur during the 1–3 months of age. The chance of deaths due to SIDS reduces after 1 year of age. It is also observed that most of the SIDS deaths occur in cold weather. Among other factors responsible for SIDS, are mothers who are of less than 20 years of age, prenatal exposure to cigarette, tobacco and nicotine. The prone or side sleep position increases the risk of rebreathing expired gases, resulting in hypercapnia and hypoxia. This position also increases the risk of overheating by decrease in heat loss and increasing body temperature compared to normal infants. The risk for SIDS is exceptionally high for infants who sleep on their stomach. This interesting observation led to a breakthrough investigation by studying the sleep postures of infants residing in East Germany *vs*. West Germany when Berlin wall was broken. Side sleeping is recommended only in exceptional cases for infants with upper airway disorder for

whom the airway protective mechanism is impaired, which may include infants with anatomic abnormalities, such as type 3 or type 4 laryngeal clefts, who have not undergone antireflux surgery [50]. Premature infants are more likely to be at risk for SIDS compared to normal infant groups. Pre-mature infants should be placed in spine position for sleeping as soon as possible after birth. Other recommendations to avoid SIDS are as follows:

1. The crib should be of safety approved.
2. Infant should not be allowed to sleep on sofa or soft beds.
3. Bed sharing with parents is not recommended in specific cases, such as where parents are using toxic drugs, alcohol, and cigarettes.
4. Car seats and other sitting devices are not recommended at home for routine sleep.
5. Wedges and positioning devices are not recommended.
6. Avoid alcohol and illicit drugs during pregnancy and after infant's birth.
7. Breastfeeding is recommended.
8. Swaddling does not reduce the chances of sleep. However, if it is applied to an infant who can roll, it can increase the risk of SIDS. There are insufficient evidences to show that swaddling should be used in routine to calm the infants. If it is correctly applied, this may avoid hazards, such as hip dysplasia (misalignment of hip joint) and strangulation.
9. Infant should be immunized as per the recommendation of hospital authorities.
10. Infants should sleep on their back instead of sleeping on their stomach.

In India, the infant mortality rate (infant deaths per 1,000 live births) is 38 which is much higher than other developed countries (3 in Australia and 6 in the USA in 2011–15) [51]. Following the above recommendations can reduce the risk of SIDS to an infant.

1.6.1 Cry characteristics of SIDS victims

It has been reported by the parents of the SIDS victims that the cries of their infants are strange or different than their siblings and other normal infants. Stark and Nathanson studied the cries of a male infant who died at the age of 6 months. They found that the cries are shorter and weaker compared to the normal infants [43]. Colton and Steinschneider reported the cry characteristics of a female infant who died at the age of 63 days and reported that the F_0 was lower, cry duration was longer and sound pressure level (SPL) was higher than the normal infant's group and SIDS sibling group [52].

The results reported by the studies are contradictory. Thus, it is difficult to say whether the identification of the SIDS prone infants on the basis of some parameters is possible. The study on SIDS is very difficult because enough statistically meaningful data for cry analysis is not available to have statistical confidence during analysis of results, and it is not known that the cries of SIDS infants are normal or abnormal *w.r.t.* characteristics embedded in their cries.

In a study reported by National Institute of Health (NIH), USA, it is found that the structural difference in a specific part of the brain (in particular, medulla oblongata which is known to control breathing functions) which causes risk for SIDS [43]. In a study reported by Harrison on SIDS infants, where he removed the larynges of the 74 infants who died of SIDS, it was found that the SIDS can be attributed to a decrease in the subglottic area (around the age of 3 months), which is highly dangerous. The reduction in subglottic airway is often secondary to an increasing mucus secreting glands, caused by upper respiratory tract infection [53]. Thereby, resulting in changes in the acoustic features of the cry.

1.7 Classification of normal and pathological infant cries

A significant amount of work has been done in the processing of adult speech for various applications like analysis of the disordered voice, development of a system for the disordered voices, classification of voice pathologies etc [54–59]. However, a marginal work is done towards the pathological infant cry classification and identification using cry as a biomarker. In this section, infant cry classification for the healthy and pathological infant cries is reported using higher order spectral analysis (HOSA). Bispectrum is used to classify the normal and the pathological cries from the HOSA family. Bispectrum of a signal is defined as the spectrum of the third order cumulant function.

1.7.1 Higher-order spectral analysis (HOSA)

Power spectral analysis is the most common spectral analysis method used by researchers across the world. However, the power spectrum of a signal ignores the phase information of the signal and provides the energy distribution across various frequency components of the spectrum. The power spectrum of a signal indicates the information contained in the autocorrelation function of the signal. The approach of using power spectrum for the spectral analysis of a signal works well

when the signal under consideration is a Gaussian signal because a Gaussian signal can be described by only first two moments i.e. mean and variance and its higher order cumulants are zero. In real life situations, a signal may not have Gaussian distribution and needs higher order moments to describe it completely.

In the following sections, higher order spectral analysis is used for the infant cry analysis, but before proceeding let us check its applicability to infant cry signals to an infant cry signal or alternatively proving that these signals are not Gaussian signals so higher order spectral analysis can be applied.

In Figure 1.35, the distribution of skewness and kurtosis of normal cries, pathological cries, adult voices and Gaussian signals of the same length is shown. For a Gaussian signal, skewness parameter is almost zero as observed from the figure. While for the infant cry signal, the skewness and kurtosis parameters are non-zero which indicates that the infant cry signals are not Gaussian. In an adult speech as well, the skewness parameter is not zero which also confirms that the speech is a non-Gaussian signal and hence, HOSA can be applied to it. The pathological voice samples are borrowed from the MEEI (Massachusetts Eye and Ear Infirmary) database [60]. In Figure 1.35, parameters are also plotted for the pathological voices of adults and it can be observed that the pathological voices have more deviation from zero of the skewness and kurtosis parameters compared to the normal voices. While in case of infants, the distinction between normal and pathological voices is difficult using these parameters as can be observed from Figure 1.36.

Higher order spectra are defined by the higher order statistics of the signal which are also called cumulants of the signal. The third order spectrum is called the bispectrum and the fourth order spectrum is termed as trispectrum of a signal. The well-known power spectrum is the second order spectrum of a signal. Higher order spectra are defined by the moments and cumulants. The

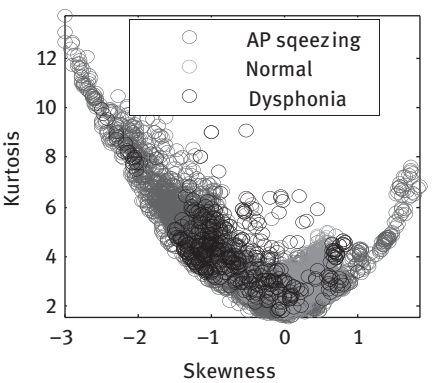

Figure 1.35: Skewness and kurtosis feature distribution for adult voices.

power spectrum is defined as the Fourier transform of the autocorrelation function where autocorrelation function is a second order moment of the signal. For a Gaussian signal, the higher order cumulants are zero which makes the higher

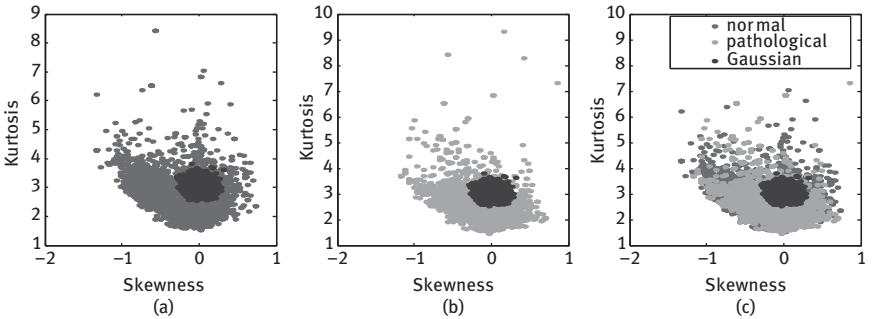

Figure 1.36: Distribution of skewness and kurtosis features for (a) normal and Gaussian signals, (b) pathological and Gaussian signals and (c) normal, pathological infant cries and Gaussian signals.

order spectrum of Gaussian signal zero. For the analysis of deterministic signals, moments and moment spectra are useful while for the stochastic signals cumulant and cumulant spectra are useful. Since speech is a stochastic signal, it can be better analyzed with higher order spectra (HOSA) of the signals. The HOSA is useful in the following cases:

1. To detect deviations from Gaussianity,
2. To identify and reconstruct non-minimum phase signals,
3. To suppress additive Gaussian noise and
4. To Detect and characterize nonlinear properties in signals and identify nonlinear systems [61].

The *nth* order moment function of a signal $X(k)$ is defined as

$$m_n^x(\tau_1, \tau_2, ..., \tau_{n-1}) = E\{X(k)X(k+\tau_1)...X(k+\tau_{n-1})\}, \tag{1.11}$$

where $\tau_1, \tau_2, ..., \tau_{n-1}$ are the time differences, and $E\{.\}$ denotes the statistical expectation. The *nth* order cumulant function of a non-Gaussian signal is given by:

$$c_n^x(\tau_1, \tau_2, ..., \tau_{n-1}) = m_n^x(\tau_1, \tau_2, ..., \tau_{n-1}) - m_n^G(\tau_1, \tau_2, ..., \tau_{n-1}), \tag{1.12}$$

where $m_n^x(\tau_1, \tau_2, ..., \tau_{n-1})$ is the *nth* order moment function of signal $X(k)$ and $m_n^G(\tau_1, \tau_2, ..., \tau_{n-1})$ is the *nth* order moment function of an equivalent Gaussian

signal that has the same mean and autocorrelation sequence as that of $X(k)$. Using the definition of cumulant, power spectrum, bispectrum and trispectrum can be defined as [61]:

$$Power\ \ Spectrum:\ P(\omega) = \sum_{\tau=-\infty}^{\infty} c_2^x(\tau)\exp(-j(\omega\tau)), \tag{1.13}$$

$$Bispectrum:\quad B(\omega_1,\omega_2) = \sum_{\tau_1=-\infty}^{\infty}\sum_{\tau_2=-\infty}^{\infty} c_3^x(\tau_1,\tau_2)\exp(-j(\omega_1\tau_1+\omega_2\tau_2)), \tag{1.14}$$

Trispectrum :

$$C(\omega_1,\omega_2,\omega_3) = \sum_{\tau_1=-\infty}^{\infty}\sum_{\tau_2=-\infty}^{\infty}\sum_{\tau_3=-\infty}^{\infty} c_4^x(\tau_1,\tau_2,\tau_3)\exp(-j(\omega_1\tau_1+\omega_2\tau_2+\omega_3\tau_3)), \tag{1.15}$$

$|\omega_1|<\pi,\ |\omega_2|<\pi,\ |\omega_3|<\pi$. For bispectrum, $|\omega_1+\omega_2|<\pi$ and for trispectrum, $|\omega_1+\omega_2+\omega_3|<\pi$. An excellent description of properties of higher-order spectrum analysis is given in [62].

1.7.2 Bispectrum estimation

There are two methods for the estimation of bispectrum of a signal, these are (1) Direct method and (2) Indirect method. Description of these methods is as follows [62]

Indirect method

For the estimation of bispectrum using indirect method, let the given dataset is $S(1)$, $S(2)$, ...,$S(k)$ and proceed as follows:

1. Segment the infant cry data of length N into K segments of M samples each, i.e., $N=KM$. Let these segments be denoted as $x(1)$, $x(2)$, . . ., $x(K)$.
2. Obtain 3rd order estimate of moment for each segment after subtraction of its mean value as given by

$$r^i(m,n) = \frac{1}{M}\sum_{l=s_1}^{s_2} x^i(l)x^i(l+n), \tag{1.16}$$

$i=1,2,...,K$, $s_1=\max(0,-m,-n)$ and $s_2=\min(M-1,M-1-m,\ M-1-n)$.

3. Average the moment function over all the K segments.

$$c_3^x(m,n) = \frac{1}{K}\sum_{i=1}^{K} r^i(m,n), \tag{1.17}$$

4. Generate the bispectrum estimate, i.e.,

$$B_3^x(\omega_1,\omega_2) = \sum_{m=-L}^{L}\sum_{n=-L}^{L} c_3^x(m,n)W(m,n)\exp(-j(\omega_1 m+\omega_2 n)), \tag{1.18}$$

where $L < M-1$ and $W(m,n)$ is a two-dimensional (i.e., 2-*D*) window function.

Direct method

1. Similar to the indirect method of bispectrum estimation, segment the infant cry signal of into K frames, each of length M. Add zeros at the end of each segment to make its length convenient for FFT computation, such that $M = 2^l, l \in Z^+$.

2. For each of the K segments, find the DFT, i.e.,

$$X^i(\lambda) = FFT(x^i(k)) \tag{1.19}$$

3. Estimate of bispectrum of each segment using the computed DFT by applying it to eq. (1.20–1.21)

$$b^i(\lambda_1,\lambda_2) = X^i(\lambda_1)X^i(\lambda_2)X^i(\lambda_1+\lambda_2) \tag{1.20}$$

The average of the bispectrum estimates is given as

$$B_3^x(\omega_1,\omega_2) = \frac{1}{K}\sum_{i=1}^{K} b^i(\omega_1,\omega_2) \tag{1.21}$$

Here, $\omega = (\frac{2\pi F_s}{N_0})\lambda$ and N_0 is the total number of samples in a segment.

In Figure 1.37, examples of the bispectrum estimated using direct and indirect methods are shown for normal and pathological infant cries. The differences in the two bispectrum can be observed from the spectral peak locations and the strength of the peaks. The strength of the bispectrum peaks is higher in normal infant cries and the bispectrum is smoother compared to the pathological infant cries. These differences motivate to use bispectrum for the classification of normal and pathological infant cries. In the following experiments, the performance of the direct and indirect method in the classification of normal and pathological cries is

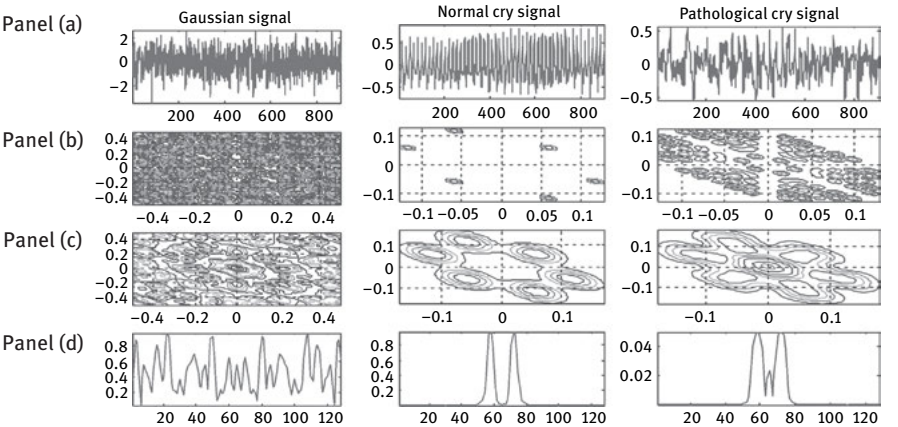

Figure 1.37: Estimated Bispectrum for a Gaussian signal, normal cry signal and pathological cry signal. In all the subfigures, Panel (a) time–domain waveforms, Panel (b) bispectrum using direct method, Panel (c) bispectrum using indirect method and Panel (d) diagonal slice derived from indirect method are shown for Gaussian signal, normal cry signal and pathological cry signal. In subfigures of Panel (a) X-axis is samples and Y-axis is amplitude, in subfigures of Panel (b) and Panel (c), X and Y-axis represents frequencies, ω_1 and ω_2, respectively and in subfigures Panel (d) X-axis is frequency and Y-axis is amplitude of bispectrum. After [63, 64].

measured and efficiency of different feature extraction methods is shown for the classification work.

1.7.3 Higher-order singular value decomposition (HOSVD)

In this section, HOSVD is used for the extraction of the features from the bispectrum of the infant cry signals. The HOSVD theorem is used to reduce the dimensions of the features space to reduce the complexity and memory requirements of the processors and reduces the classification time. HOSVD is proposed by Lathauwer et.al [65]. This feature extraction method has been used by the authors for phoneme classification and normal *vs*. pathological cry classification [66–69]. HOSVD is applied to a tensor and is a generalized form of the SVD. To form a tensor, the 2-*D* feature set, i.e., bispectrum ($F \in R^{I_1 \times I_2}$) is stacked together one after another to form a 3-*D* tensor A. Let the number of samples be I_s. The tensor A (of dimension $I_1 \times I_2 \times I_s$) can be represented in HOSVD form (as shown in Figure 1.38) by using eq. (1.22):

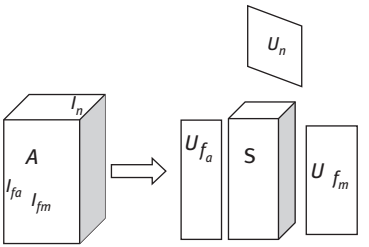

Figure 1.38: Singular value decomposition of tensor A . After [65].

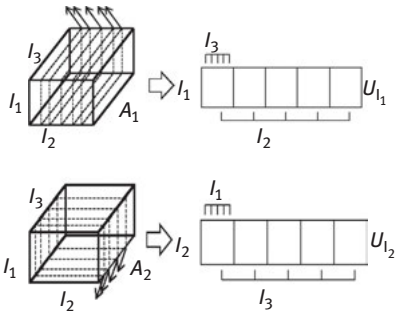

Figure 1.39: Unfolding of tensor A to matrix A_1 and matrix A_2. After [65].

$$A = S \times_1 U_{I_1} \times_2 U_{I_2} \times_3 U_S, \tag{1.22}$$

where S is the core tensor with the same dimension as A. $U_{I_1} \in R^{I_1 \times I_1}$, $U_{I_2} \in R^{I_2 \times I_2}$ and $U_{I_s} \in R^{I_s \times I_s}$ are the unitary matrices of the corresponding subspaces of I_1, I_2 and I_s. The matrices U_{I_1} and U_{I_2} contains n mode singular vectors, i.e.,

$$U^{(n)} = \left[U_1^{(n)} \quad U_2^{(n)} \quad .... \quad U_{I_n}^{(n)} \right]. \tag{1.23}$$

The matrices U_{I_1} and U_{I_2} can be obtained from the matrix unfolding of A. The unfolded matrices $A_1 \in R^{I_1 \times I_2 I_S}$ and $A_2 \in R^{I_2 \times I_1 I_S}$ are obtained (as shown in Figure 1.39), and they are decomposed in their SVD representations to give U_{I_1} and U_{I_2}. Only first R_1 and R_2 principal components are retained from these unitary matrices, respectively. Next, $\hat{U}_{I_1} \in R^{I_1 \times R_1}$ and $\hat{U}_{I_2} \in R^{I_2 \times R_2}$ are obtained, which gives reduced dimension of feature set, namely,

$$Z = B \times_1 \hat{U}_{I_1}^T \times_2 \hat{U}_{I_2}^T = \hat{U}_{I_1}^T . B . \hat{U}_{I_2}, \tag{1.24}$$

where $Z \in R^{R_1 \times R_2}$ and $B \in R^{I_1 \times I_2}$ which is taken from A. From the infant cry recording or episode (cry recorded from beginning till end), cry units are separated. The cry sound produced in one expiratory cycle is known as a cryunit. The cryunits of an infant are similar to the words spoken by an adult.

1.7.4 Experimental setup and feature extraction

To begin with, the infant cry database is divided into train and test sets in 75:25 ratio, respectively. Keeping 25% of the dataset aside for testing, such four train-test datasets were created. Using energy based algorithm (l^2 energy), voiced segments are extracted from the cryunits. The voiced data is segmented into non-overlapping frames of length 10 ms each and for each of the frame amplitude normalization is performed by mean subtraction. From each cryunit 20 frames are considered for the analysis and feature extraction.

Bispectrum of a signal is a two-dimensional feature, for the experimental purpose, we have considered FFT size of 128 which gave bispectrum of 512 × 512 dimension. The bispectrum is shown in Figure 1.40. It can be observed from the figure that the bispectrum has 12 symmetry regions. The symmetry property of bispectrum follows from the symmetry properties of the moments [62]. Because of the symmetry property of the bispectrum, information in the first or the third quadrant of the bispectrum is sufficient to consider for feature extraction. Considering only one of the quadrants reduces the computational complexity and computation time of feature extraction. Here, information in the third quadrant is considered, which reduces the feature size to 128 × 128. For each of the speech frames, similar process is repeated for the signals in train and test sets. Using these 128 × 128 feature vectors, a tensor is formed for each train and test set. On these tensors, HOSVD is applied to reduce the feature dimension from 128 × 128 to 10 × 10. For each frame, we will store the features as a 1 × 100 feature vector. On these features logarithm is applied and then entropy is used for the feature dimension reduction. The feature sizes considered in the experiments are [3 5 8 10 20 30 40 50 60 70 80 90 100]. The reduced feature set is applied to a Support Vector Machine (SVM) classifier with Radial Basis Function (RBF) kernel to determine the classification accuracy. In the analysis, LIBSVM tool is used [70].

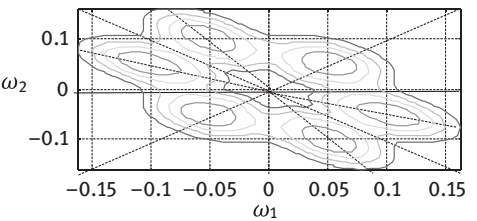

Figure 1.40: Symmetry regions of bispectrum. After [63, 64].

1.7.5 Experimental results

The performance of bispectrum features for the classification of infant cries is shown in Table 1.4 for different feature sizes. It is evident from the Table 1.4, that as the feature size increases, the classification accuracy also increases. This trend is observed till the feature size is 40, after that increasing the feature size resulted in the decrease in the classification accuracy. This happens because increasing the feature size increases redundancy after a certain value. Increasing the feature size causes more samples of pathological condition to be recognized as a normal sample by the classifier. The confusion matrix of the classification performance of the classifier is shown in Table 1.5 for the feature size of 1 × 30 for the normal and pathological cry classification. It can be observed that out of 427 samples of the normal cries 425 samples are correctly classified (99.53%), while 221 out 225 (98.43%) samples of the pathological condition are correctly classified by the classifier, resulting in a classification accuracy of 99.1% .

In Table 1.6, the performance of the proposed feature set is compared with the convention features such as MFCC, LPC and PLP. Under the same experimental conditions, it is found that the MFCC, PLP and LPC features are giving classification accuracies of 55.93%, 63.14% and 63.07% which is much lower than the proposed feature set [63]. This is due to the ability of the bispectrum features to capture the non-linearity in the infant cry signal which state-of-the-art features cannot capture. During dimension reduction technique of the bispectrum, when HOSVD is applied, the HOSVD theorem retains the principal components of the bispectrum thereby outperforms much better than other methods.

1.7.6 Robustness of the bispectrum features under noisy or signal degradation conditions

To understand the robustness of the bispectrum features in the presence of additive noise [63], let us consider a signal $s(n)$ corrupted by additive noise n(t) which can be babble noise, car noise, Gaussian noise or HF noise. The noisy signal can be represented by:

$$s_\eta(t) = s(t) + n(t). \tag{1.25}$$

where $n(t)$ can be additive babble, car, white, and HF channel noise. Let us assume that the noise under consideration has probability density function (*pdf*) which is Gaussian. The bispectrum of the noisy speech will be given as:

$$B_{s_n}(\omega_1, \omega_2) \approx B_s(\omega_1, \omega_2) + B_n(\omega_1, \omega_2) \tag{1.26}$$

Table 1.4: Classification accuracy (in %) for 4-fold cross-validation using holdout method.

Feature size		**3**	**5**	**8**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Test set	Test_1	78.88	90.46	92.81	94.72	97.07	99.12	99.56	98.68	98.24	97.654	96.77	95.30	94.13
	Test_2	84.36	96.20	97.95	98.83	99.85	100.00	99.56	99.27	99.12	98.83	98.10	97.08	96.64
	Test_3	78.62	95.02	97.22	98.24	98.98	99.41	99.56	99.41	98.54	98.39	97.80	97.80	97.66
	Test_4	78.65	91.96	94.01	95.91	97.08	97.22	97.08	97.08	97.08	96.64	95.91	94.44	92.98
	Mean	**80.13**	**93.41**	**95.50**	**96.92**	**98.24**	**98.94**	**98.94**	**98.61**	**98.24**	**97.88**	**97.14**	**96.16**	**95.35**

* In all experiments, one of the groups (Test_1, Test_2, Test_3, Test_4) is taken for testing and remaining for training the classifier.

Table 1.5: Confusion matrix of classification of normal and pathological cries using bispectrum as a feature set. Adapted from [66].

Identified as		Normal	Pathological
Actual	Normal	425	2
	Pathological	4	251

Table 1.6: Classification accuracy (in %) with MFCC, LPC, PLP, and bispectrum features. Adapted from [66].

Feature Set	Classification Accuracy (in %)
MFCC	53.99
LPC	63.07
PLP	63.14
Bispectrum	**98.94**

For noise with Gaussian distribution, the bispectrum of the noise is zero. $B_n(\omega_1, \omega_2) = 0$ and hence,

$$B_{s_n}(\omega_1, \omega_2) \approx B_s(\omega_1, \omega_2). \tag{1.27}$$

Thus, bispectrum features cause suppression of noise given the pdf of the noise is Gaussian and the noise is additive. For this reason, bispectrum features perform better than the other spectral features, such as LPCC, MFCC, etc., under noisy conditions.

The robustness of the proposed features under noisy conditions is shown in Table 1.7. In this experiment, different noises such as car noise, white noise, HF noise and babble noises are considered and the samples were taken from the NOISEX-2002 database [71]. These noises were added to the infant cry signals at different signal-to-noise-ratio (SNR) levels and then classification experiment were conducted using the bispectrum features. From the results shown in Table 1.7, we can observe that the classification performance of the bispectrum features remains almost the same for both the feature extraction methods even if the signal is corrupted by noise. The performance of the bispectrum features is good at SNR as low as –10 dB. Hence, bispectrum

Table 1.7: Effect of different noises on classification performance (in %) when the indirect method of bispectrum is used.

SNR (in dB)	Babble Noise (direct method)	Babble Noise (Indirect method)	White Noise	Car Noise	HF Channel
Without noise	**82.44 ± 4.04**	**81.65 ± 4.28**	**81.65 ± 4.28**	**81.65 ± 4.28**	**81.65 ± 4.28**
20	82.29 ± 3.81	81.37 ± 4.31	81.82 ± 4.29	81.57 ± 3.84	81.44 ± 4.30
15	82.44 ± 3.84	81.39 ± 4.59	81.82 ± 4.05	81.43 ± 4.21	81.66 ± 3.87
10	82.08 ± 3.82	81.73 ± 4.57	81.92 ± 4.27	81.39 ± 4.56	81.74 ± 3.64
5	81.81 ± 3.60	81.65 ± 4.12	81.64 ± 3.94	81.73 ± 4.25	81.79 ± 3.62
0	82.41 ± 3.54	81.83 ± 4.40	81.75 ± 3.95	81.16 ± 3.67	81.11 ± 3.41
−5	81.84 ± 4.35	80.72 ± 3.55	81.83 ± 3.17	80.38 ± 3.86	81.68 ± 3.66
−10	81.47 ± 4.61	81.22 ± 3.58	81.67 ± 4.05	80.27 ± 3.35	79.98 ± 4.11

features are more robust under noisy conditions and are suitable for infant cry classification task because in hospitals, the environment for data collection is noisy.

1.8 Summary and conclusions

In this chapter, the results of applying conventional signal processing methods on infant cry signals and the characteristics and classification of the normal and pathological infant cries are reported. From the STFT analysis of a cry signal, it is observed that the formants and harmonics are difficult to identify in an infant cry spectrum because of poor spectral resolution. This effect occurs due to the serious interaction of the fundamental frequency with the formant frequencies. The effect of increasing the LP order on the spectral matching problem is also observed. Increasing the LP order 'p' causes the LP spectrum peaks to match the pitch harmonics. The order of LP analysis is very low in infant cry analysis compared to adult speech because of the smaller vocal tract length. In the Cepstral analysis also, similar effect is observed where the lifter length is found to be very small compared to adults. Continuously changing vocal tract length in early days of life of an infant makes infant cry analysis more challenging because it causes continuous changes in LP order and lifter size. Finally, TEO analysis proved that

the infant cry production is a nonlinear mechanism. TEO also supports the identification of the glottal activity and no glottal activity in the infant cry signals.

In our experiments of infant cry classification task, it has been observed that the higher order spectral features perform much better than the conventional features. Bispectrum features also found to be noise robust which makes these features ideal for infant cry classification tasks in hospital environment. In order to reduce dimensions of the bispectrum features, other feature extraction methods can also be tried for this work. This extensive book chapter has tried to explore future research directions in this exciting and challenging field of infant cry analysis and classification.

Bibliography

[1] B. Lester, and C. Boukydis. No language but a cry, Nonverbal vocal communication, H. Papousek and U. Jurgens, Eds., 2008, 145–173.
[2] E. Gustafsson, F. Levrero, D. Reby, and N. Mathevon. Fathers are as good as mothers at recognizing the cries of their baby, Nature Communication, 4(2713), 1–6, Oct 2012.
[3] S. Gordon. Infant crying can trigger abuse in some parents, [Online]. Available:, https://consumer.healthday.com/caregiving-information-6/infant-and-child-care-health-news-410/infant-crying-can-trigger-abuse-in-some-parents-521678.html. [Accessed March 2019].
[4] P.S. Zeskind. Infant crying and synchrony of arousal, Evolution of Emotional Communication, January, 2013, 155–172.
[5] L.L. LaGasse, A.R. Neal, and B.M. Lester. Assessment of infant cry : Acoustic cry analysis and parental perception, Mental Retardation and Developmental Disabilities Research Reviews, 11(1), 83–93, 2005.
[6] J. Soltis. The signal functions of early infant crying, Journal of Behavioral and Brain Sciences, 27(1), 443–490, 2004.
[7] O. Wasz-Hockert, T. Partanen, V. Vuorenkoski, and E. Valanne. The identification of some specific meanings in the newborn and infant vocalization, Experientia, 20, 154, 1964.
[8] H.L. Golub, and M.J. Corwin. A physioacoustic model of the infant cry, Infant Crying-Theoretical and Research Perspective, B. M. Lester and C. Z. Boukydis, Eds., New York and London, Plenum Publishing Corporation, 1985, 59–82.
[9] O. Wasz-Hockert, K. Michelson, and J. Lind. Twenty five years of Scandinavian cry research, Infant Crying- Theoritical and Research Perspective, B. M. Lester and C. Z. Boukydis, Eds., New York and London, Plenum Publising corporation, 1985, 83–104.
[10] H.A. Patil. Infant identification from their cry, in 7th Int. Conf. on Advances in Pattern Recognition, Kolkata, India, 2009, 107–110.
[11] A. Messaoud, and C. Tadj. A cry based infant identification system, in 4th Int. conf. on Image and Signal Process., 2010, 192–199.
[12] Q. Xie. Automatic infant cry analysis and recognition, University of British Columbia (UBC), Canada, Dept. of Electrical engineering, Doctoral Thesis, 1993.

[13] Q. Xie, R.K. Ward, and C.A. Laszlo. Automatic assessment of infants' level of distress from the cry signals, IEEE Transactions on Speech and Audio Process., 4(4), 253–265, July 1996.
[14] Q. Xie, R.K. Ward, and C.A. Laszlo. Determing normal infant's level of distress from cry sounds, in Canadian Conf. on Elect. and Comp. Eng., Vancouver, BC, 2009, 1094–1097.
[15] A.K. Singh, J. Mukhopadhyay, and K.S. Rao. Classification of infant cries using epoch and spectral features in National Conf. on Comm., 2013, 1–5.
[16] S. Barajas-Montiel, and C. Reyes-Garcia. Identifying pain and hunger in infant cry with classifiers ensembles, in International Conference on Intelligent Agents, Web Technologies and Internet Commerce, 2, Vienna, 2005, 770–775.
[17] M. Petroni, A. Malowany, C. Johnston, and B. Stevens. Classification of infant cry vocalizations using artificial neural networks, in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 5, Detroit, MI, 1995, 3475–3478.
[18] M. Hariharan, R. Sindhu, and S. Yaacob. Normal and hypoacoustic infant cry signal classification using time-frequency analysis and general regression neural networks, Journal of Computer Methods and Programs in Biomedicine, 108(2), 559–569, 2012.
[19] J. Garcia, and C.A.R. Garcia. Mel frequency cepstrum coefficients extraction from infant cry for classification of normal and pathological cry with feed-forward neural networks, in Proceedings of the International Joint Conference on Neural Networks (IJCNN), July, 2003, 4, 3140–3145.
[20] J. Orozco, and C.A. Reyes García. Detecting pathologies from infant cry applying scaled conjugate gradient neural networks, European Symposium on Artificial Neural Network (ESANN), Bruges, Belgium, 2003, 349–354.
[21] A. Rosales-Perez, C.A. Reyes-Garcia, J.A. Gonzalez, and E. Arch-Tirado. Infant cry classification using genetic selection of a fuzzy model, 17th Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP), Argentina, Vol. 7441, 2012, 212–219, 2012.
[22] M. Hariharan, J. Saraswathy, R. Sindhu, W. Khairunizam, and S. Yaacob. Infant cry classification to identify asphyxia using time-frequency analysis and radial basis neural networks, Expert Systems with Applications, 39(10), 9515–9523, 2012.
[23] M. Hariharan, S. Yaacob, and S. Ardeen. Pathological infant cry analysis using wavelet packet transform and probabilistic neural network, Expert Systems with Applications, 38(12), 15377–15382, 2011.
[24] R. Sahak, W. Mansor, Y.K. Lee, A.I. Mohd Yassin, and A. Zabidi. Orthogonal least square based support vector machine for the classification of infant cry with asphyxia, in 3rd International Conference on Biomedical Engineering and Informatics (BMEI), Yantai, 2010, 986–990.
[25] O.F.R. Galaviz, and C.A.R. García. Infant cry classification to identify hypo acoustics and asphyxia comparing an evolutionary-neural system with a neural network system, in Advances in Artificial Intelligence: Proc. of 4th Mexican International Conference on Artificial Intelligence (MICAI), A. Gelbukh, A. Albornoz and H. Terashima-Marin (Eds.), Monterrey, Springer Berlin Heidelberg, 2005, 949–958.
[26] O.F. Reyes-Galaviz, and C.A. Reyes-Garcia. A system for the processing of infant cry to recognize pathologies in recently born babies with neural networks, in 9th Conference on Speech and Computer (SPECOM), St. Petersburg, Russia, 2004, 1–4.

[27] O. Reyes-Galaviz, S. Cano-Ortiz, and C. Reyes-Garcia. Validation of the cry unit as primary element for cry analysis using an evolutionary-neural approach, in Mexican International Conference on Computer Science (ENC), Baja California, 2008, 261–267.
[28] D. Lederman, A. Cohen, E. Zmora, K. Wermke, S. Hauschildt, and A. Stellzig-Eisenhauer. On the use of hidden Markov models in infants' cry classification, in The 22nd Convention of Electrical and Electronics Engineers, Israel, 2002, 350–352.
[29] H.F. Alaie, and C. Tadj. Cry-based classification of healthy and sick infants using adapted boosting mixture learning method for Gaussian mixture models, Modelling and Simulation in Engineering, 1–10, 2012.
[30] N. Buddha, and H.A. Patil. Corpora for analysis of infant cry, in Int. conf. on Speech Databases and Assessments, Oriental COCOSDA, Hanoi, Vietnam, 2007, 43–48.
[31] T.F. Quatieri. Discrete Time Speech Signal Processing, Pearson Education, 2004.
[32] "Factors influencing fundamental frequency," [Online]. Available: http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/influence.html. [Last Accessed March 2019].
[33] H. Patil, P. Dutta, and T. Basu. On the investigation of spectral resolution problem for identification of female speakers in Bengali, IEEE International Conference on Industrial technology (ICIT), Mumbai, 2006, 375–380.
[34] H.A. Patil. Speaker Recognition in Indian Languages: A feature based approach, Department of Electrical Engg., Indian Institute of Technology (IIT), Kharagpur, Doctoral Thesis, 2005.
[35] A.E. Rosenberg. Automatic speaker verification: A review, Proceedings of the IEEE, 64(4), 475–487, 1976.
[36] H. Teager. Some observations on oral air flow during phonation, IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(5), 599–601, Oct 1980.
[37] H. Teager, and S. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract, Speech Production and Speech Modelling, W. J. Hardcastle and A. Marchal, Eds., Springer, 1990, 241–261.
[38] J.F. Kaiser. On a simple algorithm to calculate the 'energy' of a signal, in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Albuquerque, NM, 1990, 381–384.
[39] K. Michelson, P. Sirvio, and O. Wasz-Hokert. Sound spectrographic cry analysis of infants with bacterial meningitis, Developmental Medicine and Child Neurology, 19(3), 309–315, 1977.
[40] O. Wasz-Hockert, K. Michelsson, and J. Lind. Twenty five years of Scandinavian cry research, Infant Crying- Theoritical and Research perspective, B. M. Lester and Z. Boukydis, Eds., New York, Plenum Publishing Corporation, 1985, 83–101.
[41] D. Gabor. Theory of communication, Journal of the Institution of Radion and Communication Engineering, 93(26), 429–441, November 1946.
[42] N. Samudrarajan, and N. Vasudha. Genesis of wavlet transform types and applications, Wavelets and Fractals in Earth System Sciences, E. Chandrasekhar, V. Dimri and V. M. Gadre, Eds., CRC Press, Taylor and Francis Group, 2014, 95–96.
[43] H.A. Patil. Cry Baby: Using spectrographic analysis to assess neonatal health from an infant's cry Advances in Speech Recognition, Mobile Environments, Call Centres and Clinics, A. Neustein, Ed, Springer-Verlag, 2010, 323–348.
[44] "Laryngomalacia details," [Online]. Available: http://emedicine.medscape.com/article/1002527-overview. [Last Accessed March 2019].
[45] J. Raes, K. Michelsson, and M. Despontin. Spectrographic analysis of the crying of infants with laryngeal disorders, Acta Oto-rhino-laryngologica Belgica, 34(3), 224–237, 1980.

[46] "Hearing loss – infants," [Online]. Available: https://www.nlm.nih.gov/medlineplus/ency/article/007322.htm. [Last Accessed March 2019].
[47] Autocorrelation Method [Online Available]. http://www.fit.vutbr.cz/~grezl/ZRE/lectures/05_pitch_en.pdf. {Last Accessed March 2019}.
[48] N.A. Weiss, and M.J. Hasset. Introductory Statistics, Addison-Wesley, 1993.
[49] L. LaGasse, A. Neal, and B. Lester. Assessment of infant cry: acoustic cry analysis and parental perception, Mental retardation and developmental disabilities research reviews, 11(1), 83–93, 2005.
[50] R. Y. Moon "SIDS and other sleep related infant deaths: Expansion of recommendations for a safe infant sleeping environment", Pediatrics, 128(5), 1341–1367, Nov 2011.
[51] "The World Bank," [Online Available]. http://data.worldbank.org/indicator/SP.DYN.IMRT.IN. {Last Accessed March 2019}.
[52] R.H. Colton, and A. Steinschneider. The cry characteristics of an infant who died of the sudden infant death syndrome, Journal of Speech and Hearing Disorders, 4(46), 359–363, 1981.
[53] D. Harrison. Histologic evaluation of the larynx in sudden infant death syndrome, Annals of Otology, Rhinology and Laryngology, 100, 173–175, 1991.
[54] D. Bone, T. Chaspari, K. Audkhasi, J. Gibson, A. Tsiartas, M.V. Segbroeck, M. Li, S. Lee, and S. Narayanan. Classifying language related developmental disorders from speech cues: the promise and the potential confounds, INTERSPEECH, Lyon, France, 2013, 182–186.
[55] D. Mehta, and R.E. Hillman. Use of aerodynamic measures in clinical voice assessment, Perspectives on Voice and Voice Disorders, 17(3), 14–18, Nov 2007.
[56] D.D. Mehta, and R.D. Hillman. Current role of stroboscopy in laryngeal imaging, Current Opinion in Otolaryngology and Head and Neck Surgery, 20(6), 429, Dec 2012.
[57] S.A. Selouani, M.S. Yakoub, and D.D. O'Shaughnessy. Alternative speech communication system for person with severe speech disorders, EURASIP Journal of Advance in Signal Process, 1(1), 249–254, Jun 2009.
[58] T. Binzoni, C.S. Seelamantula, and D.V.D. Ville. A fast time-domain algorithm for the assessment of tissue blood flow in laser Doppler flowmetry, Journal of Physics in Medicine and Biology, 55(13), N383, 2010.
[59] C.S. Seelamantula, and T.V. Sreenivas. Blocking artifacts in speech/ audio: Dynamic auditory model based characterization and optimal time-freqency smoothing, Elsvier Journal of Signal Processing, 89(4), 523–531, 2009.
[60] K. Elemetrics. Massachusetts Eye & Ear Infirmary Voice Disorder Databse (MEEI Database): Elemetrics Disordered Voice Database (version 1.03), 1994.
[61] C. Nikias, and R.R. Mysore. Bispectrum estimation: A digital signal processing framework, Proceedings of the IEEE, 75(7), 869–891, July 1987.
[62] C. Nikias, and J. Mendel. Signal processing with higher-order spectra, IEEE Signal Processing Magazine, 10(3), 10–37, July 1993.
[63] A. Chittora. Crying for a reason: A signal processing based approach for infant cry analysis and classification, Ph.D. Thesis, Dhirubhai Ambani Institute of Informations and Communication Technology, Gandhinagar, India, 2017.
[64] A. Chittora, and H.A. Patil. Significance of Higher-Order Spectral Analysis in Infant Cry Classification, Circuits, Systems, and Signal Processing, Springer, Vol. 37, 1, 232–254, 2018.

[65] L.D. Lathauwer, B.D. Moor, and J. Vandewalle. A multilinear singular value decomposition, SIAM Journal Matrix Analysis Applications, 21(4), 1253–1278, 2000.
[66] A. Chittora, and H.A. Patil. "Analysis of normal and pathological infant cries using bispectrum features derived using HOSVD," in Int. Conf. on Biosig. Ana., Proc. and Systems (ICBAPS), Kuala Lumpur, 2015, 151–155.
[67] A. Chittora, and H.A. Patil. "Classification of normal and pathological infant cries using bispectrum features," in 23rd Euro. Sig. Proc. Conf. (EUSIPCO), Nice, 2015, 639–643.
[68] A. Chittora, and H.A. Patil. "Classification of pathological infant cries using modulation spectrogram features," in 9th Int. Symp. Chinese Spoken Lang. Proc. (ISCSLP), Singapore, 2014, 541–545.
[69] A. Chittora, and H.A. Patil. "Classification of phonemes using modulation spectrogram based features for Gujarati languages", Int. Conf. on Asian Lang. Proc. (IALP), Kuching, 2014, 46–49.
[70] C.C. Chang, and C.-J. Lin. LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, 27(2), 1–27, 2011.
[71] NOISEX-92 [Online]. http://www.speech.cs.cmu.edu/comp.speech/Section/Data/noisex.html. {Last Accessed on March 2019}.

Hardik B. Sailor and Hemant A. Patil

2 Unsupervised auditory filterbank learning for infant cry classification

Abstract: The infant cry classification is a socially relevant problem where the task is to classify the normal versus pathological cry signals. Since the cry signals are very different from the speech signals, there is a need of better feature representation for infant cry signals. Recently, representation learning is very popular in various signal processing areas including the medical domain. In this chapter, we propose to use unsupervised auditory filterbank learning using convolutional restricted Boltzmann machine (ConvRBM). Analysis of the subband filters shows that they are very distinct compared to the subband filters learned from the speech signals. Various cry models were analyzed using ConvRBM spectrogram for normal and pathological cry signals. The infant cry classification experiments were performed on the two databases, namely, DA-IICT Infant Cry and Baby Chillanto. The experimental results show that the proposed features perform better than the standard mel-frequency cepstral coefficients (MFCC) using various statistically meaningful performance measures. In particular, our proposed ConvRBM-based features obtained an absolute improvement of 2% on the DA-IICT Infant Cry database and 0.58% on the Baby Chillanto database in the classification accuracy. Since, the auditory filterbanks are learned from the infant cry signals, it is optimal to represent the statistical structures in the infant cry signals. Hence, it performs better then standard handcrafted feature sets such as the MFCC.

Keywords: Auditory representation learning, Convolutional Restricted Boltzmann Machine, auditory filterbank, subband filters, and infant cry classification

2.1 Introduction

Humans cry to express a range and degree of emotions, such as from happiness after passing a tough exam or meeting a beloved one to grief after the death of

Note: Hardik B. Sailor is now at Samsung Research and Development Institute Bengaluru, Karnataka 560037, India. The work was done while the author was at DA-IICT Gandhinagar, India, and it does not contain any Samsung Research and Development Institute proprietary information.

Hardik B. Sailor, Speech and Hearing Research Group, The University of Sheffield, UK
Hemant A. Patil, Speech Research Lab, DA-IICT, Gandhinagar, Gujarat, India

https://doi.org/10.1515/9781501513138-002

a person or difficult situations in life [1]. On the whole, the crying is not just a simple reaction to any feeling or emotional state but rather a multifaceted behavior that can offer clues to how we process and regulate our feelings, and how we experience the world around us [1]. The evolutionary background of crying is discussed in a book [2], where it is shown that only humans have the ability to cry not other mammals. In humans, infants communicate their needs such as feeding, distress or pain by crying [3]. Intra-individual variation in the infant cries is known to encode qualitative and quantitative information on the condition, needs, emotional status and the degree of urgency. Infant cry carries multiple levels of information as shown in Figure 2.1. Based on the perception of the cry, the parents or caretakers empirically try to understand the reason for the crying and even identify their newborn [3]. Recently, there is an increasing effort to investigate the reasons for sudden infant death syndrome (SIDS) [4] through the analysis of infant cry signals. The infant cry analysis is also valuable in the clinical diagnostics in order to know, whether the disease to the newborn is due to the central nervous system (CNS) [5]. From a signal processing perspective, our goal is to classify whether the infant is crying due to the pain, hunger or some medical diseases collectively called *pathology*.

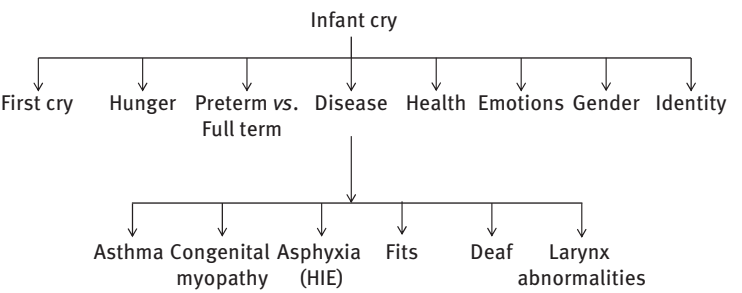

Figure 2.1: Multiple levels of the information present in the infant cry signal. Adapted from [4].

The basic block diagram of infant cry classification task is shown in Figure 2.2. In the training phase, the pattern classifier is trained using acoustic features extracted from the infant cry signals. The model obtained from the training phase is used to test the cry signal for the presence of pathology based on the decision logic (e.g., log-likelihood ratio). Here, we briefly discuss the recent approaches for infant cry databases and classification using signal processing and pattern recognition techniques. To date, there is no standard publicly available database for infant cry classification. Many researchers collected their own data including our Speech Research Group at DA-IICT [6, 7]. Most of the studies used Baby Chillanto infant cry database, which is a property of INAOE-CONACyT, Mexico [8]. Other studies

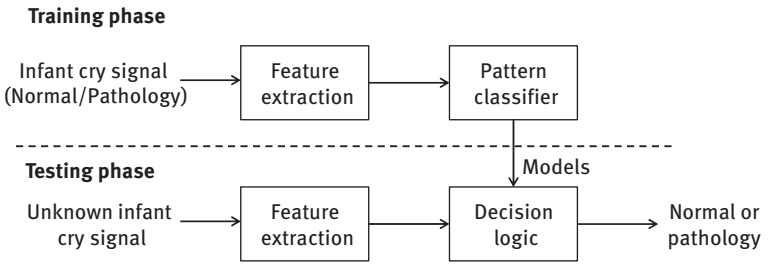

Figure 2.2: The basic block diagram of the infant cry classification.

include the work in [9–13]. The literature on the infant cry classification using this database is given in [8].The detailed discussion on the topic of infant cry and the literature of infant cry classification is found in [4] and in [14], the first Ph.D. thesis from India in this area (to the best of authors' knowledge). Most of the studies used mel-frequency cepstral coefficients (MFCC) as an auditory-based features. Recently, representation learning (RL) is very popular to learn meaningful feature representation directly from the raw audio signals. Various approaches were proposed for RL that shows significant improvements compared to the handcrafted features, such as MFCC. In this chapter, the objective is to use our proposed convolutional restricted Boltzmann machine (ConvRBM) for auditory-like filterbank learning from the raw audio signal [15, 16]. We used two databases, namely, (1) Baby Chillanto [8] and (2) DA-IICT infant Cry database collected by our group [6]. The experimental results showed improved performance with the proposed feature representation.

The organization of the rest of the chapter is as follows: Section 2.2 describes how the problem of infant cry classification is socially relevant. The significance of representation learning for auditory modeling is presented in Section 2.3. The architecture, training methodology and feature representation using ConvRBM are presented in Section 2.4. The experimental setup for infant cry classification task is given in Section 2.5. The analysis of the ConvRBM filterbank and experimental results are presented in Section 2.6 and Section 2.7, respectively. Finally, the summary and conclusions of the presented work are given in Section 2.9.

2.2 Social relevance of the infant cry classification

The infant cry classification task is very helpful to the parents, caretakers and pediatricians in the diagnosis of a pathology at very early stage. This may be beneficial to reduce or completely eliminate symptoms of a pathology. Many

times avoiding pathology at an early stage of infant development leads to the severe conditions including death. The research work in this direction is also important to identify the appropriate reasons for SIDS [17, 18]. The first study of SIDS case was analyzed in [19]. The SIDS is the sudden unexplained death of a child less than one year of age that remains unexplained even after a complete investigation [20]. Data from the Centers for Disease Control and Prevention (CDC) show that 1,545 infants died from SIDS in 2014 (the most recent year for which data is available) [20]. The research evidence suggests that infants who die from the SIDS are born with brain abnormalities or physiological defects [20]. Hence, to prevent SIDS cases the study of infant cry analysis will be very much helpful. Another important social relevance is to the families where the literacy level is lower and access to the good hospitals is difficult specifically in the remote village places. The infant cry classification if implemented in the mobiles (since mobiles are nowadays available almost everywhere) can be beneficial in the initial detection (as early warning signs) of the pathology in the infants through cry signals.

2.3 Representation learning for auditory modeling

The features of human speech perception, vision and in other cognition tasks do not exist rather they are learned through the experience as we grow [21]. The new area of machine learning has emerged since 2006 that has a significant impact on many signal processing applications, such as speech, audio, image, etc. Since the early techniques started with learning from unlabeled data, it is called the representation learning (RL). Later impressive results also achieved using supervised learning techniques by using many parallel data processing units with nonlinearities. It is now called deep learning or hierarchical learning. RL is defined as learning the representations of the data that makes it easier to extract meaningful and useful information when building classifiers or other predictors for signal processing applications. In this chapter, we use the RL-based technique to model auditory processing. The advantage of using RL to learn auditory-like filterbank is that the subband filters obtained are *optimal* for the given task since it uses statistics of the underlying database. Unsupervised representation learning is currently a very active research area where we do not have labels or have very limited amount of database. Supervised RL techniques are very effective when we have a large amount of data available to train the classifier. Unsupervised learning is the most important form of representation learning since human learning is largely unsupervised [12]. An example is the language acquisition by the infants during initial stages of their growth, which is also type of unsupervised learning [22]. Most work on unsupervised learning for speech and audio

signals are based on cochlear filterbank learning to model auditory processing. Some of the studies model auditory cortex-level mechanisms by learning spectro-temporal receptive fields (STRFs). The first approach to model cochlear filterbank is presented in the remarkable study by M. S. Lewicki where the subband filters learned from various natural sounds are analyzed. We have developed unsupervised auditory filterbank learning using ConvRBM directly from the full-length raw speech and audio signals. The ConvRBM filterbanks successfully applied in the automatic speech recognition (ASR) [15, 16, 23], environmental sound classification (ESC) [24] and spoof speech detection (SSD) task [25, 26]. Motivated by these studies, in this chapter, we explore the potential of the ConvRBM filterbank learning for the infant cry classification task. In the next section, we present the architecture and feature representation using ConvRBM.

2.4 Unsupervised auditory filterbank learning

In this section, we present the architecture of the proposed filterbank learning model and feature extraction.

2.4.1 Convolutional restricted Boltzmann machine

ConvRBM is an undirected probabilistic graphical model with two layers, namely, a visible layer and a hidden layer [16]. The block diagram of the arrangement of the hidden units is shown in Figure 2.3. The input to the visible layer (denoted

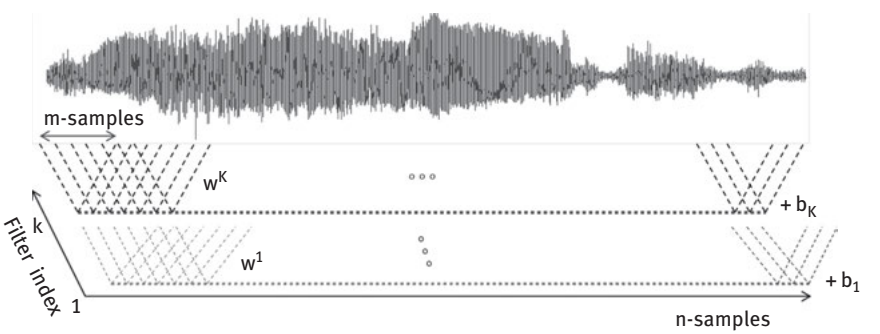

Figure 2.3: The arrangement of hidden units in the K groups, and corresponding weight connections. The filter index-axis is perpendicular to the plane of this page. Each hidden unit (red dots) in the kth group is wired such that it resulted in a valid convolution between the infant cry signal and weights, W^k. After [16].

as $\mathbf{x}$) is an infant cry signal of length n-samples. Hidden layer (denoted as $\mathbf{h}$) consists of K-groups (i.e., number of filters) with filter length m-samples in each. Weights (also called subband filters) are shared between visible and hidden units among all the locations in each group [16]. Denoting b_k as the hidden bias for the kth group, the convolutional response for the kth group is given as [16],

$$\mathbf{I}_k = (\mathbf{x}^* \tilde{\mathbf{W}}^k) + b_k, \tag{2.1}$$

where $\mathbf{x} = [x_1, x_2, ..., x_n]$ are samples of the infant cry signal, $\mathbf{W}^k = [w_1^k, w_2^k, ..., w_m^k]$ is a weight vector (i.e., kth subband filter) and $\tilde{\mathbf{W}}$ denote a *flipped* array [16]. For ConvRBM with visible units $\mathbf{x}$, and hidden units $\mathbf{h}$, the energy function of the model is given as [15, 16]:

$$\begin{aligned} E(\mathbf{x}, \mathbf{h}) = \frac{1}{2\sigma_x^2} \sum_{i=1}^{n} x_i^2 - \frac{1}{\sigma_x} \sum_{k=1}^{K} \sum_{j=1}^{l} \sum_{r=1}^{m} \left(h_j^k w_r^k x_{j+r-1} \right) \\ - \sum_{k=1}^{K} b_k \sum_{j=1}^{l} h_j^k - \frac{1}{\sigma_x^2} c \sum_{i=1}^{n} x_i, \end{aligned} \tag{2.2}$$

where c is a visible bias, which is also shared. We have used "valid" length convolution and, hence, the length of each group is $l = n - m + 1$. Each infant cry signal is normalized to a zero-mean and a unit variance so that the variance parameter (σ_x) in eq. (2.2) is set to 1 as suggested in [27]. The probability of joint distribution of visible and hidden units is,

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{x}, \mathbf{h})}, \tag{2.3}$$

where Z is the partition function, $Z = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-E(\mathbf{x}, \mathbf{h})} d\mathbf{x} d\mathbf{h}$, which normalizes the energy, and thereby making it a probability distribution function (PDF). With noisy rectifier linear units (NReLU), the sampling equations for hidden and visible units are given as [15, 16]:

$$\begin{aligned} \mathbf{h}^k \quad &\sim max(0, \mathbf{I}_k + N(0, \sigma(\mathbf{I}_k))), \\ \mathbf{x}_{recon} \quad &\sim N\left(\sum_{k=1}^{K} (\mathbf{h}^k * \mathbf{W}^k) + c, 1 \right), \end{aligned} \tag{2.4}$$

where c is a visible bias which is also shared, $N(0, \sigma(I_k))$ is a Gaussian noise with mean-zero and sigmoid of $\mathbf{I}_k$ as a variance. While calculating the relationship between hidden and visible units, a deterministic ReLU (i.e., $max(0, \mathbf{I}_k)$) is used as an activation function [16]. In ConvRBM training, a dropout is applied before sampling the hidden units in both positive and negative phase of contrastive divergence (CD) learning. Applying a dropout to ConvRBM can be thought of as multiplying each unit in a kth group with a binary mask (called

as the *dropout mask*). The dropout mask for the kth group is defined as random variables drawn from the Bernoulli distribution, that is,

$$\mathbf{m}_k = Bernoulli(p), \tag{2.5}$$

where $P(m_k = 1) = p$ and $P(m_k = 0) = 1 - p$. The sampling equation for the hidden units is now given as:

$$\mathbf{h}^k \sim max(0, \mathbf{m}_k \odot \mathbf{I}_k + N(0, \sigma(\mathbf{m}_k \odot \mathbf{I}_k))), \tag{2.6}$$

where $\odot$ indicates an element-wise multiplication. We have explored an annealed dropout training of ConvRBM that was proposed for supervised deep networks in [28]. Our earlier works showed that the use of annealed dropout resulted in an improved performance in speech recognition [23] and audio classification task [24, 26]. In an annealed dropout, the dropout probability of the units in the network is gradually decreased over the training period. We have used the following annealing dropout schedule as suggested in [28]:

$$p[t] = max\left(0, \left(1 - \frac{t}{N}\right) \cdot p[0]\right), \quad t \in [0, N], \tag{2.7}$$

where $p[0]$ is the initial dropout rate at training iteration, $t = 0$. The dropout rate is decayed from $p[0]$ to a small value or zero for $t = N$ iterations. After N iterations, $p[t]$ is kept constant as 0, that is, no dropout. The log-likelihood of the ConvRBM is denoted as $\ell(\mathbf{x}; \theta)$, where $\theta = (\mathbf{W}^k, b_k, c)$ are model parameters. With the notations used in [29], we can write the log-likelihood of the ConvRBM in terms of expectations as [16]:

$$\begin{aligned} \frac{\partial}{\partial \theta} \ell(\mathbf{x}; \theta) &= -\mathbb{E}_{p(\mathbf{h}|\mathbf{x})}\left[\frac{\partial}{\partial \theta} E(\mathbf{x}, \mathbf{h})\right] + \mathbb{E}_{p(\mathbf{h}, \mathbf{x})}\left[\frac{\partial}{\partial \theta} E(\mathbf{x}, \mathbf{h})\right], \\ &\approx -\left\langle \frac{\partial}{\partial \theta} E(\mathbf{x}, \mathbf{h}) \right\rangle_{data} + \left\langle \frac{\partial}{\partial \theta} E(\mathbf{x}, \mathbf{h}) \right\rangle_{model}, \end{aligned} \tag{2.8}$$

where $\langle \cdot \rangle$ is the sample mean under distribution used to calculate expectations. Here, $\langle \cdot \rangle_{data}$ is the sample mean estimated, when the visible units are clamped to the signal (i.e., input data), and $\langle \cdot \rangle_{model}$ is the sample mean estimated, when visible and hidden units are sampled from a model distribution. For the weights of the model, eq. (2.8) can now be written as [16]:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{W}^k} \ell(\mathbf{x}; \theta) &= \left\langle conv(\mathbf{x}, \tilde{\mathbf{h}}^k) \right\rangle_{data} - \left\langle conv(\underline{\mathbf{x}}, \underline{\tilde{\mathbf{h}}^k}) \right\rangle_{model}, \\ \frac{\partial}{\partial b_k} \ell(\mathbf{x}; \theta) &= \left\langle \sum_{j=1}^{l} h_j^k \right\rangle_{data} - \left\langle \sum_{j=1}^{l} \underline{\tilde{h}^k} \right\rangle_{model}, \\ \frac{\partial}{\partial c} \ell(\mathbf{x}; \theta) &= \left\langle \sum_{i=1}^{n} x_i \right\rangle_{data} - \left\langle \sum_{i=1}^{n} \underline{x_i} \right\rangle_{model}. \end{aligned} \tag{2.9}$$

where the underline symbol denotes visible ($\underline{\mathbf{x}} = \mathbf{x}_{recon}$), and the hidden states ($\underline{\tilde{\mathbf{h}}^k}$) in the CD-1 stage (negative phase). The model parameters are updated using the Adam optimization method [30]. In the next section, we discuss how to extract features once the ConvRBM is trained.

2.4.2 Auditory feature representation

After ConvRBM is trained, the pooling is applied to reduce the representation of ConvRBM filter responses in the temporal domain. Here, pooling in the time domain is equivalent to short-time averaging in spectral features, such as MFCC and low-pass filtering in scattering wavelets. For an audio signal of sampling frequency, Fs = 16 kHz, pooling is applied using 25 ms (i.e., 400 audio samples) window length (*wl*) and 10 ms (i.e., 160 audio samples) shift (*ws*). We used this setup to compare standard spectral features (e.g., MFCC) extracted using same windowing parameters. The infant cry signal with *n*-samples has, $F = \frac{n - wl + ws}{ws}$ number of frames. We have experimented with both average and max-pooling and found better results with the max-pooling. After the pooling operation, stabilized logarithm $\log(\cdot + \delta)$ (with $\delta = 0.0001$) is applied as a compressive nonlinearity. The block diagram for feature extraction procedure (described above) is shown in Figure 2.4.

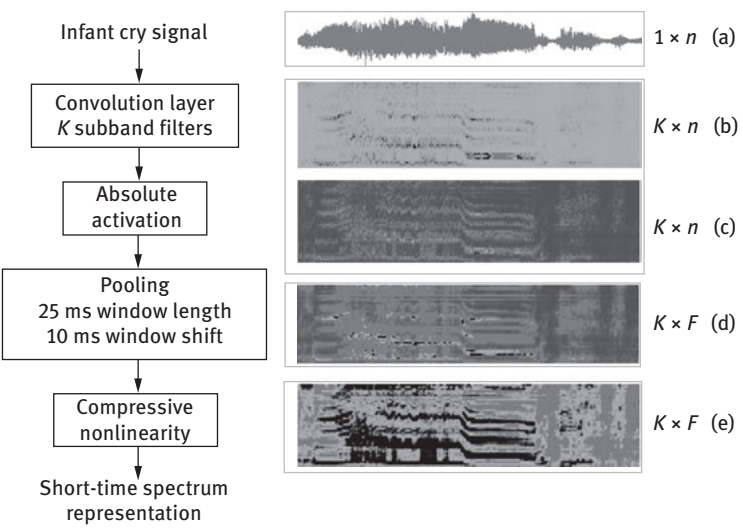

Figure 2.4: Feature extraction using trained ConvRBM: (a) infant cry signal, (b) and (c) responses from the convolutional layer and ReLU nonlinearity, respectively, (d) representation after pooling and (e) logarithmic compression. After [16].

To obtain the same length as the audio signal, "same" length convolution is used. During feature extraction stage, we have used absolute nonlinearity $|\mathbf{I}_k|$ as an activation function of the hidden units as done in [26]. The pooling operation reduces the temporal resolution from $K \times n$ samples to the $K \times F$ frames. Logarithmic nonlinearity compresses the dynamic range of features. The feature extraction steps involved in this ordering resemble the auditory processing in the human auditory system (HAS) [31, 32].

2.5 Experimental setup

2.5.1 Databases

The experiments were performed with two databases: DA-IICT Infant Cry database developed by our group, and Baby Chillanto discussed as follows.

DA-IICT Infant Cry database

The DA-IICT Infant Cry database was collected as a part of B.Tech project work, Ph.D. thesis work and the DST fast-track award for young scientist to Prof. Hemant A. Patil for the project "Development of Infant Cry Analyzer using Source and System Features" [6]. The infant cry data was collected from three hospitals in Visakhapatnam, namely, (1) King George Hospital, (2) Prabha Nursing Home and (3) Child Clinic. The sampling frequency of the original recordings was 12 kHz, quantized at 16-bit PCM. For our experiments, we downsample it to 11.025 kHz since at a later stage, we will compare the experimental results with another database. The statistics of the DA-IICT Infant Cry database is shown in Table 2.1. The healthy cry signals consist of normal and hunger cry signals. The pathology cry includes two types of pathologies, namely, asphyxia (also called Hypoxic Ischemic Encephalopathy (HIE)) and Asthma.

Table 2.1: Description of DA-IICT Infant Cry database. After [6].

Class	Category	No. of samples
Healthy	Normal, hunger	793
Pathology	Asphyxia	215
	Asthma	182

Baby Chillanto database

Baby Chillanto database was developed by the recordings conducted by the medical doctors. The infant cry signals were carefully labeled at the time of the recording with the references, such as the reason for crying, sick or not, and infant age. Each cry signal was segmented into one second duration (that represents one sample) and have been grouped into five categories as shown in Table 2.2. Since the sampling rate of cry signals is different in all the categories, we kept the sampling rate of 11.025 kHz. Two groups were formed for the binary classification of healthy versus pathology. Healthy cry signals include three categories, namely, normal, hungry and pain resulting in 1,049 cry samples. Pathology cry signals include two categories, namely, asphyxia and deaf resulting in 1,219 cry samples.

Table 2.2: Description of Baby Chillanto database. After [8].

Class	Category	No. of samples
Healthy	Normal	507
	Hungry	350
	Pain	192
Pathology	Asphyxia	340
	Deaf	879

2.5.2 Training of ConvRBM and feature extraction

The ConvRBM is trained with an annealed dropout using $p = 0.3$ that decayed to zero (i.e., $p = 0$) during training. The learning rate was chosen to be 0.001 and decayed according to the learning rate schedule as suggested in [30]. The moment parameters of Adam optimization chosen to be $\beta_1 = 0.9$ and $\beta_2 = 0.999$ similar to other audio classification experiments [24]. The model is trained with 40 number of subband filters (i.e., K) with convolution window length $m = 88$ samples (i.e., 8 ms). After the model was trained, the features were extracted from the infant cry signals. The Discrete Cosine Transform (DCT) was applied to reduce the dimension retaining only first 13-D coefficients and compare the proposed features with MFCC feature set in the Gaussian Mixture Model (GMM) framework. The delta and delta-delta features were also appended resulting in 39-D cepstral features (denoted as ConvRBM-CC). The baseline MFCC features are extracted from the cry signals with 25 ms window length and 10 ms window shift.

2.5.3 Binary classification framework

Since the infant cry databases are very small in size, the GMM is used for binary classification. Healthy cry features belong to one class, and pathology cry features belong to another class. The GMMs with different mixture components were trained using MFCC and ConvRBM-CC features. Final scores are represented in terms of the log-likelihood ratio (LLR). The decision of the test cry signal being normal or pathology is based on the LLR, that is,

$$LLR = \log \frac{p(\mathbf{X} \mid H_0)}{p(\mathbf{X} \mid H_1)}, \tag{2.10}$$

where $p(\mathbf{X} \mid H_0)$ and $p(\mathbf{X} \mid H_1)$ are the likelihood scores from the GMM for a normal and pathology trials (with hypothesis H_0 and H_1), respectively, for features X. The results are predicted using log-likelihood scores with 10-fold cross-validation (CV). Since the number of samples in the two classes is different, we applied 10-fold CV separately for each class and then combine respective training and test folds. For each fold, we noted % classification accuracy. The final result is presented as averaged % classification accuracy over all the 10-folds. Along with classification accuracy, other performance measures are also used described in the following section.

2.5.4 Performance measures

A significance of the proposed features is evaluated using various performance measures. The confusion matrix of the binary classification task shows how errors are distributed across the classes [33]. The example of a confusion matrix for classification task is shown in Figure 2.5. The rows indicate the actual classes, and columns indicate the predicted outcome of the classifier [33]. Since our task is to detect the pathology in an infant cry, we denote the results associated with pathology as positive and negative for vice-versa. Given the labels of actual and predicted classes by the classifier, there are four outcomes possible [33]:

		Predicted outcomes	
		Normal	Pathology
Actual classes	Normal	TN	FP
	Pathology	FN	TP

Figure 2.5: The details of a confusion matrix for the binary classification task.

- True positive (TP): Actual class is pathology and predicted pathology
- True negative (TN): Actual class is normal and predicted normal
- False positive (FP): Actual class is normal and predicted pathology
- False negative (FN): Actual class is pathology and predicted normal

In the case of *k*-fold CV, we find the combined confusion matrix (i.e., all the entries in the matrix are summed for all folds). Various other performance measures can be obtained from the confusion matrix. The numbers along the major diagonal indicate (TP and TN) the correct decisions made by the classifier [33]. The classification accuracy can also be obtained from TP, TN and a total number of instants of both the classes (i.e., P + N) as follows [33]:

$$\text{Classification accuracy}\,(\%) = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}. \tag{2.11}$$

Another important performance measure is F1-score also known as *F*-measure. The range of *F*-measure is between 1 and 0, where 1 represents the perfect prediction and 0 means the worst. The *F*-measure is defined as follows [33]:

$$F - \text{measure} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{2.12}$$

F-measure does not take TN into account. Hence, we also used another performance measure called Youden's *J*-statistic or informedness [34]. The range of *J*-statistic is between −1 and +1, where −1 indicates no agreement between the observation and the prediction, and +1 represents a perfect prediction. *J*-statistic estimates the probability of an informed decision and is given by [34]:

$$J - \text{statistic} = \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} - 1. \tag{2.13}$$

Another important performance measure is the Matthews correlation coefficient (MCC) [35]. It takes into account TP, TN, FP, FN and is generally regarded as a balanced measure, which can be used even if the classes are of very different sizes. This is very important in our case due to a different number of samples in each class (i.e., normal *vs*. pathological). The range of MCC is between −1 and +1 where +1 indicates a perfect prediction, 0 means no better than just a random prediction, and 1 indicates a total disagreement between the observation and the prediction. MCC is expressed as [35]:

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TP} + \text{FN})(\text{TN} + \text{FP})}}. \tag{2.14}$$

The feature visualization is performed using t-SNE technique [36]. t-SNE (Stochastic Neighbor Embedding) is a high-dimensional data visualization technique using student's t-distribution [36]. It maps the high-dimensional data onto two or three dimensions. We used t-SNE to visualize 39-D MFCC and ConvRBM-CC feature vectors into three dimensions. This visualization helps us to see the class separability of MFCC and ConvRBM-CC features.

2.6 Analysis of infant cry signals

2.6.1 Analysis of subband filters and frequency scale

The subband filters learned from the DA-IICT Infant Cry database and Baby Chillanto infant cry database are shown in Figure 2.6 in the time and frequency domain. We have also shown subband filters obtained from the TIMIT speech database. It is very interesting to note an intriguing observation that, these subband filters were learned from only 37 min and 50 s duration of cry signals from the Baby Chillanto and 30 min of cry signals from the DA-IICT Infant Cry database (which is all the more the case in medical scenarios). It shows the applicability of our proposed model even in the very small scale database scenarios. The time-domain subband filters are significantly different than speech database. The subband filters of the infant cry databases contain more Fourier-like basis functions due to harmonic nature of the infant cry signals as shown in Figure 2.7. The analysis of the frequency-domain subband filters revealed that many subband filters are not localized and contain harmonic structures. This may be due to more harmonic content present in infant cry signals. On comparing the subband filters learned from the two different databases, the subband filters from the baby Chillanto database has even lower frequency filters. However, the filter shapes of most of the subband filters are similar.

The frequency scales obtained using ConvRBM are compared with the standard auditory frequency scales in Figure 2.8. Unlike the frequency scale obtained through the speech database [16], here we observed two linear segments in the frequency scale, from 0 to 1 kHz and from 1 to 3 kHz. After 3 kHz, it is nonlinear and follows the ERB, and Bark scales. However, the frequency scale from the DA-IICT Infant Cry database is more away from the standard scales. The minimum center frequency is 500 Hz and 4 kHz it follows the other frequency scales. The difference in the frequency scales of both the databases is due to variability in cry signal production through language perception (Indian languages *vs*. English in the Baby Chillanto), data recording conditions, background noise, channel

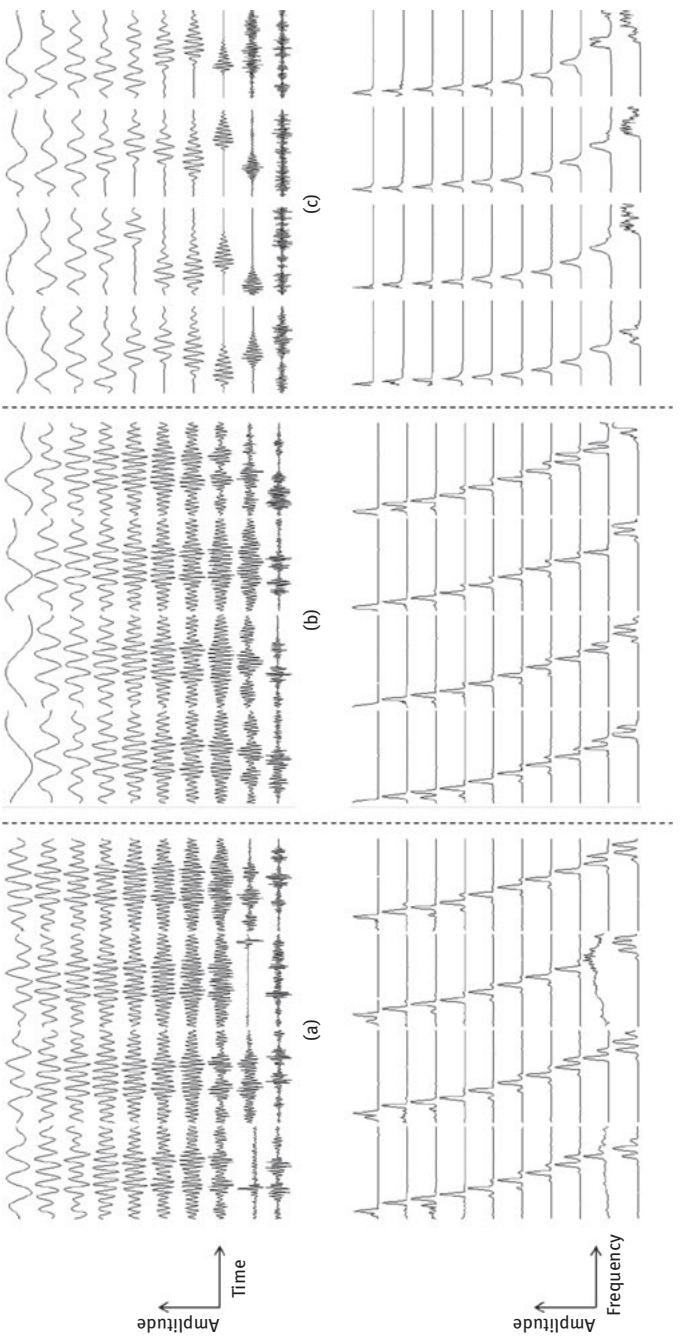

Figure 2.6: The examples of subband filters trained on DA-IICT Infant Cry (Panel I), Baby Chillanto (Panel II) and TIMIT (Panel III) databases, respectively: (a)–(c) subband filters in the time domain (i.e., impulse responses), (d)–(f) corresponding subband filters in the frequency domain (i.e., frequency responses).

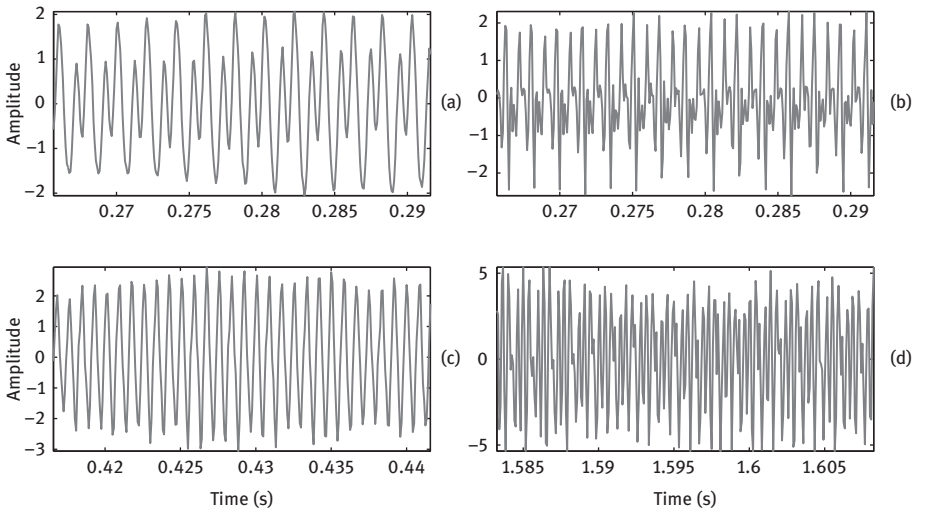

Figure 2.7: Segments of the infant cry signals showing the harmonic nature of the cry signals for different categories: (a) normal, (b) deaf, (c) asphyxia and (d) asthma.

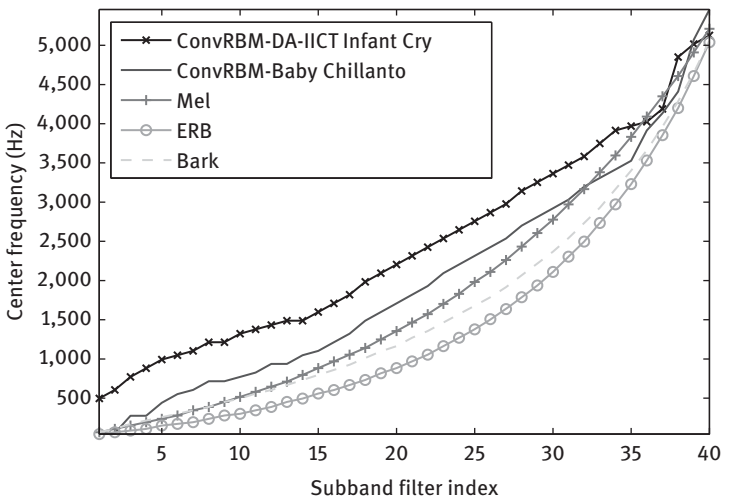

Figure 2.8: Comparison of the filterbanks learned using the ConvRBM with the standard auditory filterbanks.

characteristics, microphone specifications, etc. Overall, the frequency scale of ConvRBM still follows standard auditory frequency scales (with piecewise linear segments up to 3 kHz).

2.6.2 Analysis of ConvRBM spectrograms

The spectrogram representation for the cry signals using the ConvRBM filterbank is shown in Figure 2.9–2.13 along with the mel spectrograms. We will discuss each case in detail as follows:

Normal infant cry signal

The spectrograms from the three normal infant cry signals taken from the Baby Chillanto database are shown in Figure 2.9. A better time-frequency resolution is obtained using ConvRBM filterbank as marked in the spectrograms, specifically in the high frequency regions. We can see the slowly varying harmonic structures and some noise (this is predominantly due to the turbulent excitation source and not due to the environmental noise) in the normal cry signals that are related to the *cry modes* as observed in [6, 14]. Figure 2.9(a) is an example of falling, (b) is an example of flat and (c) is an example of rising with vibration cry mode. We can also observe dysphonation cry mode in Figure 2.9(a) after 0.2 s along with the falling cry mode. The spectrograms from the three normal infant cry signals taken from the DA-IICT Infant Cry database are shown in Figure 2.10. The resolution of the ConvRBM spectrograms is higher than the mel spectrograms as shown in marked regions in Figure 2.10. The harmonics are clearly resolved in the ConvRBM spectrograms. The cry modes, such as series of rising, falling and flat can be observed in Figure 2.10(d) and (g). The dysphonation cry mode is observed in Figure 2.10(e), (h) with harmonic vibration mode (shown by a circle). Our observations for the normal cry signals are similar to as observed in [6, 14] for the normal infant cry signals.

Asphyxia infant cry signal

The asphyxia or HIE is a disease caused to the newborn due to the lack of supply of oxygen or blood to the brain that arises due to abnormal breathing. In very serious conditions, asphyxia can cause coma or even death. The infants suffering from asphyxia are not able to produce a normal cry that resulted in pathological signs in the cry signals. The spectrograms from the three asphyxia infant cry signals taken from the Baby Chillanto database are shown in Figure 2.11. The time-frequency resolution is significantly better compared to the mel spectrograms as can be seen from Figure 2.11. The difference between normal and asphyxia cry can clearly visible from the spectrograms. There are no continuous harmonic

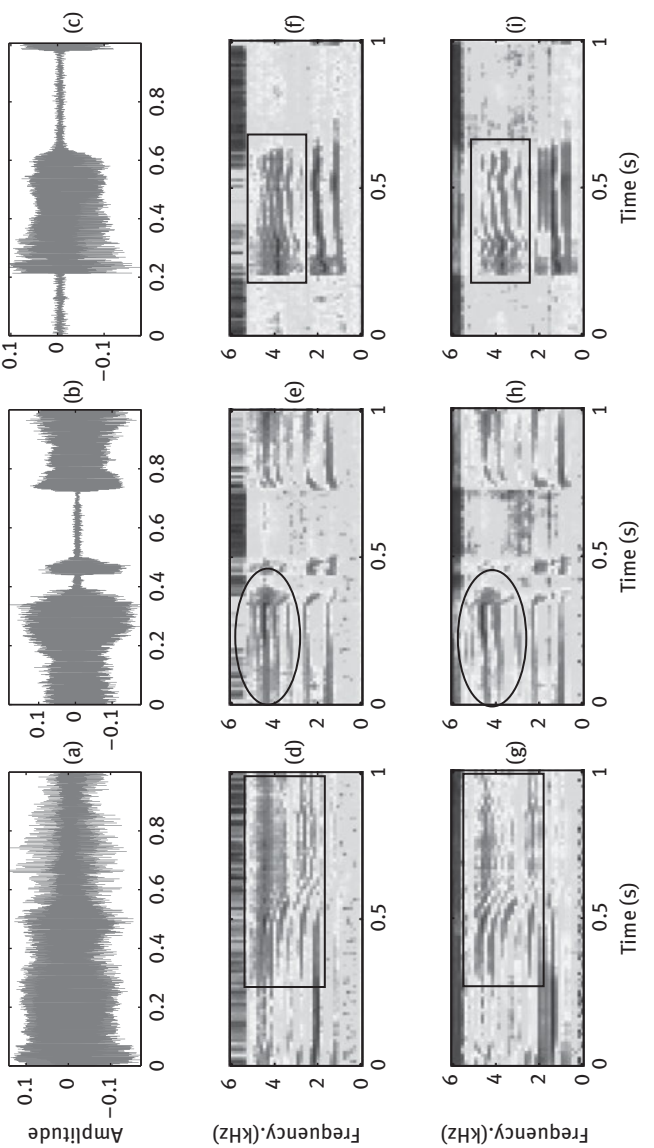

Figure 2.9: Comparison of spectrograms for normal cry signals from the Baby Chillanto database: (a)–(c) time-domain signals, (d)–(f) mel spectrograms and (g)–(i) ConvRBM spectrograms. The rectangular and circular regions indicate the differences in two spectrograms.

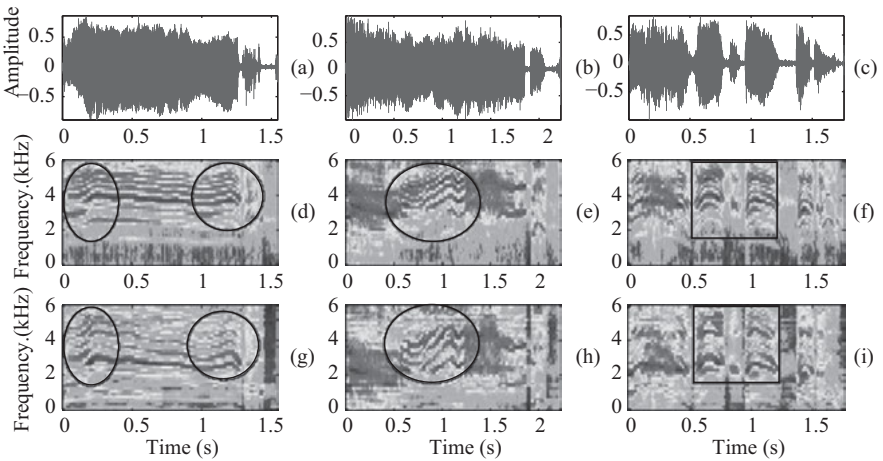

Figure 2.10: Comparison of spectrograms for normal cry signals from the DA-IICT Infant Cry database: (a)–(c) time-domain signals, (d)–(f) mel spectrograms, and (g)–(i) ConvRBM spectrograms. The rectangular and circular regions indicate the differences in two spectrograms.

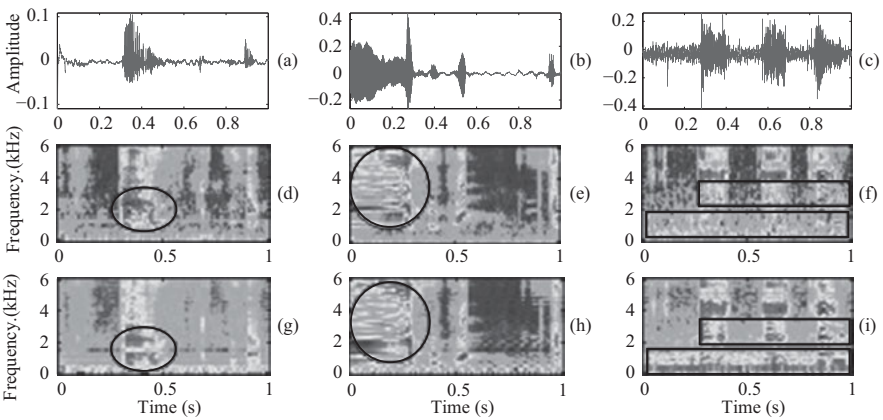

Figure 2.11: Comparison of spectrograms for asphyxia cry signals from the Baby Chillanto database: (a)–(c) time-domain signals, (d)–(f) mel spectrograms and (g)–(i) ConvRBM spectrograms. The rectangular and circular regions indicate the differences in two spectrograms.

structures present in the asphyxia cry, rather, it is of very short duration and noisy. This is due to lack of oxygen that infant is not able to vocalize due to an inadequate supply of oxygen or blood to his/her brain. Many of the cry modes related to harmonics are absent in asphyxia cry. The blurred harmonics can be

seen from asphyxia cry signals in Figure 2.11(a)–(c). ConvRBM spectrogram can show continuous dysphonation cry mode for one of the asphyxia cry signals in Figure 2.11(i) which cannot be revealed by the mel spectrogram in Figure 2.11(f). Similar observations are made from the asphyxia cry signals taken from the DA-IICT Infant Cry database as shown in Figure 2.12. The continuous dysphonation cry mode is present in the asphyxia cry signals shown in Figure 2.12(g). One can see that the mel spectrograms can not able to resolve leading and trailing harmonics on both sides of dysphonation cry mode. The asphyxia cry signals from the DA-IICT Infant Cry database also show very less spectral energies or dysphonation cry mode in the spectrograms.

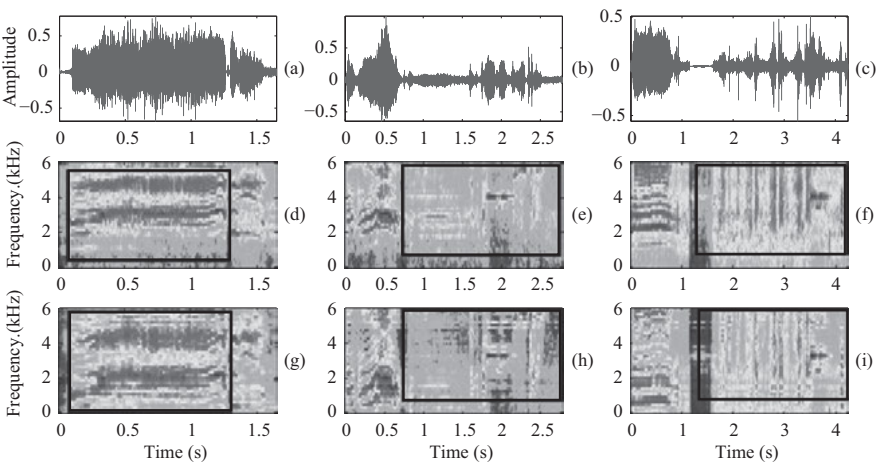

Figure 2.12: Comparison of spectrograms for asphyxia cry signals from the DA-IICT Infant Cry database: (a)–(c) time-domain signals, (d)–(f) mel spectrograms and (g)–(i) ConvRBM spectrograms. The rectangular regions indicate the differences in two spectrograms.

Deaf infant cry

There are several reasons for deafness in newborns or they become deaf early in life. It is not always possible to identify the reason for such cases, however, they are two possible cases, namely, prenatal causes and postnatal causes [37]. Prenatal cases include genetic reasons, complications during pregnancy, illnesses, such as rubella, cytomegalovirus (CMV), toxoplasmosis and herpes can cause newborns deaf [37]. Postnatal causes include infection, specifically in prematurely born babies and exposure to loud noise [37]. The deaf infant's cry signals differ from the normal infant cry signals. The onset of crying or canonical

babbling is delayed in deaf infants and cry signals differ in duration and timing [38]. Moreover, vocal cry inventories are very limited in deaf infants. The deaf infants rely on only sounds that are visually prominent, such as /ba/ and /ma/. This has a significant impact on the acquisition of language where sound perception plays a critical role [22]. Hence, early detection of deafness in infancy may help in reducing or providing a hearing aid may benefit for the better development of infants. The spectrograms from the three deaf infant cry signals from the baby Chillanto database are shown in Figure 2.13. One can see more resolved harmonics in the high frequency regions in ConvRBM spectrograms (as marked in Figure 2.13) compared to the mel spectrograms. In all the deaf cry samples, dysphonation cry mode is present in the high frequency regions. There are vibration cry modes also present as seen in the cry signals in Figure 2.13(a) and (b).

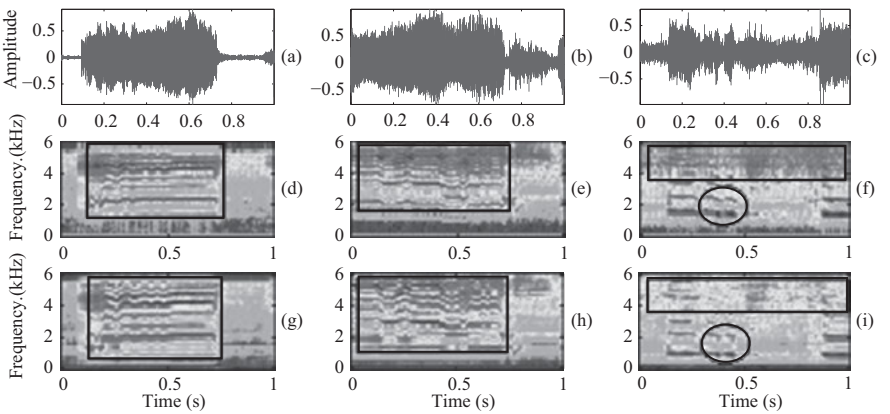

Figure 2.13: Comparison of spectrograms for deaf cry signals from the Baby Chillanto database: (a)–(c) time-domain signals, (d)–(f) mel spectrograms and (g)–(i) ConvRBM spectrograms. The rectangular and circular regions indicate the differences in two spectrograms.

Asthma cry

Asthma is a chronic inflammatory disease that inflames and narrows the airways. These airways allow air to come in and out of the lungs. Asthma causes recurring periods of wheezing (a whistling sound when you breath), shortness of breath (i.e., difficulty in breathing), chest tightness and coughing. The symptoms of asthma seen in people of all ages, but it most often starts during childhood or in the infant stage. Asthma is thought to be caused by a combination of

genetic and environmental factors that include allergens or air pollution. There is no cure for asthma till now; however, early symptoms can be prevented by avoiding triggers such as allergens and irritants, etc. An infant suffering from asthma face difficulties in breathing, and hence proper treatment must be conducted to reduce the symptoms. The spectrograms from the three deaf infant cry signals from the DA-IICT Infant Cry database are shown in Figure 2.14. Due to frequency inhalation, distorted harmonic structures are seen in the spectrograms in Figure 2.14(d) and (g). Abrupt dysphonation cry modes are present in Figure 2.14(e) and (h). Due to breathing difficulty, sometimes acoustic energy levels, and harmonic frequency range changes abruptly Figure 2.14(f) and (i).

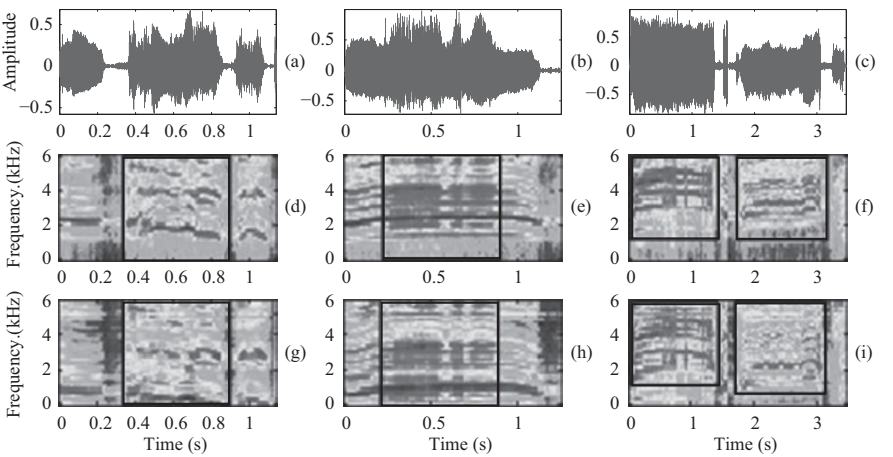

Figure 2.14: Comparison of spectrograms for asthma cry signals from the DA-IICT Infant Cry database: (a)–(c) time-domain signals, (d)–(f) mel spectrograms and (g)–(i) ConvRBM spectrograms. The rectangular regions indicate the differences in two spectrograms.

2.7 Experimental results

In this section, the classification results and evaluation using various performance measures are presented.

2.7.1 Experimental results on the DA-IICT Infant Cry database

The classification accuracies for the DA-IICT Infant Cry database using the MFCC and ConvRBM-CC feature sets are shown in Figure 2.15. The ConvRBM-CC

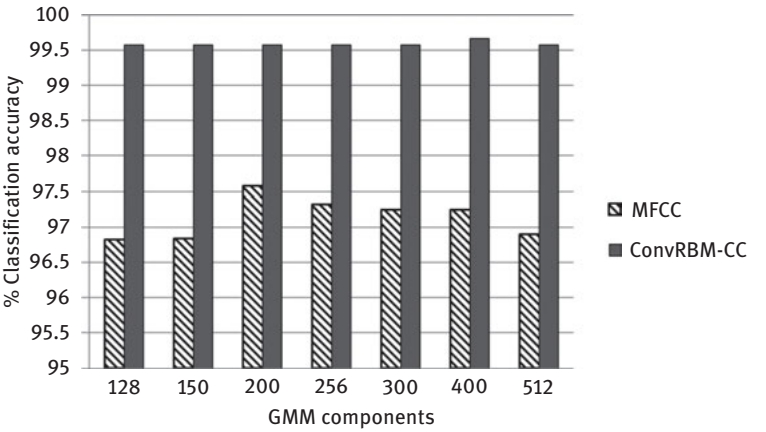

Figure 2.15: The % classification accuracies using for various GMM components the DA-IICT Infant Cry database.

obtained higher % classification accuracy compared to the MFCC for all the GMM components. For the MFCC, optimal results obtained using 200 GMM components. For the ConvRBM-CC, the optimal results obtained using 400 GMM components. We achieved an absolute improvement of 2% in the classification accuracy compared to the MFCC feature set. The confusion matrices for the classification experiment are shown in Figure 2.16. The FP and FN rate of the MFCC is quite high compared to the ConvRBM-CC feature set. From Figure 2.16(b), it can be seen that the ConvRBM-CC has no FP and only 4 FN compared to the MFCC with 21 FN (Figure 2.16(a)). Hence, with the ConvRBM-CC, there is no chance that a normal cry signal is considered as pathological cry signal.

	Normal	Pathology
Normal	791	8
Pathology	21	378

(a)

	Normal	Pathology
Normal	799	0
Pathology	4	395

(b)

Figure 2.16: Confusion matrices for experiments on the DA-IICT Infant Cry database using: (a) MFCC, and (b) ConvRBM-CC.

The performance measures of the classification experiments on the DA-IICT Infant Cry database are shown in Table 2.3. The ConvRBM-CC obtains significantly high values for all the measures compared to the MFCC. Since *F*-measure does

not consider the true negatives into account, the values of *F*-measure are very similar for both the feature sets. The MCC and *J*-statistic values are higher for the ConvRBM-CC compared to the MFCC. The % accuracy does not consider false positive and false negatives. From Table 2.3, one can see that the difference in MCC and *J*-statistic for the MFCC and ConvRBM-CC is higher compared to % accuracy. Hence, MCC and *J*-statistic are more meaningful performance measures than just % classification accuracy. We have also shown the 3-D t-SNE visualization in Figure 2.17 for the MFCC and ConvRBM-CC feature set extracted from the DA-IICT Infant Cry database. The MFCC feature vectors of both normal and pathology classes are scattered over all the dimensions in the tSNE plot. However, the tSNE plot of ConvRBM shows that features related to normal and pathology are grouped in separate clusters with very small overlap in them. Hence, ConvRBM-CC discriminates normal and pathology classes in the DA-IICT Infant Cry database more significantly than the MFCC.

Table 2.3: Performance measures for the classification experiments on the DA-IICT Infant Cry database.

Feature set	MCC	*F*-measure	*J*-statistic
MFCC	0.945	0.963	0.937
ConvRBM-CC	**0.993**	**0.995**	**0.99**

2.7.2 Experimental results on the Baby Chillanto database

The experimental results using the Baby Chillanto Database are shown in Figure 2.18 for the MFCC and ConvRBM-CC with different GMM mixture components. Compared to the DA-IICT Infant Cry database, both the feature sets were able to perform well in the classification of normal and pathology cry signals. However, ConvRBM-CC consistently performs better than the MFCC for all the GMM mixture components. The best classification accuracy of 99.87% was achieved using ConvRBM-CC (0.58% absolute improvement compared to the MFCC) obtained with 300 GMM mixture components. The confusion matrices for both feature sets are shown in Figure 2.19. The false positive rate of the MFCC is quite high than the ConvRBM-CC (15 *vs*. 1), while there are no false negative when the ConvRBM-CC is used in the classification task. Hence, with the ConvRBM-CC feature set, all the cry samples are correctly classified with only one false negative.

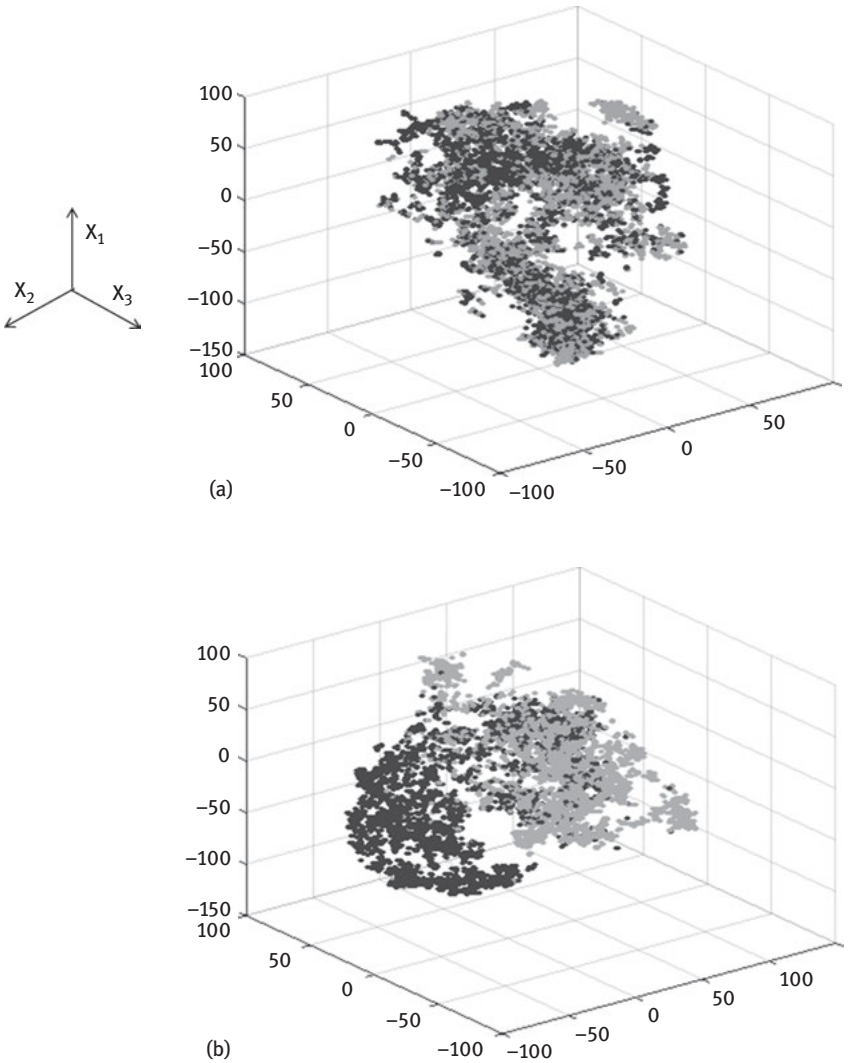

Figure 2.17: t-SNE plots for 3-D visualizations of (a) MFCC, and (b) ConvRBM-CC feature sets extracted from the DA-IICT Infant Cry database. The 39-D of both feature sets is mapped to 3-D using the tSNE technique. The black color corresponds to normal and gray color corresponds to pathology class.

The significance of this improvement using the ConvRBM-CC feature set can also be seen from the performance measures in Table 2.4. Here, again the *F*-measure is similar for both ConvRBM-CC and the MFCC. The MCC and *J*-statistic

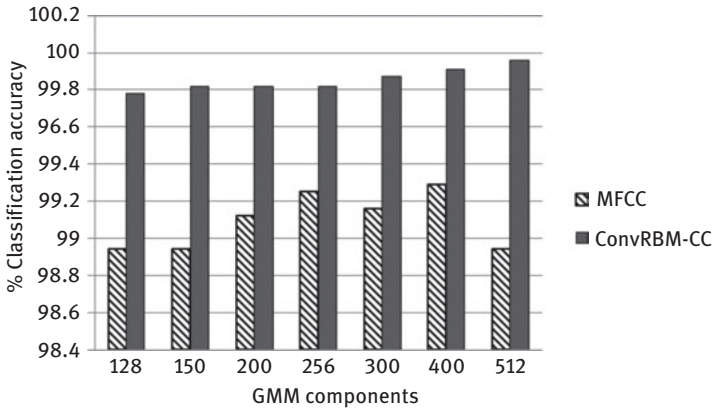

Figure 2.18: The % classification accuracies using for various GMM components the Baby Chillanto database.

	Normal	Pathology
Normal	1034	15
Pathology	1	1218

(a)

	Normal	Pathology
Normal	1048	1
Pathology	0	1219

(b)

Figure 2.19: Confusion matrices for experiments on the Baby Chillanto database using: (a) MFCC and (b) ConvRBM-CC.

Table 2.4: Performance measures for the classification experiments on the baby Chillanto database.

Feature set	MCC	*F*-measure	*J*-statistic
MFCC	0.986	0.994	0.985
ConvRBM-CC	**0.999**	**0.999**	**0.999**

are quite high for the ConvRBM-CC with value 0.999 (close to 1). The difference in values of MCC and *J*-statistic indicates that the ConvRBM-CC performs better than the MFCC even though % accuracy is quite similar. The tSNE visualization of the MFCC and ConvRBM-CC feature sets are shown in Figure 2.20. Compared to the DA-IICT Infant Cry database, the MFCC features of the Baby Chillanto database show more class separability in the tSNE 3-D feature space as shown in Figure 2.20(a). ConvRBM-CC feature set also shows separate clusters of feature

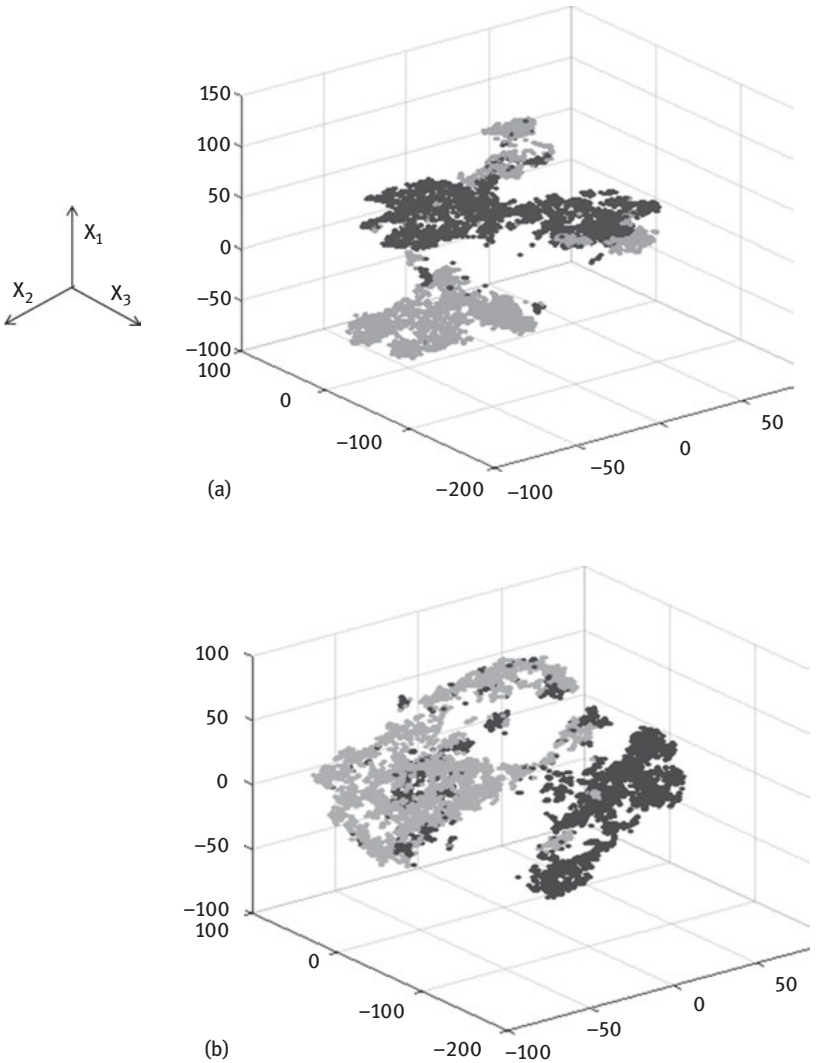

Figure 2.20: t-SNE plots for 3-D visualizations of (a) MFCC and (b) ConvRBM-CC feature sets extracted from the baby Chillanto database. The 39-D of both feature sets is mapped to 3-D using the tSNE technique. The black color corresponds to normal and gray color corresponds to pathology class.

vectors of the normal and pathological infant cry signals in Figure 2.20(b). The class-specific grouping of feature vectors is more dominant in the ConvRBM-CC compared to the MFCC.

2.8 Discussions

Most of the pathology in the infants is due to neurological diseases (including SIDS) or changes in neurological responses due to physical abnormalities, such as asthma and asphyxia. In one of the studies, it was observed that babies who die of SIDS have abnormalities in the specific brain region called *medulla oblongata* which helps in control functions like breathing, blood pressure and abnormalities in serotonin signaling [4]. The ConvRBM is a statistical auditory model incorporating responses to the auditory nerve fibers (ANF). Hence, the learned filterbank reflects the perceptual cues regarding pathology in the infant cry signals. The spectrogram analysis of the filterbank revealed that good time-frequency resolution is required to clearly see different cry modes in the infant cry signals. Though the mel filterbank is based on perceptual mel scale, the frequency resolution in progressively higher frequency region is poor. The frequency scale obtained from the ConvRBM is learned through the infant cry signals that are optimal to represent the infant cry signals in time and frequency domain. Hence, we can analyze different cry modes clearly in the ConvRBM spectrograms. Further extension of this work is to classify the type of pathology class. However, there are potential challenges in doing so, such as very limited cry samples for each pathology case and highly imbalance classes (e.g., large samples of normal cry *vs*. few samples of pathology classes). Such difficulty may be overcome by using data augmentation technique in the infant cry signals.

2.9 Summary and conclusions

In this chapter, we proposed to use ConvRBM-based auditory filterbank learning for the infant cry classification task. The subband filters learned from the two infant cry databases shows that most of the learned subband filters are the Fourier-like basis. This is due to the fact that infant cry signals contain greater harmonic structures. The filterbank scale also follows the standard auditory frequency scales. The analysis of the ConvRBM spectrograms shows various cry modes more clearly present in the ConvRBM spectrograms compared to the mel spectrograms. The classification experiments for the normal versus three pathologies, namely, deaf, asthma and asphyxia cry signals are presented. The experimental results using standard performance measures show that the proposed ConvRBM-based features perform significantly well in the infant cry classification task. The current limitation of the present work is the use of imbalanced classes and justification of the proposed approach through only two databases. We would like to obtain

more statistically meaningful infant cry databases to further signify our proposed approach. The future work also includes to classify pathological cry signals and make better infant cry classier that will be helpful to the doctors and the society.

References

[1] O. Aragn. Why do we cry?, Scientific American Mind, 28(2), 74, April 2017.
[2] A. Vingerhoets. Why only humans weep: Unravelling the mysteries of tears, Oxford University Press, First Edition, 2013.
[3] E. Gustafsson, F. Levrro, D. Reby, and N. Mathevon. Fathers are just as good as mothers at recognizing the cries of their baby, Nature Communications, 4(1698), 1–6, 2013.
[4] H.A. Patil. Cry baby: Using spectrographic analysis to assess neonatal health status from an infant's cry, Advances in Speech Recognition Mobile Environments, Call Centers and Clinics, A. Neustein, Ed, Springer, 323–348, 2010.
[5] O. Wasz-Hckert. et. al. Twenty five years of Scandinavian cry research, Infant Crying: Theoretical and Research Perspectives, C. Boukydis and B. Lester, Eds, Springer, 83–104, 1985.
[6] N. Buddha, and H.A. Patil. "Corpora for analysis of infant cry," in International Conference on Speech Databases and Assessments, Oriental COCOSDA, Hanoi, Vietnam, Dec. 2007, 43–48.
[7] A. Chittora, and H.A. Patil. Data collection of infant cries for research and analysis, Journal of Voice, Elsevier, 31(2), 252.e15–252.e26, 2017.
[8] A. Rosales-Perez, C.A. Reyes-Garcia, J.A. Gonzalez, O.F. Reyes-Galaviz, H.J. Escalante, and S. Orlandi. Classifying infant cry patterns by the genetic selection of a fuzzy model, Biomedical Signal Processing and Control, 17(1), 38–46, 2015.
[9] R. Prescott. Infant cry sound: developmental features, The Journal of the Acoustical Society of America (JASA), 57(5), 1186–1191, 1975.
[10] T. Etz, H. Reetz, C. Wegener, and F. Bahlmann. Infant cry reliability: Acoustic homogeneity of spontaneous cries and pain-induced cries, Speech Communication, 58(1), 91–100, 2014.
[11] H.F. Alaie, L. Abou-Abbas, and C. Tadj. Cry-based infant pathology classification using GMMs, Speech Communication, 77(1), 28–52, 2016.
[12] S. Orlandi, C.A.R. Garcia, A. Bandini, G. Donzelli, and C. Manfredi. Application of pattern recognition techniques to the classification of full-term and preterm infant cry, Journal of Voice, Elsevier, 30(6), 656–663, 2016.
[13] L. Abou-Abbas, C. Tadj, C. Gargour, and L. Montazeri. Expiratory and inspiratory cries detection using different signals' decomposition techniques, Journal of Voice, Elsevier, 31(2), 259.e13–259.e28, 2017.
[14] A. Chittrora. "Crying for a reason: A signal processing based approach for infant cry analysis and classification," Ph.D. Thesis, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India, June 2016.
[15] H.B. Sailor, and H.A. Patil. "Filterbank learning using convolutional restricted Boltzmann machine for speech recognition," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016, 5895–5899.

[16] H.B. Sailor, and H.A. Patil. Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition, IEEE/ACM Transactions on Audio, Speech and Language Processing, 24(12), 2341–2353, Dec 2016.
[17] M.J. Corwin, B.M. Lester, C. Sepkoski, M. Peucker, H. Kayne, and H.L. Golub. Newborn acoustic cry characteristics of infants subsequently dying of sudden infant death syndrome, Pediatrics, 96(1), 73–77, 1995.
[18] M.P. Robb, D.H. Crowell, and P. Dunn-Rankin. Sudden infant death syndrome: Cry characteristics, International Journal of Pediatric Otorhinolaryngology, Elsevier, 77(8), 1263–1267, 2013.
[19] R.H. Colton, and A. Steinschneider. The cry characteristics of an infant who died of the sudden infant death syndrome, Journal of Speech and Hearing Disorders, 46(4), 359–363, 1981.
[20] U. D. of Health and H. Services, "Sudden infant death syndrome (SIDS)," URL: https://www.nichd.nih.gov/health/topics/sids/Pages/default.aspx, {Last Accessed on 7 Oct., 2017}.
[21] G. Hinton. Where do features come from?, Cognitive Science, 38(6), 1078–1101, 2014.
[22] P.K. Kuhl. Early language acquisition: cracking the speech code, Nature Reviews Neuroscience, 5(11), 831–843, Nov 2004.
[23] H.B. Sailor, and H.A. Patil. Auditory feature representation using convolutional restricted Boltzmann machine and Teager energy operator for speech recognition, Journal of Acoustical Society of America Express Letters (JASA-EL), 141(6), EL500–EL506, June 2017.
[24] H.B. Sailor, D.M. Agrawal, and H.A. Patil. Unsupervised filterbank learning using convolutional restricted Boltzmann machine for environmental sound classification, Interspeech, Stockholm, Sweden, 3107–3111, 2017.
[25] H.B. Sailor, M.R. Kamble, and H.A. Patil. Unsupervised representation learning using convolutional restricted Boltzmann machine for spoof speech detection, Interspeech, Stockholm, Sweden, 2601–2605, 2017.
[26] Hardik B. Sailor, Madhu Kamble and Hemant Patil, "Auditory Filterbank Learning for Temporal Modulation Features in Replay Spoof Speech Detection", in INTERSPEECH 2018, Hyderabad, India, September 2018, pp. 666–670.
[27] G.E. Hinton. A practical guide to training restricted Boltzmann machines, Neural Networks: Tricks of the Trade, Springer, 599–619, 2012.
[28] S.J. Rennie, V. Goel, and S. Thomas. Annealed dropout training of deep networks, IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, California and Nevada, 159–164, 2014.
[29] A. Fischer, and C. Igel. An introduction to restricted Boltzmann machines, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer, 14–36, 2012.
[30] D. Kingma, and J. Ba. 2015, "Adam: A method for stochastic optimization," in International Conference on Learning Representations (ICLR), San Diego, 1–11.
[31] R.M. Stern, and N. Morgan. Features based on auditory physiology and perception, Techniques for Noise Robustness in Automatic Speech Recognition, T. Virtanen, B. Raj, and R. Singh, Eds., John Wiley and Sons, Ltd, New York, NY, USA, 193–227, 2012.
[32] X. Yang, K. Wang, and S. Shamma. Auditory representations of acoustic signals, IEEE Transactions on Information Theory, 38(2), 824–839, March 1992.
[33] T. Fawcett. An introduction to ROC analysis, Pattern Recognition Letters, Elsevier, 27(8), 861–874, 2006.

[34] W.J. Youden. Index for rating diagnostic tests, Cancer, Wiley Subscription Services, Inc., A Wiley Company, 3(1), 32–35, 1950.
[35] B. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta (BBA) – Protein Structure, Elsevier, 405(2), 442–451, 1975.
[36] L. van der Maaten, and G.E. Hinton. Visualizing high-dimensional data using t-SNE, Journal of Machine Learning Research (JMLR), 9, 2579–2605, 2008.
[37] N. D. C. Society, "Causes of deafness," URL: http://www.deafchildworldwide.info, {Last Accessed on 20 August, 2017}.
[38] P.K. Kuhl, and A.N. Meltzoff. Infant vocalizations in response to speech: Vocal imitation and developmental change, Journal of Acoustical Society of America (JASA), 100, 2425–2438, Oct 1996.

Stefany Bedoya, Nirit Brosh Katz, Jessica Brian,
Douglas O'Shaughnessy and Tiago H Falk

3 Acoustic and prosodic analysis of vocalizations of 18-month-old toddlers with autism spectrum disorder

Abstract: Autism spectrum disorder (ASD) covers a wide spectrum of symptoms with the main ones relating to problems with social communication and interaction. Definite ASD diagnosis is based on the presence of certain symptoms and their severity levels and, according to current standards, occurs typically at 48 months of age. Recent statistics show that about 1 in 68 children are diagnosed with autism and there is a recurrence rate of 18.7% for the biological siblings of individuals with ASD. As such, early detection is critical, as it may allow for intense therapy to be initiated, thus tapping into a young brain's plasticity properties and increasing odds of success. Today, researchers and clinicians have joined efforts to understand and identify new markers of the disorder, thus allowing for early diagnosis, ideally around 18 months of age. To this end, acoustic analysis of toddler vocalizations has emerged as a promising area, even for preverbal children. Prosodic and acoustic disorders have been reported for babble and speech-like vocalizations. As such, pitch, energy and voice quality related features have been explored for early ASD diagnosis. In this work, we build upon these findings and propose the use of wavelet-based and speech modulation spectral features for ASD diagnosis based not only on speech-like verbalizations, but also on cries, laughs and other sounds made by the toddlers. We show that the proposed features are complementary to existing ones and, on a cohort of 43 18-month-old toddlers, a support vector machine classifier was capable of correctly discriminating the ASD group from the typically developing toddlers with accuracies above 80%, thus outperforming existing methods. More importantly, we show that with these new features, vocalizations such as cries, squeals, whines and shouts showed to

Note: Stefany Bedoya is now at Nexalogy, Montreal, Canada. The work was done while the author was a graduate student at INRS-EMT, Montreal, Canada. The work does not contain any Nexalogy proprietary information.

Stefany Bedoya, Nexalogy, Montreal, Canada
Nirit Brosh Katz, Afula Child Development Center, Clalit Health Services, Israel
Jessica Brian, Holland-Bloorview Kids Rehab, Toronto, Canada
Douglas O'Shaughnessy, Tiago H Falk, INRS-EMT, University of Quebec, Montreal, Canada

https://doi.org/10.1515/9781501513138-003

be more discriminative than babble and speech-like vocalizations. It is hoped that these findings will lead to more accurate early diagnosis of ASD symptoms.

Keywords: autism spectrum disorder, diagnosis, prosody, wavelets, speech modulation spectrum

3.1 Introduction

The American Psychiatric Association defines autism as a pervasive developmental disorder that is related to a triad of impairments: (1) atypical development in reciprocal social interaction; (2) atypical communication; and (3) restricted, stereotyped and repetitive behaviors [1]. In fact, the definition has recently been updated to include a wide spectrum of symptoms and impairment levels, thus the terminology autism spectrum disorder (ASD) has been incorporated [1]. The definite (or stable) diagnosis of ASD is based on the presence of certain symptoms and their severity levels, and typically occurs around 48 months of age [2]. Recent statistics suggest that roughly 1 in 68 children are diagnosed with autism and there is a recurrence rate of 18.7% for the biological siblings of children with ASD [3, 4].

Recent research, however, has suggested that diagnosis can be made around 18 months in many cases [5]. Early diagnosis can allow parents to move forward with appropriate educational support [6] and clinicians to initiate interventions that take advantage of the young brain's plasticity properties [7]. New tools are needed, however, in order to accurately diagnose ASD at such an early age. The Autism Diagnostic Observation Schedule (ADOS) is one of the gold-standard assessment tools used in the diagnosis of ASD. In the most recent version of the ADOS [8], the item that captures atypical intonation, pitch, stress, tone, volume, rhythm and rate of vocalizations was added to the diagnostic algorithm, due it demonstrated clinical utility [8–11].

These studies have shown that vocalizations of participants with ASD are more difficult to interpret in terms of affective meaning and function than in their typically developing (TD) and developmentally delayed (DD) peers [12, 13]. Moreover, individuals on the autism spectrum have been observed to have more hoarse, harsh, hyper-nasal voice quality, with higher recurrence of squeals, growls and yells [14]. Several terms have been used to described prosody in ASD including: monotonous, exaggerated, robotic, pedantic and wooden, to name a few. Whatever the unusual feature might be, it is often noticed by social contact and may represent a significant barrier to peer acceptance. Atypicalities in the prosody of individuals with ASD were noted in the earliest description of autism, and studies show that impairments in vocal behavior are evident even in preverbal individuals with ASD and may represent an early feature of ASD [12, 15–18].

Early studies have examined both perception and production of prosody in individuals with autism; however, an issue often addressed in the literature of ASD is the unknown prevalence of atypical prosody and the heterogeneity of the ASD population. Additionally, the lack of standardized investigative methods used to quantify vocal prosody and the shortcomings of perceptual judgment have produced inconsistent and sometimes contradictory findings [14, 19, 20]. As such, recent research has focused on incorporating computerized acoustic analysis within the detection of ASD, thus overcoming some of these limitations involved in research studies using qualitative methods [9–11]. Acoustic characteristics of speech production have been measured and quantified across several studies as possible markers of ASD. Most existing studies, however, have been conducted with older children or with younger children but with a wide age range between 18 and 36 months. Given the natural changes in vocal characteristics (e.g., growing and maturation of the vocal tract) that occur during childhood, it is important that acoustic analysis focus on individuals of roughly the same age, thus removing the potential bias from natural age-related acoustic changes and variability, and allowing the system to focus solely on disorder related differences. The work described herein fills this gap.

3.2 Prosody in ASD

Disordered prosody has long been considered a hallmark of ASD and atypical intonation has been recognized as an important diagnostic ASD marker. Prosodic characterization of children with ASD is an under-researched area, particularly for very young and preverbal children, although studies suggest vocal atypicality may represent an early appearing symptom of ASD. Studies have traditionally measured and quantified parameters of intonation, volume and vocal quality to identify differences between ASD and comparison groups. The most persisting finding reported in the literature is that individuals with ASD have greater variability and a wider range in fundamental frequency (F0) measures [9, 14, 18, 21–26]. This finding is to an extent counter-intuitive as children on the autism spectrum are often described as having monotone or robotic speech. Few studies have also suggested that impairments in the use of pitch related features in individuals with ASD are linked to an abnormal processing of auditory stimuli at the brainstem level [27–29]. More recently, a higher pitch coefficient of variation (CV) per word was found for the TD group in comparison with the ASD group at school age, while no significant differences between the two classes could be observed at a preschool age [11].

In addition to disordered pitch, ASD individuals have been reported to have speech that is described as “too fast” or “too slow” [14]. Most studies that have

measured sentence and word duration have reported longer duration in individuals with ASD [10, 14, 30–32]. In [10] for example, a perceptual and acoustic analysis in children with Asperger syndrome was employed. The study showed that even when children with ASD and TD performed similarly in the perception of intonation patterns and the use of prosodic features to express grammatical meaning, children with ASD showed an alteration in duration, mean and maximum pitch. These findings are in line with the results shown in [32], where a significant difference in duration for emotional speech (sad utterances having a longer duration than happy or angry utterances) was found for TD and Asperger Syndrome (AS) groups, while participants with high-functioning autism (HFA) did not show any difference. However, two recent studies did not find significant differences in duration between ASD and TD children [21, 33]. Again, such contradictory results may be due to a wide age-range of the study participants, thus results may be biased due to natural age-related acoustic changes and variability.

Moreover, a few studies have employed multivariate analysis using machine learning techniques. In [34], for example, four groups of features: voice quality, energy-related, spectral and cepstral features were compared for a database collected to assess child abilities in imitation of different types of prosody contours. The results showed that voice quality features improved classification performance over the other feature groups. Other studies measured voice quality related measures such as jitter, harmonic-to-noise ratio (HNR) and cepstral peak prominence (CPP), which tend to increase in the children with ASD with increasing symptom severity [9, 24, 25, 35].

3.3 Early vocal patterns in ASD

Parents of children with ASD may have a difficult time recognizing the affective meaning of their infants' vocalizations, and recent evidence is emerging to indicate that vocal atypicality may be apparent in very young infants with ASD [12, 13, 16, 36, 37]. However, most of the studies have tended to focus on verbal adolescents and adults, though some have studied school and preschool age children as well. Infant vocalizations are the earliest form of vocal communication. They play an important role in the development of the parent-child relationship and language acquisition. Infants begin to produce vegetative or reflexive sounds such as coughing or crying soon after birth, and through several stages they expand their sounds and their vocalizations to become more speech-like [38]. The growth and anatomic restructuring of the vocal tract during the first half year of life and the vocal learning are the main factors that induce a change in child vocalizations [38]. Recently, a few studies have been analyzing the development of acoustic

parameters in infants' vocalizations as a noninvasive tool to measure vocal-muscular maturation [12, 16, 39, 40]. Atypicalities in the production of vocal patterns could involve abnormal processing of auditory feedback or problems in the speech production mechanisms. In the next subsections we explore some studies employed in children with ASD. Specifically crying and canonical babbling vocalizations have been studied in the ASD literature.

3.3.1 Crying

Crying is the infant's earliest form of vocal communication. It has been explored due to its relationship with the central nervous system [12, 16, 17, 37, 41–45]. Acoustic abnormalities in infants' cries have been associated with some disorder such as asphyxiation, low birth weight, metabolic disorder, neurological symptoms and lead exposure, among others. Most of the studies evidence a high and variable pitch [41]. In the ASD literature, the cries of children later diagnosed with autism exhibited a higher F_0 value and shorter pauses than the cries of developmentally delayed or typically developing children [12, 16, 36, 37]. Moreover, the fundamental frequency was shown to decrease for healthy children during the first and second years of life, unlike the case with children later diagnosed with ASD [16, 36]. Sheinkopf and colleagues employed a cry acoustic analysis of 6-month-old infants [46] and showed that at 6 months, high risk children started to show a higher and more variable pitch value than those with low risk. Later, when the same high risk participants were analyzed at the age of 36 months, they showed an even higher F0. Additionally, other studies have analyzed the variability in the fundamental frequency (pitch range) but no differences were found between the two groups [37, 46].

A more recent study in ASD employed a reaction time (RT) categorical task to analyze adults' responses to cries of children between 36 and 40 months of age with ASD [16]. They found that differences in vocal behavior in children later diagnosed with ASD caused adults to perceive the cries as distressed and more difficult to process and interpret than the cries of typically developing children, as well as mammalian animal cries and environmental noise control sounds. Esposito and colleagues, in turn, examined acoustic features of infant-cry vocalizations of a group of typically developing children and children with ASD, both aged 13 months [12], and found that the pause length was more important for the perception of distress in children with ASD than the fundamental frequency or the frequency of cry sounds per unit time across an episode of crying.

3.3.2 Canonical babbling

Babbling begins shortly after birth and is well established by 10 months of age. Any delay in the onset of canonical babbling is related to language delay or other developmental disabilities [17]. Just a few studies have focused on the analysis of canonical babbling in children with ASD [17, 43, 44]. Patten and colleagues studied the canonical status and vocalization frequency of a group of 37 infants with ASD compared with a typical development group at 9–12 and 15–18 months [17]. Individuals later diagnosed with autism produced significantly lower rates and had a later onset of canonical babbling than the control group. These findings are congruent with the results obtained in the analysis of canonical syllable production of infants aged between 16 and 48 months [44]. However, some reported findings have been contradictory. For example, Sheinkopf and colleagues studied the nature of early vocal behaviors in young children with autism with a focus on canonical babbling and atypical vocal quality (defined as the rate of production of atypical phonation) [43] and the group with ASD did not display significant differences compared with developmentally delayed control group in terms of rate of canonical babbling, despite their vocalizations showing atypical vocal quality [17, 44].

3.4 Methods and materials

3.4.1 Data collection

Data used in this study were extracted from a set of videotaped ADOS – Module 1 sessions, which are part of the longitudinal prospective Canadian "Infant Sibling Study" from the Autism Research Unit at Toronto's Sick Kids Hospital [47]. The study monitors younger siblings of probands with ASD, recruited due to a known higher risk to exhibit the disorder, estimated at an 18.7% recurrence rate [3], as well as low-risk age-matched "control" children of families without a history of ASD. Participants are followed from the age of 6–24 months and every 3–12 months undergo a series of (re)assessments, including the ADOS and other standardized developmental and language tests. At the 36-month follow-up visit, a well experienced clinician, blinded to previous outcomes and impressions, assesses them for ASD utilizing gold standard clinical tools, such as medical history, ADOS and the Autism Diagnostic Interview – Revised (ADI-R) [48].

Our analysis was conducted on a subset of the Infant Sibling data set and relied on audio recordings of 43 participants during their 18-month assessment. The

ASD group includes 23 (15 male and 8 female) children independently diagnosed with ASD at the age of 36 months and encompassed both Asperger syndrome and Autism disorder diagnoses. An age-matched comparison group of 20 (13 male and 7 female) low-risk TD children was used. Children in the control group received the same follow-up assessments as the ASD group and were determined at 36 months of age not to have ASD. As per the larger study, participants who were born premature or with low birth-weight are excluded from the study.

Audio content was extracted from the video recordings and toddlers' vocalizations were manually segmented and labelled according to vocalization type. Instances of vocalizations with overlapping adult speech (parents, clinician) were discarded from our analysis. Overall, the total audio segments extracted from *only* the toddler vocalizations resulted in 127.0 and 194.5 seconds for control and ASD groups, respectively.

3.4.2 Preprocessing

In the literature of infant phonology, a vocalization occurs on expiration and when an inspiration occurs, it is perceived as a break that separates the vocal events [44]. Each vocal event separated by a breath is called an utterance. In order to extract the utterances of each one of the vocalizations in our database and to avoid processing silent/noise-only intervals, an automated energy-thresholding method was used, as in [44]. Using this segmentation method, a total of 2,647 utterances were obtained for both ASD and non-ASD groups. Table 3.1 summarizes the number of vocalization utterances for each group and type of vocalization. The class labeled as "negative emotions" combines vocalization utterances such as cry, squeal, whine and shout. As shown, the number of vocalizations per group

Table 3.1: Summary of vocalization utterances for ASD and control groups.

Group	ASD (n = 23)	Control (n = 20)	Total
Babble	451	333	784
Speech	375	527	902
Laugh	49	14	63
Negative emotions	224	96	320
Others	378	200	578
Total	1477	1170	2647

and per vocalization type are not balanced with babble and speech being the most prominent and laugh the least.

3.4.3 Feature extraction

Wavelet packet decomposition

Wavelet packet decomposition (WPD) is a generalization method of the discrete wavelet transform (DWT) that allows a time-frequency multiresolution analysis of an input signal. WPD has been used in previous studies for emotion recognition, speech analysis and also has shown to be useful in the analysis of pathological speech and pathological infant cry [49–51]. Due to the highly nonstationary characteristics of some vocalizations such as cry, squeal and shout, the wavelet analysis was more suitable for detection and classification than the traditional Fourier methods [52]. The decomposition process and multiresolution analysis can be viewed as the application of a filter bank. More specifically, the input signal is passed through a low-pass and high-pass filter, which corresponds to the scaling and wavelet function [49, 53]. The lower frequency band gives the approximation coefficients and higher frequency band the detail coefficients. In wavelet packet decomposition, the process is recursively applied to both frequency subbands to generate the next level of decomposition. The wavelet packets coefficients can be computed as follows:

$$C_{n,k}^{P} = \sqrt{2^{P}} \sum_{m=-\infty}^{\infty} f(m) \cdot W_n(2^{P}m - k), \tag{3.1}$$

$$C_{2n,l}^{P-1} = \sum_{m} h(m-2l) \cdot C_{n,m}^{P}, \tag{3.2}$$

$$C_{2n+1,l}^{P-1} = \sum_{m} g(m-2l) \cdot C_{n,m}^{P}. \tag{3.3}$$

where P is the scale index, l the translation index, h low-pass filter and g high-pass filter, and K is filter length.

Figure 3.1 shows an example of a two-level wavelet packet decomposition for a babble signal generated from a control and an ASD participant. For n levels of decomposition, the WPD produces 2^n finer equal-width frequency subbands or nodes. The frequency ranges given in Hertz for the level n of decomposition are described by:

$$\left[\frac{kfs}{2^{n+1}} \frac{(k+1)fs}{2^{n+1}}\right] \quad k = 0, 1, ..2^{n-1}, \tag{3.4}$$

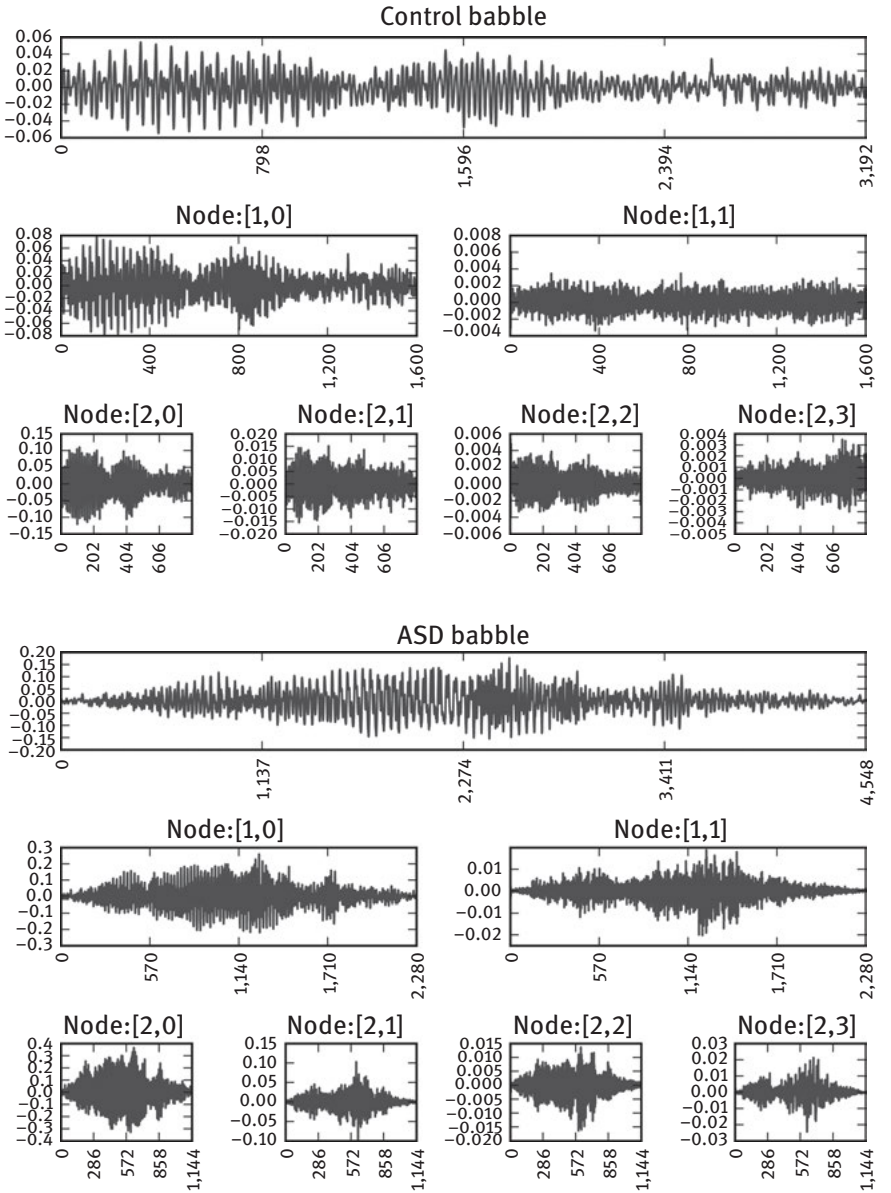

Figure 3.1: Two-level wavelet packet decomposition of (a) control and (b) ASD babble signals with bior2.6 mother wavelet.

where f_s is the sampling frequency of the original signal [50]. Table 3.2 shows the frequency ranges for each level of decomposition for a signal with a sampling frequency of 16 kHz, as is the case in the database used herein. These frequency ranges are used in our experiments to decompose vocalization utterances.

Table 3.2: Frequency bands for wavelet packet decomposition of a signal with sampling frequency of 16 kHz.

Decomposition level	Frequency band (Hz)
1	0–4,000, 4,000–8,000
2	0–2,000, 2,000–4,000, 4,000–8,000
3	0–1,000, 1,000–2,000, 2,000–4,000, 4,000–8,000
4	0–500, 500–1,000, 1,000–2,000, 2,000–4,000, 4,000–8,000

We propose the use of energy and entropy based features for our analysis, computed from each wavelet packet coefficient $C^P_{n,k}$ as follows [49–51], respectively:

$$E_{(n,k,l)} = \sum_{l=-\infty}^{+\infty} |C^P_{n,k}(m)|^2 w(l-m), \tag{3.5}$$

$$S_{(n,k,l)} = \sum_{l=-\infty}^{+\infty} -|C^P_{n,k}(m)|^2 \log |C^P_{n,k}(m)|^2 w(l-m), \tag{3.6}$$

where l is the number of window frames, P is the scale index, n represents the decomposition level and $k = 0, 1, ..., 2^{n-1}$ is the node number. In our experiments, each subband wavelet packet coefficient is divided into frames of 40 ms and successive frames were overlapped by 50%. Finally, statistical measures such as mean ($\bar{X}_{(n,k)}$), standard deviation ($std_{(n,k)}$), skewness ($g_{(n,k)}$) and kurtosis ($G_{(n,k)}$) are computed overall per-frame measures over the entire vocalization utterance. The final feature vector is an 8-dimensional feature vector computed per node k and decomposition level i, comprised of the mean, standard deviation, skewness, and kurtosis of the energy and the entropy values.

3.4.4 Speech modulation spectral representation

The so-called speech modulation spectral signal representation is an auditory-inspired spectro-temporal representation that captures both acoustic frequency and temporal modulation frequency properties of the analyzed signal

[54]. Figure 3.2 shows the steps used in our approach to compute the spectro-temporal (ST) representation of an input signal. In the first step, the active signal level is normalized to −26 dBov (dB overload) [55]. Next, a bank of 23 critical-band gammatone filters is used to model the frequency response of the basilar membrane [56]. The first filter is centered at 125 Hz and the last one at half of the sampling frequency of the analyzed signal. The bandwidth of each filter is described by a psychoacoustic measure called equivalent rectangular bandwidth (ERB) and is computed as follows:

$$ERB_i = \frac{f_i}{Q_{ear}} + B_{min}, \tag{3.7}$$

where f_i is the center frequency given in Hz of the ith critical-band filter, and Q_{ear} and B_{min} are constants set to 9.26449 and 24.7, respectively.

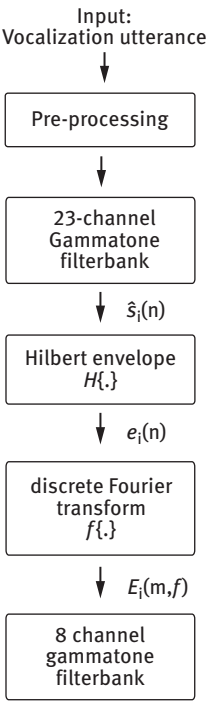

Figure 3.2: General scheme to compute ST representation.

The envelope $e_i(n)$ is computed for each one of the filterbank outputs $\hat{s}_i(n)$ using the Hilbert transform $H\{\cdot\}$. The envelope of the ith bandpass filter signal is given by:

$$e_i(n) = \sqrt{\hat{s}_i(n)^2 + H\{\hat{s}_i(n)\}^2} \qquad i = 1 \cdots, 23 \tag{3.8}$$

The modulation spectrum of the signal is computed using the discrete Fourier transform (DFT). Specifically, the envelope signal $e_i(n)$ is divided into frames of 256-ms every 40-ms using a Hamming window. The notation $e_i(m)$ is used to indicate the frame m of the windowed envelope. The next step is to take the DFT $F\{\cdot\}$ of each frame. Then, the modulation spectrum is defined by

$$E_i(m, f) = |F(e_i(m))|, \tag{3.9}$$

where f is the modulation frequency. Finally, an auditory-inspired modulation filter bank allows us to build a representation as a function of the acoustic frequency and temporal modulation frequency elements. The modulation energy of the ith critical-band signal is grouped into 8 bands; each band is denoted as $\varepsilon_{i,k}(m), k = 1, \ldots, 8$, where k notes the kth modulation filter.

Given the modulation spectral representation above, the set of features originally proposed in [54] is extracted. The first feature set, $\Phi_{1,m}(k)$, represents the energy distribution along the modulation frequency. It is defined as the mean of the energy samples with respect to the kth modulation channel:

$$\Phi_{1,m}(k) = \frac{\sum_{i=1}^{N} \varepsilon_{i,k}(m)}{N}. \tag{3.10}$$

The second set, $\Phi_{2,m}(k)$, is defined as the ratio of the geometric mean of a spectral energy measure and its arithmetic mean value, thus representing the spectral flatness of the spectrum. A spectral flatness value close to 1 is related to a flat spectrum, while a value close to 0 suggests a spectrum with high variations in its spectral amplitude. This measure is computed as follows:

$$\Phi_{2,m}(k) = \frac{\sqrt[N]{\Pi_{i=1}^{N} \varepsilon_{i,k}(m)}}{\Phi_{1,m}(k)}. \tag{3.11}$$

The third $\Phi_{3,m}(k)$ measure corresponds to the center of mass of each modulation channel, where $f(i)$ is the index of the critical band. The spectral centroid for the *ith* modulation channel is given by:

$$\Phi_{3,m}(k) = \frac{\sum_{i=1}^{N} f(i)\varepsilon_{i,k}(m)}{\varepsilon_{i,k}(m)}. \tag{3.12}$$

In order to measure the relationship of different modulation channels, the 23 acoustic channels are grouped in five levels: $D_1 = [1-4]$, $D_2 = [5-8]$, $D_3 = [9-12]$, $D_4 = [13-18]$ and $D_5 = [19-23]$. The modulation channels into each category are summed and used to compute the spectral centroid $\Phi_{4,m}(k)$ in the modulation frequency domain for D_l as follows:

$$\mathrm{E}_m(l,k) = \sum_{i \in D_l} \varepsilon_{i,k}(m), \tag{3.13}$$

$$\Phi_{4,m}(k) = \frac{\sum_{k=1}^{8} k\mathrm{E}_m(l,k)}{\sum_{k=1}^{8} \mathrm{E}_m(l,k)}. \tag{3.14}$$

The two final measurements capture the rate of change of each acoustic frequency region, thus providing an indication of the temporal dynamics of the utterances. The linear regression coefficient $\Phi_{5,m}(k)$ (slope) and the corresponding regression error $\Phi_{6,m}(k)$ (root mean squared error, RMSE) are computed. Those measures are associated with the first-degree polynomial model used to fit $\mathrm{E}_m(l,k)$. The final feature vector includes 184 modulation spectrum energy features $\varepsilon_{i,k}(m), k = 1, \ldots, 8$ plus the 39 features described above, thus totaling 223 features.

Acoustic-prosodic measures

Acoustic-prosodic features and their variations between groups have been the most widely used features in the analysis of autism spectrum disorder. Here, vocalization utterances were acoustically analyzed using the VoiceSauce MATLAB toolbox from the UCLA SPAP laboratory. Many of the parameters estimated by VoiceSauce depend on F0 and the formant range of the input signal. In our experiments, both measures are optimized for children's vocalizations. Pitch and formant-related parameters were computed using a fundamental frequency (F0) range between 60 and 1,600 Hz and a nominal frequency F1 of 1,250 Hz, which correspond to the nominal frequency of a 7 cm vocal tract. The features were extracted from 25 ms frames every 10 ms.

In total, 26 acoustic parameters were extracted, as listed in Table 3.3. The final feature group includes those related to intonation (pitch), maturity of speech (first formant frequencies and amplitudes), volume (energy) and measures of vocal quality such as voice breathiness and harshness/creakiness (harmonics, spectral tilt and cepstral peak prominence). In order to explore the variations of prosodic features between ASD and control groups, three different prosodic feature combinations are proposed. The first group (PF1) includes the

Table 3.3: List of extracted acoustic-prosodic parameters.

Parameter	Acronym
Fundamental frequency	F0
Formant frequencies	F1, F2, F3, F4
Formant frequency bandwidths	BW1, BW2, BW3, BW4
Harmonic spectra (location and magnitude)	H1, H2, H4, A1, A2, A3
Differences of harmonic spectra at corrected formant frequencies	$H1^*$-$H2^*$, $H2^*$-$H4^*$, $H1^*$-$A1^*$, $H1^*$-$A2^*$, $H1^*$-$A3^*$
Volume	Energy
Cepstral peak prominence	CPP
Harmonic to noise ratio	HNR5, HNR15, HNR25, HNR35

mean value of the features reported in Table 3.3 for each vocalization utterance. The second group (PF2) combined the distribution mean and standard deviation, and finally, mean, standard deviation and range are included in the third group (PF3). Such a partition was shown useful in [35].

3.4.5 Classification

In our experiments, three SVMs are trained separately on vocalization utterances from the three different feature groups, namely wavelets, modulation spectral and acoustic-prosodic. The SVM implementation in [57] was adopted and an RBF kernel was chosen as it resulted in improved performance during our pilot experiments. In order to find the optimal parameters for each SVM, a fourfold grid search methodology was used. Classifier performance is measured using stratified 10-fold cross-validation. In stratified 10-fold CV, the feature vector is divided into partitions of approximately four individuals, where nine sets are used for training and the rest are left for testing only. The sets are designed to ensure the classes are equally represented across each test fold. The process is repeated 10 times, and the performance is computed on a per-participant basis. With this approach, an infant is labeled as control or ASD using a score-based scheme of the decisions made by the SVM. This method was chosen empirically due to its superior performance when compared with the common method of plurality vote. More specifically, SVM outputs of the vocalization utterances for each participant are compared and the vocalization with the highest likelihood score decides the

final class prediction. Thus, if $c(x_i)$ corresponds to the prediction score for sample x_i, then the final prediction score can be computed as:

$$C = \arg\max[c(x_1) \cdots c(x_i)], \tag{3.15}$$

where $c(x_i)$ corresponds to the distance of the sample x_i to the separating hyperplane.

3.4.6 Fusion schemes

Decision-level fusion

Decision-level fusion schemes combine the decisions from different classifiers in order to achieve higher robustness and to improve the performance of single-classifier systems. The combination problem consists of finding the combination function accepting N-dimensional input vectors from M classifiers and outputting N final classification decisions, where the optimal function is the function that minimizes the misclassification cost. The input vector depends on the combination function and it can be probabilities, scores or labels from the classifiers. In our experiments, all the samples x_i belonging to the participant s_l from the different classifiers are used for the combination problem in order to make a per-participant diagnosis. Three different combination functions are proposed:

1. Plurality vote (PV): This is the simplest and most common fusion method. The participant s_l is assigned to the class c_j that obtained the highest number of votes. In this case, all the classifier weights are equal, that is, $w_k = 1/K \ \forall K$.
2. Maximum probability vote (MPV): In this fusion scheme, the probabilities for the samples x_i belonging to the participant s_l are compared and the sample with the highest probability decides the final prediction:

$$C = \arg\max[p(x_1) \cdots p(x_i)]. \tag{3.16}$$

3. Average probability vote (APV): The per-sample conditional probabilities per each class are averaged and the class c_j that obtained the highest average probability decides the final prediction. Thus if $p(c_j/x_i)$ is the conditional probability that x_i belongs to the class c_j, the final prediction is made as follows:

$$C = \arg\max\left[\frac{\sum_i p(c_1/x_i)}{i}, \frac{\sum_i p(c_2/x_i)}{i}\right]. \tag{3.17}$$

In our experiments, the classifiers in the ensembles are comparable in the sense that they have been trained on the same data sets and using the same partitioning.

Feature-level fusion

In feature-level fusion, the feature sets from different sources are concatenated into a single feature vector before the classification process. The main advantage of this method is that correlated features within and between different feature sets can be removed via dimensionality reduction tools, thus improving the generability of the system. In our experiments, a mutual information (MI)-based algorithm was used in order to measure the degree of relatedness between the feature values [58, 59]. The MI of the feature set is computed for the algorithm using the nearest-neighbor method. The details of the method are presented in [58].

3.5 Results

3.5.1 Experiment 1: Wavelet mother selection

Our first aim was to investigate and compare the effectiveness of different types of mother wavelets and the distribution of information in several decomposition levels for the discrimination of autism spectrum disorder. In order to do so, the extraction feature methodology proposed in the section "Wavelet packet decomposition" is employed using different wavelet families, such as Daubechies (db1-db10), coiflet (coif1-coif10), symlet (sym2-sym10), biorthogonal (bior1.1, bior1.3, bior1.5, bior2.2, bior2.4, bior2.6, bior,2.8, bior3.1, bior3.3) and reverse biorthogonal (rbior1.1, rbior1.3, rbior1.5, rbior2.2, rbior2.4, rbior2.6, rbior,2.8, rbior3.1, rbior3.3). Energy and entropy features are extracted from the wavelet-packet coefficient at several decomposition levels and compared between them for each kind of mother wavelet.

A summary of the best performance results per wavelet family is presented in Table 3.4. Columns labelled Acc, Sens and Spec correspond to classifier accuracy, sensitivity and specificity. The index number specified in each wavelet family refers to the vanishing order of the wavelet, which is related to the length of the filter. Maximum recognition accuracy of 81.5% was achieved using a first level decomposition and an eighteenth order Daubechies mother

wavelet. Additionally, it was found that increasing the wavelet decomposition level improved the general performance for most mother wavelets.

Table 3.4: Summary of best results per wavelet family.

Wavelet	Acc %	Sen %	Spec %	Level of decomposition
db8	81.5	91.6	70.0	1
coif7	76.5	91.6	60.0	3
sym10	76.5	83.3	70.0	3
bior2.6	77.0	91.6	60.0	4
rbior3.1	71.5	73.3	70.0	4

3.5.2 Experiment 2: Feature set comparisons

Experimental results for the different feature sets proposed in Section 3.4 are reported in Table 3.5 As can be seen, the proposed wavelet features achieved the best accuracy, reaching up to 81.5% average recognition rate. They are followed by the modulation spectral features with a performance of 79.0%. The three benchmark prosodic feature sets achieved similar performances with PF1 achieving the best overall performance among the PF1-PF3 (see Section 3.4.4 for details).

Table 3.5: Recognition results for the different feature sets proposed.

Features group	Acc (%)	Sen (%)	Spec (%)	AUC
Prosodic-PF1	71.5	90.0	50.0	0.71
Prosodic-PF2	65.0	86.6	40.0	0.69
Prosodic-PF3	65.0	91.6	35.0	0.56
Wavelet features “db8”	81.5	91.6	70.0	0.81
Modulation spectral features	79.0	80.0	75.0	0.72

The average recognition rates in Table 3.5 are reported on a per-participant basis as is described in Section 3.4.5. A total of 43 children (23 with ASD, 20 control) were classified for each classification model separately. The vocalization with the

highest score per individual made the final diagnosis. Tables 3.6 and 3.7 present the details of the diagnosis made between control and ASD groups for each set of features. Specifically, Table 3.6 shows the number of children correctly classified (true positives and true negatives) and the vocalization that most contributed to the final classification per child. Table 3.7, in turn, follows the same methodology and shows the children that were incorrectly classified for each model (false positive and false negative).

Table 3.6: Children correctly classified per model.

	Prosodic-PF1 features		Wavelet features "db8"		Modulation spectral features	
Vocalization	ASD	Control	ASD	Control	ASD	Control
Babble	0(0%)	1(5%)	2(8.7%)	2(10%)	3(13%)	1(5%)
Speech	2(8.6%)	0(0%)	6(26%)	0(0%)	1(4.3%)	4(20%)
Laugh	3(13%)	1(5%)	2(8.6%)	1(5%)	1(4.3%)	0(0%)
Others	1(4.3%)	4(20%)	3(13%)	7(35%)	4 (17.3%)	6(30%)
Negative emotions	15(65.3%)	4(20%)	8(34.7%)	4(20%)	10(43.4%)	4(20%)
TP/TN	21(91.3%)	10 (50%)	21(91.3%)	14(70%)	19(82.6%)	15(75%)

Table 3.7: Children incorrectly classified per model.

	Prosodic-PF1 features		Wavelet features "db8"		Modulation spectral features	
Vocalization	ASD	Control	ASD	Control	ASD	Control
Babble	1(4.3%)	0 (0%)	1(4.3%)	1(5%)	2(8.6%)	2(10%)
Speech	0(0%)	1(5%)	0(0%)	1(5%)	0(0%)	1(5%)
Laugh	0(0%)	2(10%)	0(0%)	0(0%)	0(0%)	0(0%)
Others	0(0%)	1 (5%)	0(0%)	2(10%)	0(0%)	1(5%)
Negative emotions	1(4.3%)	6(30%)	1(4.3%)	2(10%)	2(8.6%)	1(5%)
FP/FN	2(8.6%)	10(50%)	2(8.6%)	6(30%)	4(17.3%)	5(25%)

As shown in Table 3.6, the vocalizations with the highest score among all the feature groups are the negative emotions. This group includes pain/anger related

vocalizations such as cry, squeal, whine and shout. The prosodic features resulted in the highest number of individuals with ASD classified correctly (n = 15) through "negative emotions" vocalizations, followed by modulation spectral features (n = 10). In turn, modulation spectral and wavelet features contributed mostly to correctly assigning the control label within the "others" vocalization class. On the other hand, while "negative emotions" vocalizations were shown to be helpful in the classification of ASD, they contributed negatively to the labelling of control cases, as shown in Table 3.7.

3.5.3 Experiment 3: Decision- and feature-level fusion

First, decision-level fusion was performed by combining decisions from the classifiers that were trained and tested by prosodic, wavelet and modulation features independently. Three different decision-level fusion schemes, as described in Section 3.4.6, were employed. Table 3.8 has the classification results for plurality (PV), maximum probability (MPV) and average probability (APV) fusion schemes. Also, each fusion scheme was tested under different ensembles where WF, MF and FC1 correspond to the wavelet, modulation and (mean) prosodic features, respectively. As shown in Table 3.8, the combination of wavelet and modulation features achieved the highest performance over all methods tested. Notwithstanding, decision-level fusion did not improve the performance obtained with the individual classifiers, as reported in Table 3.5.

Next, feature-level fusion combined with an MI dimensionality reduction scheme was employed. In the end, an SVM classifier was trained on the top-17 features and the results are reported at the bottom of Table 3.8. As can be seen, feature-level fusion was able to improve the accuracy and specificity of the best individual ASD versus non-ASD classifier, while maintaining the sensitivity level at around 90%. Table 3.9 lists the top 17 features used by this classifier. As can be seen, most of the features are from the modulation spectral class. Finally, Table 3.10 has the details of the classification made between control and ASD groups using the top 17 features selected.

3.6 Discussion

Over the last decade, acoustic-prosodic characterization of children on the autism spectrum has been explored as a possible marker for very early detection. Here, we have explored two new features sets, namely features derived from a

Table 3.8: Recognition results for different decision level fusion and feature-level schemes.

Decision level fusion			
Plurality vote (PV)			
Features group	Acc (%)	Sen (%)	Spec (%)
WF+MF+PF1	71.5	81.6	60.0
WF+MF	74.0	91.6	55.0
WF+PF1	71.5	76.6	65.0
MF+PF1	63.0	81.6	40.0
Maximum probability vote (MPV)			
Features group	Acc (%)	Sen (%)	Spec (%)
WF+MF+PF1	69.5	90.0	45.0
WF+MF	79.0	90.0	65.0
WF+PF1	74.0	95.0	50.0
MF+PF1	67.5	95.0	35.0
Average probability vote (APV)			
Features group	Acc (%)	Sen (%)	Spec (%)
WF+MF+PF1	69.0	90.0	45.0
WF+MF	74.0	90.0	55.0
WF+PF1	61.5	90.0	30.0
MF+PF1	69.5	95.0	40.0
Feature-level fusion			
Features group	Acc (%)	Sen (%)	Spec (%)
WF+MF+PF1	86.5	90.0	80.0

wavelet packet decomposition and features derived from an auditory-inspired spectro-temporal feature representation. We showed that in a cohort of 18-month-old toddlers, we were able to accurately discriminate between toddlers with ASD and controls with accuracies higher than those achieved with previously proposed prosodic features [35]. Such findings are important as early detection can allow for early interventions to commence, notably improving

Table 3.9: Top 17 features chosen using the mutual information-based algorithm for classification of ASD and control groups.

Feature selected	Type of feature
Entropy[1,1]_mean	Wavelet
$\varepsilon_{23,6}$	MSF
$\Phi_{2,m}(6)$	MSF
$\varepsilon_{8,5}$	MSF
Energy[1,1]_mean	Wavelet
H1A3C_mean	Prosodic
$\varepsilon_{1,8}$	MSF
$\Phi_{1,m}(3)$	MSF
$\varepsilon_{13,6}$	MSF
$\varepsilon_{4,6}$	MSF
$\varepsilon_{4,4}$	MSF
Energy[1,0]_mean	Wavelet
A1_mean	Prosodic
Energy[1,0]_std	Wavelet
$\varepsilon_{5,5}$	MS(n = 20)F
$\varepsilon_{4,7}$	MSF
$\varepsilon_{14,8}$	MSF

prognosis [6]. In the sections to follow, we discuss in detail the major findings of our study in light of the existing literature.

3.6.1 Vocalization types

Infants produce a wide variety of vocal expressions during the first years of life. Children use these vocal expressions as a communicative means to express different emotions or different communicative functions to caregivers. From an acoustic point of view, these vocalizations also exhibit different patterns that can be the subject of further analyses [60–62]. Several studies in language

Table 3.10: Children correctly classified after feature-level fusion.

	Selected features	
Vocalization	**ASD**	**Control**
Babble	4(17.4%)	1(5%)
Speech	1(4.3%)	6(30%)
Laugh	4(17.4%)	1(5%)
Others	4(17.4%)	5(25%)
Negative emotions	8(34.7%)	3(15%)
FP/FN	21(91.3%)	16(80%)

acquisition and early identification of pathologies have analyzed the specific characteristics of different vocalizations such as cries, babbles, laughter and grunts due to their relevance during language and communicative skills development [12, 16, 17, 37, 38, 41–45, 60–63].

Vocal development in infants is considered a continuous, but nonlinear process. Early vocalizations are precursors of speech and language development. In the literature, vocalizations such as crying and babbling have received significant attention. For example, babbling is likely to influence the development of spoken language due to the fact that words are composed of canonical syllables [17, 60, 64]. Infants start to produce canonical babbling around 10 months and it has been shown that any delay in the onset of canonical babbling is a significant predictor of language delay or other disabilities [65, 66]. Given the relationship between language acquisition and babbling, a few studies have explored the production of babbling in individuals with ASD [17, 43, 44]. Infants diagnosed with ASD display low rates of canonical babbling, lower number of total syllables produced (volubility) and a later onset in canonical babbling stage compared with typically developing children [17, 44].

Crying, in contrast, is the first method of communication for an infant. It is used to express different needs, states and demands. Frequency vibration of the vocal cords has been related to the dominance of laryngeal processes in early sound production [39]. Previous studies, including autism spectrum disorder analysis, have shown that low birth weight infants and infants with neurological symptoms have different acoustic patterns such as fundamental frequency (F0), vocal tract resonance frequencies, pause length, amplitude modulations and number of utterances compared with typically developing children [12, 16, 17, 37, 41–45]. In babies later diagnosed with ASD, cries were shown to convey high

levels of distress, a factor later attributed to modulation deficits and unnatural F0 and formant values [36].

In our study, vocalizations such as cry, squeal, whine and shouts (called here "negative emotions") were grouped in order to allow for a more balanced class relative to, for example, speech and babble. Table 3.6, in fact, suggests that such vocalization types were more important than speech or babble in helping correctly classify between ASD and controls, regardless of the feature used. Similar findings could be seen with feature fusion, as reported in Table 3.10.

Laughter has also been studied within the infant population to convey information about the child's mental/affective state [67]. Individuals with ASD have been reported as having laughter episodes without an apparent motivating stimulus [61, 68]. Other studies have shown that patients with neurological disorder exhibit uncontrollable episodes of pathological crying, pathological laughing or both, potentially related to impairment in the control and use of their emotions [69, 70]. Within our study, laughter was shown to be a useful vocalization type to help correctly detect ASD, particularly within the prosodic feature space (see Table 3.6).

3.6.2 Features

Mother wavelets

With wavelet packet decomposition, the signal is decomposed into scaled and translated versions of a mother wavelet. As each family of mother wavelets has different characteristics such as symmetry, orthogonality, filter length and vanishing order, different signal properties may be captured by different mother wavelets. In this investigation, Daubechies 8 ("db8") was deemed the best mother wavelet to discriminate between control and ASD groups among other tested mother wavelets, including: coiflet, symlet, biorthogonal and reverse biorthogonal. Additionally, our results showed that increasing the level of decomposition led to more detailed features and, consequently, better classification performance. This was true for all tested mother wavelets, except db8, in which a one-level decomposition showed to be optimal (see Table 3.4).

In the speech processing literature, the db8 mother wavelet has been shown to be widely used across numerous applications, including enhancement, compression and recognition, to name a few [71]. Wavelet decomposition has also been used in the past for pathological cry and pathological speech analysis [49, 50, 72, 73]. This is the first time, however, that wavelet features have been explored for autism spectrum diagnosis. In [50], for example, cry

signals were decomposed into five levels using four different mother wavelets from the Daubechies family, namely: “db1”, “db4”, “db10” and “db20”. The highest classification accuracy was achieved at the fifth level of decomposition using the “db20” mother wavelet.

While higher decomposition levels may have assisted with pathological cry detection, a simple one-level decomposition showed here to be optimal for the task at hand. Such decomposition likely sufficed to measure the energy differences typically reported within the ASD literature (e.g., [44]), the high frequency entropy representative of breathy and harsh sounds, as well as high frequency energies representative of squeal/cry quality [44]. As shown in Table 3.6, wavelet features were useful for the speech, other and negative emotion vocalization classes, thus likely capturing these qualities, respectively.

Prosodic features

According to Table 3.6, prosodic features were shown to be particularly useful in discriminating between ASD and controls within the negative emotion vocalization class, contributing to the correct classification of roughly half the participants. Such findings corroborate those previously published in the literature, which have shown cries to have different F0 and formant frequencies between ASD and controls [12, 16, 17, 37, 41–45, 74]. The “others” and “laughter” categories were the second to contribute mostly to correct classification. Such findings also corroborate those in the literature that have shown laughter to an affect spectral tilt, F0 and first formant amplitudes [75]. Prosodic features have been typically explored in the literature and are used here as a benchmark to the proposed system, as well as providing complementary information to the proposed wavelet and modulation spectral features.

Wavelet features

As per Table 3.6, negative emotion, others vocalizations and speech were the top three vocalization classes, respectively, contributing to correct classification when using only wavelet features. Wavelet features computed from the one-level decomposition basically explore energy levels and variability in high and low frequency ranges, as well as spectral entropy. In the past, such details, while not computed via WPD, were shown to discriminate between the two groups. Spectral entropy, for example, was related to rhythmic cues, breathiness, and harshness and could discriminate between typically developing children and

those with ASD [40]. Energy variability, in turn, was shown to be a correlate of perceived prosody atypicality in ASD [76]. Wavelet features have also been shown useful in pathological cry detection and in emotion recognition from speech [77, 78]. In [77], for example, wavelet-based features were shown to be useful for anger detection, whereas in [78], they were shown to be useful in discriminating angry and disgust emotions. By computing wavelet features for different vocalization classes separately, different attributes could be measured, thus contributing positively toward ASD detection.

Modulation features

Auditory-inspired modulation features have been used in the past for pathological speech characterization [79–81] and speech emotion recognition [54]. High frequency modulations have also been linked to turbulence noise in no-pain cries [40], whereas certain cry modulation frequencies have been linked to central nervous system disorder [40]. Additionally, recent research has suggested impaired extraction of speech rhythm from temporal modulation patterns in speech in developmental dyslexia, an impaired neural representation of the sound structure of words typically observed with individuals on the spectrum [82]. The work described herein exemplifies the first attempt at using modulation features for ASD detection. Such findings corroborate the observation that the modulation spectral features that contributed the most toward the task at hand were computed from negative emotion, other and speech classes (see Table 3.6).

3.6.3 Overall accuracy

Table 3.5 shows that wavelet features achieved the best overall accuracy and sensitivity, outperforming the benchmark prosodic features and proposed modulation features. The modulation features, in turn, resulted in the highest specificity. These findings suggest that different features may contribute complementary information to overall ASD detection. The next section discusses the obtained findings aimed at fusing information at the decision and feature levels.

Decision-level fusion

Decision-level fusion is a widely explored method in machine learning and pattern recognition that typically improves the performance over single classifiers

[83, 84]. Here, three decision-level schemes were explored. However, as reported in Table 3.8, none of the ensemble methods outperformed the results achieved with a single classifier. Plurality vote, for example, helped improve the specificity of the benchmark prosodic classifier when combined with decisions from the wavelet classifier and from the three feature classes together, but the overall performance was below that achieved with wavelet and modulation spectral classifiers. Within the maximum probability fusion scheme, in turn, combining decisions from wavelet and modulation feature classifiers helped improve the sensitivity of the overall system, but at the cost of reduced specificity. Similar findings were observed for the average probability vote fusion scheme. From Tables 3.6 and 3.7, it can be seen that the majority of the correct decisions have been made based on negative emotion classes, regardless of the tested feature. As such, fusing decisions of individual classifiers may be overlooking the complementarity of the different feature sets and placing most weight on features that convey similar information. To overcome this potential limitation, feature fusion with feature selection has been explored.

Feature-level Fusion

Feature fusion combined with a mutual information-based selection algorithm should be able to sift out top features that convey complementary information from different vocalization classes. Results in Table 3.10 corroborate this hypothesis and show that, after feature fusion, the different classes are contributing to the overall accuracy in a more balanced manner. While negative emotions still play a crucial role, other classes such as babble and laughter have stood out. As expected, by attending to complementary details extracted from the different feature sets, improved overall performance is achieved. Overall, relative to using only prosodic features, improvements of 21% and 60% could be seen in accuracy and specificity, respectively. Relative to using only the wavelet features, improvements of 6.1% and 14.3% could be seen in accuracy and specificity, respectively, with a small drop in sensitivity of 1.7%. Lastly, relative to using only the modulation spectral features, gains of 9.5%, 12.5% and 6.7% in accuracy, sensitivity and specificity could be observed, respectively.

Close inspection of the top features reported in Table 3.9 further validate the claim that feature-level fusion has allowed different features to extract complementary information from different vocalization classes. For example, the first formant amplitude (A1_mean) and spectral tilt (H1A3C_mean) measures have been used within laughter research [75]. The entropy[1,1]_mean feature, in turn, conveys detail about high frequency entropy, a metric previously related

to vocalizations that fall under the "other" class. The authors in [49] also argue that the vibration of the vocal cords is reflected in the entropy of the wavelet parameters, a common finding reported in the ASD cry literature but computed via the more complex method of fundamental frequency tracking [46]. The energy[1,0]_mean and energy [1,1]_mean features, on the other hand, measure low and high frequency energy, respectively. These features, in turn, have been shown useful in speech vocalization discrimination between ASD and controls, and characterizing squeal quality [44], respectively. Energy variability, characterized by energy[1,0]_std, in turn, has shown to be useful for discriminating speech from ASD and control individuals from older, verbally fluent kids [76].

Interestingly, of the top 17 features chosen by the selection algorithm, 11 correspond to modulation spectral features. Of the modulation energy features $\varepsilon_{i,j}$, all features are from modulation channels higher than 4, corresponding to a modulation frequency greater than 17.6 Hz. It is a well-supported finding that, in running speech, modulation frequencies below 16 Hz contribute toward intelligibility [85]. In the case of preverbal toddlers, such information is not deemed useful and higher modulation frequencies seem to stand out. Within no-pain cries, higher modulation frequencies have been related to turbulent noise [40]. In another study looking at modulations of cries, frequencies around 10–70 Hz were reported and significant differences around 40 Hz were seen for children with brain damage relative to controls [40]. Interestingly, four of the top-selected modulation features convey information about the 6th modulation channel centered near 40 Hz (47.5 Hz, to be more exact). In another recent study, modulation below 20 Hz was shown to significantly differ between children with dyslexia and controls [82]. Three top-features capture such a modulation frequency range near modulation channels 4 and 5. Lastly, the energy distribution feature $\Phi_{1,m}(3)$ has been shown for running speech to be a top discriminative feature to detect sadness and anger emotions [54]. Such features could be detecting distress cues in cries, a finding that has been widely reported in the ASD literature [37].

3.7 Conclusions

Acoustic-prosodic characterization of toddler vocalization utterances has been shown, in some studies, useful to discriminate between children with and without ASD. Existing studies have typically explored irregularities during vocal fold vibration and inappropriate use of volume in individuals with autism [9, 14, 18, 21–26]. Given the wide range of age of the participants, however, and the fast changing vocal tract characteristics during childhood, many of the reported findings have been contradictory [14]. Moreover, findings have typically been reported using

only speech-like utterances (e.g., babble) [17, 43, 44] or cries [12, 16, 17, 37, 41–45], thus it is still not clear which types of vocalization contribute the most to classification. Lastly, a vast body of literature has explored the use of wavelet features for infant cry and pathological speech analysis (e.g., [49–51, 86]), as well as the speech modulation spectrum (e.g., [54, 80, 87]). To the best of the authors' knowledge, however, such features have yet to be explored in ASD. This chapter aims to fill these three gaps.

More specifically, we explore the use of wavelet, modulation spectral and prosodic features to classify 43 18-month-old toddlers (23 children of which were diagnosed as having autism at the 36-month assessment and 20 children age-matched control group) into ASD and non-ASD groups. By focusing only on 18-month-old data, variability from vocal tract maturation is minimized, thus shedding light on features truly discriminative of ASD. Lastly, we explore the contributions of different vocalization types – namely, babble, speech-like, laughter, negative emotions (grouping vocalizations such as cries, whines, squeals and shouts) and others – and also explore the effects different features have on certain vocalization types for overall ASD diagnosis.

Overall, it was found that an accuracy of 81.5%, a sensitivity of 91.6% and a specificity of 70% could be achieved with an individual SVM classifier trained on wavelet based energy and entropy features. When trained with speech modulation spectral features, an accuracy of 79%, a sensitivity of 80% and a specificity of 75% could be achieved. These accuracies compared favorably against prosodic features previously proposed in the literature [35]. Moreover, while decision-level fusion did not improve overall performance, feature-level fusion combined with feature selection achieved an accuracy of 86.5%, a sensitivity of 90% and a specificity of 80%, thus representing a relative improvement over the individual classifier of 5% and 10% in terms of accuracy and specificity, respectively. Close inspection of the top 17 features selected showed that the most important features corresponded to modulation spectral features. Interestingly, it was observed that vocalizations such as cries, squeals, whines and shouts were more discriminative between groups than speech, babble or laugh vocalizations. Such findings could assist clinicians in future assessments, which currently place focus on prosodic nuances during speech-like utterances.

Acknowledgments: This work was supported by funding from the National Science and Engineering Research Council of Canada, the Canadian Institutes of Health Research, Autism Speaks, and NeuroDevNet. We wish to thank the children and their families for participating in the Canadian Infant Sibling Study, as well as Drs. Lonnie Zwaigenbaum, Susan E. Bryson, Wendy Roberts, Isabel M. Smith and Peter Szatmari.

References

[1] Association AP, on Nomenclature C, Statistics, et al, 1960, Diagnostic and statistical manual of mental disorders (DSM-5®), American Psychiatric Association.

[2] A.M. Daniels, and D.S. Mandell Explaining differences in age at autism spectrum disorder diagnosis: A critical review, Autism, 18(5), 583–597, 2014.

[3] S. Ozonoff, G.S. Young, A. Carter, D. Messinger, N. Yirmiya, L. Zwaigenbaum, S. Bryson, L.J. Carver, J.N. Constantino, K. Dobkins et al. Recurrence risk for autism spectrum disorders: A baby siblings research consortium study, Pediatrics, 128(3), e488–e495, 2011.

[4] D.L. Christensen Prevalence and characteristics of autism spectrum disorder among children aged 8 yearsautism and developmental disabilities monitoring network, 11 sites, united states, 2012, MMWR Surveillance Summaries, 65, 2016.

[5] M. Sigman, A. Dijamco, M. Gratier, and A. Rozga Early detection of core deficits in autism, Developmental Disabilities Research Reviews, 10(4), 221–233, 2004.

[6] E.C. Fenske, S. Zalenski, P.J. Krantz, and L.E. McClannahan Age at intervention and treatment outcome for autistic children in a comprehensive intervention program, Analysis and Intervention in Developmental Disabilities, 5(1–2), 49–58, 1985.

[7] K. Sullivan, W.L. Stone, and G. Dawson Potential neural mechanisms underlying the effectiveness of early intervention for children with autism spectrum disorder, Research in Developmental Disabilities, 35(11), 2921–2932, 2014.

[8] C. Lord, S. Risi, L. Lambrecht, J.E.H. Cook, B.L. Leventhal, P.C. DiLavore, A. Pickles, and M. Rutter The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism, Journal of Autism and Developmental Disorders, 30(3), 205–223, 2000.

[9] D. Bone, S. Bishop, R. Gupta, S. Lee, and S. Narayanan Acoustic-prosodic and turn-taking features in interactions with children with neurodevelopmental disorders, Interspeech, 2016, 1185–1189, 2016.

[10] M.G. Filipe, S. Frota, S.L. Castro, and S.G. Vicente Atypical prosody in Asperger syndrome: Perceptual and acoustic measurements, Journal of Autism and Developmental Disorders, 44(8), 1972–1981, 2014.

[11] Y. Nakai, R. Takashima, T. Takiguchi, and S. Takada Speech intonation in children with autism spectrum disorder, Brain and Development, 36(6), 516–522, 2014.

[12] G. Esposito, J. Nakazawa, P. Venuti, and M.H. Bornstein Componential deconstruction of infant distress vocalizations via tree-based models: A study of cry in autism spectrum disorder and typical development, Research in Developmental Disabilities, 34(9), 2717–2724, 2013.

[13] J.S. Stephen, J. Iverson, and B.M. Lester Atypical cry characteristics in infants at risk for autism, 2008.

[14] R. Fusaroli, A. Lambrechts, D. Bang, D.M. Bowler, and S.B. Gaigg Is voice a marker for autism spectrum disorder? A systematic review and meta-analysis, Autism Research, 2016.

[15] L. Kanner Autistic disturbances of affective contact, Acta Paedopsychiatrica, 35(4), 100–136, 1967.

[16] M. Bornstein, K. Costlow, A. Truzzi, and G. Esposito Categorizing the cries of infants with asd versus typically developing infants: A study of adult accuracy and reaction time, Research in Autism Spectrum Disorders, 31, 66–72, 2016.
[17] E. Patten, K. Belardi, G.T. Baranek, L.R. Watson, J.D. Labban, and D.K. Oller Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency, Journal of Autism and Developmental Disorders, 44(10), 2413–2428, 2014.
[18] C. Lord, P.C. DiLavore, A. Pickles, M.J. Elliot, C. Hellreigel, S. Arnold, and L. Tao Pre-linguistic vocalizations and social directedness in autistic, developmentally delayed and typical toddlers, Infant Behavior and Development, 19:61, 1996.
[19] J. McCann, and S. Peppé Prosody in autism spectrum disorders: a critical review, International Journal of Language & Communication Disorders, 38(4), 325–350, 2003.
[20] L.D. Shriberg, R. Paul, J.L. McSweeny, A. Klin, D.J. Cohen, and F.R. Volkmar Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. Journal of Speech, Language, and Hearing Research, 44(5), 1097–1115, 2001.
[21] J. Quigley, S. McNally, and S. Lawson Prosodic patterns in interaction of low-risk and at-risk-of-autism spectrum disorders infants and their mothers at 12 and 18 months, Language Learning and Development, 12(3), 295–310, 2016.
[22] J. Parish-Morris, M. Liberman, N. Ryant, C. Cieri, L. Bateman, E. Ferguson, and R.T. Schultz Exploring autism spectrum disorders using hlt, CLPsych@ HLT-NAACL, 74–84, 2016.
[23] K. Kary, L. Chan, K. Carol, and S. To do individuals with high-functioning autism who speak a tone language show intonation deficits?, Journal of autism and developmental disorders, 46(5), 1784, 2016.
[24] Y.S. Bonneh, Y. Levanon, O. Dean-Pardo, L. Lossos, and Y. Adini Abnormal speech spectrum and increased pitch variability in young autistic children, Frontiers in Human Neuroscience, 4, 237, 2011.
[25] D. Bone, C.C. Lee, M.P. Black, M.E. Williams, S. Lee, P. Levitt, and S. Narayanan The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. Journal of Speech, Language, and Hearing Research, 57 (4), 1162–1177, 2014.
[26] Ricks DM, Wing L Language, communication, and the use of symbols in normal and autistic children, Journal of Autism and Childhood Schizophrenia, 5(3), 191–221, 1975.
[27] R. Paul, A. Augustyn, A. Klin, and F.R. Volkmar Perception and production of prosody by speakers with autism spectrum disorders, Journal of Autism and Developmental Disorders, 35(2), 205–220, 2005.
[28] S. Peppé, J. McCann, F. Gibbon, A. OHare, and M. Rutherford Receptive and expressive prosodic ability in children with high-functioning autism, Journal of Speech, Language, and Hearing Research, 50(4), 1015–1028, 2007.
[29] H. Tager-Flusberg A psycholinguistic perspective on language development in the autistic child, Autism: Nature, Diagnosis, and Treatment, 92–115, 1989.
[30] J. Demouy, M. Plaza, J. Xavier, F. Ringeval, M. Chetouani, D. Perisse, D. Chauvin, S. Viaux, B. Golse, D. Cohen et al. Differential language markers of pathology in autism, pervasive developmental disorder not otherwise specified and specific language impairment, Research in Autism Spectrum Disorders, 5(4), 1402–1412, 2011.

[31] J.J. Diehl, and R. Paul Acoustic and perceptual measurements of prosody production on the profiling elements of prosodic systems in children by children with autism spectrum disorders, Applied Psycholinguistics, 34(1), 135–161, 2013.
[32] K. Hubbard, and D.A. Trauner Intonation and emotion in autistic spectrum disorders, Journal of Psycholinguistic Research, 36(2), 159–173, 2007.
[33] L.M. Morett, K. OHearn, B. Luna, and A.S. Ghuman Altered gesture and speech production in asd detract from in-person communicative quality, Journal of Autism and Developmental Disorders, 46(3), 998, 2016.
[34] M. Asgari, A. Bayestehtashk, and I. Shafran Robust and accurate features for detecting and diagnosing autism spectrum disorders, Interspeech, 191–194, 2013.
[35] J.F. Santos, N. Brosh, T.H. Falk, Zwaigenbaum L, S.E. Bryson, W. Roberts, I.M. Smith, P. Szatmari, and J.A. Brian (2013) Very early detection of autism spectrum disorders based on acoustic analysis of pre-verbal vocalizations of 18-month old toddlers. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 7567–7571
[36] G. Esposito, and P. Venuti Understanding early communication signals in autism: A study of the perception of infants' cry, Journal of Intellectual Disability Research, 54(3), 216–223, 2010.
[37] G. Esposito, M. Del Carmen Rostagno, P. Venuti, J.D. Haltigan, and D.S. Messinger Brief report: Atypical expression of distress during the separation phase of the strange situation procedure in infant siblings at high risk for ASD, Journal of Autism and Developmental Disorders, 44(4), 975–980, 2014.
[38] P.K. Kuhl, and A.N. Meltzoff Infant vocalizations in response to speech: Vocal imitation and developmental change, The Journal of the Acoustical Society of America, 100(4), 2425–2438, 1996.
[39] K. Wermke, W. Mende, C. Manfredi, and P. Bruscaglioni Developmental aspects of infants cry melody and formants, Medical Engineering & Physics, 24(7), 501–514, 2002.
[40] W. Mende, K. Wermke, S. Schindler, K. Wilzopolski, and S. Hock Variability of the cry melody and the melody spectrum as indicators for certain CNS disorders, Early Child Development and Care, 65(1), 95–107, 1990.
[41] F.B. Furlow Human neonatal cry quality as an honest signal of fitness, Evolution and Human Behavior, 18(3), 175–193, 1997.
[42] G. Esposito, and P. Venuti Comparative analysis of crying in children with autism, developmental delays, and typical development, Focus on Autism and Other Developmental Disabilities, 24(4), 240–247, 2009.
[43] S.J. Sheinkopf, P. Mundy, D.K. Oller, and M. Steffens Vocal atypicalities of preverbal autistic children, Journal of Autism and Developmental Disorders, 30(4), 345–354, 2000.
[44] D.K. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. Warren Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development, Proceedings of the National Academy of Sciences, 107(30), 13,354–13,359, 2010.
[45] G. Esposito, and P. Venuti Developmental changes in the fundamental frequency (f0) of infants cries: A study of children with autism spectrum disorder, Early Child Development and Care, 180(8), 1093–1102, 2010.
[46] S.J. Sheinkopf, J.M. Iverson, M.L. Rinaldi, and B.M. Lester Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder, Autism Research, 5(5), 331–339, 2012.

[47] L. Zwaigenbaum, S. Bryson, T. Rogers, W. Roberts, J. Brian, and P. Szatmari Behavioral manifestations of autism in the first year of life, International Journal of Developmental Neuroscience, 23(2), 143–152, 2005.
[48] C. Lord, M. Rutter, and A. Le Couteur Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders, Journal of Autism and Developmental Disorders, 24(5), 659–685, 1994.
[49] R. Behroozmand, and F. Almasganj Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients speech signal with unilateral vocal fold paralysis, Computers in Biology and Medicine, 37(4), 474–485, 2007.
[50] M. Hariharan, S. Yaacob, and S.A. Awang Pathological infant cry analysis using wavelet packet transform and probabilistic neural network, Expert Systems with Applications, 38(12), 15,377–15,382, 2011.
[51] Y. Huang, A. Wu, G. Zhang, and Y. Li Speech emotion recognition based on coiflet wavelet packet cepstral coefficients, Chinese Conference on Pattern Recognition, Springer, 436–443, 2014.
[52] S. Mallat A wavelet tour of signal processing, Academic press, 1999.
[53] C.S. Burrus, R.A. Gopinath, and H. Guo Introduction to wavelets and wavelet transforms: A primer, 1997.
[54] S. Wu, T.H. Falk, and W.Y. Chan Automatic speech emotion recognition using modulation spectral features, Speech Communication, 53(5), 768–785, 2011.
[55] I. Ree P. 56: Objective measurement of active speech level, 1993.
[56] M. Slaney et al. An efficient implementation of the Patterson-Holdsworth auditory filter bank, Apple Computer, Perception Group, Tech Rep, 35, 8, 1993.
[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.
[58] B.C. Ross Mutual information between discrete and continuous data sets, PloS one, 9(2), e87,357, 2014.
[59] A. Kraskov, H. Stögbauer, and P. Grassberger Estimating mutual information, Physical Review E, 69(6), 066,138, 2004.
[60] H. Rothgänger Analysis of the sounds of the child in the first year of age and a comparison to the language, Early Human Development, 75(1), 55–69, 2003.
[61] E. Schoen, R. Paul, and K. Chawarska Phonology and vocal behavior in toddlers with autism spectrum disorders, Autism Research, 4(3), 177–188, 2011.
[62] H.C. Hsu, A. Fogel, and R.B. Cooper Infant vocal development during the first 6 months: Speech quality and melodic complexity, Infant and Child Development, 9(1), 1–16, 2000.
[63] C. Papaeliou, G. Minadakis, and D. Cavouras Acoustic patterns of infant vocalizations expressing emotions and communicative functions. Journal of Speech, Language, and Hearing Research, 45(2), 311–317, 2002.
[64] S. Nakazima A comparative study of the speech developments of Japanese and American English in childhood (1): A comparison of the developments of voices at the prelinguistic period, 1962.

[65] I. Van den Dikkenberg-pot, B.F. Koopmans-van, and C. Clement Influence of lack of auditory speech perception on sound productions of deaf infants, Proceedings of the Institute of Phonetic Sciences Amsterdam, 22, 47–60, 1998.
[66] N. Masataka Why early linguistic milestones are delayed in children with Williams syndrome: Late onset of hand banging as a possible rate–limiting constraint on the emergence of canonical babbling, Developmental Science, 4(2), 158–164, 2001.
[67] H. Rao, J.C. Kim, A. Rozga, and M.A. Clements Detection of laughter in children's speech using spectral and prosodic acoustic features, INTERSPEECH, 1399–1403, 2013.
[68] T.J. Runyon (2015) Function of laughter from a student with autism. PhD thesis, San Francisco State University.
[69] J. Parvizi, S.W. Anderson, C.O. Martin, H. Damasio, and A.R. Damasio Pathological laughter and crying: a link to the cerebellum, Brain, 124(9), 1708–1719, 2001.
[70] J. Parvizi, D.B. Arciniegas, G.L. Bernardini, M.W. Hoffmann, J.P. Mohr, M.J. Rapoport, J.D. Schmahmann, J.M. Silver, and S. Tuhrim Diagnosis and management of pathological laughter and crying, Mayo Clinic Proceedings, Elsevier, 81, 1482–1486, 2006.
[71] M.H. Farouk Application of wavelets in speech processing, Springer, 2014.
[72] J. Saraswathy, M. Hariharan, T. Nadarajaw, W. Khairunizam, and S. Yaacob Optimal selection of mother wavelet for accurate infant cry classification, Australasian Physical & Engineering Sciences in Medicine, 37(2), 439–456, 2014.
[73] Y. Long, L. Gang, and G. Jun Selection of the best wavelet base for speech signal, Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing, IEEE, 218–221, 2004.
[74] S. Orlandi, L. Bocchi, C. Manfredi, M. Puopolo, A. Guzzetta, S. Vicari, and M.L. Scattoni (2011) Study of cry patterns in infants at high risk for autism. In: International Workshop MAVEBA, pp. 7–10.
[75] C. Menezes, and Y. Igarashi The speech laugh spectrum, Proc Speech Production, Brazil, 157–164, 2006.
[76] D. Bone, M.P. Black, C.C. Lee, M.E. Williams, P. Levitt, S. Lee, and S. Narayanan Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist, InterSpeech, 1043–1046, 2012.
[77] V.N. Degaonkar, and S.D. Apte Emotion modeling from speech signal based on wavelet packet transform, International Journal of Speech Technology, 16(1), 1–5, 2013.
[78] H.K. Palo, and M.N. Mohanty Wavelet based feature combination for recognition of emotions, Ain Shams Engineering Journal, 2017.
[79] T.H. Falk, W.Y. Chan, and F. Shein Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility, Speech Communication, 54(5), 622–631, 2012.
[80] M. Markaki, and Y. Stylianou Voice pathology detection and discrimination based on modulation spectral features, IEEE Transactions on Audio, Speech, and Language Processing, 19(7), 1938–1948, 2011.
[81] M. Markaki, and Y. Stylianou (2009) Normalized modulation spectral features for cross-database voice pathology detection. In: Tenth Annual Conference of the International Speech Communication Association.
[82] V. Leong, and U. Goswami Impaired extraction of speech rhythm from temporal modulation patterns in speech in developmental dyslexia, Frontiers in Human Neuroscience, 8, 2014.

[83] D.W. Opitz, and R. Maclin Popular ensemble methods: An empirical study, Journal of Artificial Intelligence Research(JAIR), 11, 169–198, 1999.
[84] N.M. Baba, M. Makhtar, S.A. Fadzli, and M.K. Awang Current issues in ensemble methods and its applications, Journal of Theoretical and Applied Information Technology, 81(2), 266, 2015.
[85] R. Drullman, J.M. Festen, and R. Plomp Effect of reducing slow temporal modulations on speech reception, The Journal of the Acoustical Society of America, 95(5), 2670–2680, 1994.
[86] J. Saraswathy, M. Hariharan, V. Vijean, S. Yaacob, and W. Khairunizam Performance comparison of Daubechies wavelet family in infant cry classification, 2012 IEEE 8th International Colloquium on Signal Processing and Its Applications (CSPA), IEEE, 451–455, 2012.
[87] N. Malyska, T.F. Quatieri, and D. Sturim (2005) Automatic dysphonia recognition using biologically-inspired amplitude-modulation features. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP'05), IEEE, vol 1, pp I–873.

Shou-Chun Yin and Richard Rose

4 Computer-aided speech therapy for dysarthric speakers: Statistical acoustic modeling for automated verification of pronunciation accuracy

Abstract: The objective of this study is to develop statistical modeling techniques for characterizing phonetic variation in automatic speech recognition (ASR). One issue addressed in this domain is to reliably detect the phoneme-level mispronunciations in speech utterances that arise from speech therapy applications. Another issue addressed in this work is to study the ability of ASR systems to model the phonetic variation that often exists in speaker-independent recognition tasks. In order to address these issues, first, a phoneme-level pronunciation verification (PV) scenario is investigated for detecting the mispronunciation occurrences in speech utterances recorded from a population of impaired children with neuromuscular disorders. The well-known continuous density hidden Markov model (CDHMM) is used as a phoneme decoder which generates a finite state network of phoneme string hypotheses for input speech utterances. The phoneme-level confidence measures can be constructed from this network, and PV decision can be made by comparing the confidence measures with a pre-selected threshold. Second, the subspace Gaussian mixture model (SGMM) formalism is incorporated into a new PV scenario. A new kind of pronunciation confidence measure used for making mispronunciation verification decisions is extracted directly from the state-level model parameters. Both session-level and utterance-level PV scenarios based on the SGMM-based confidence measures are proposed. In the session-level PV task, the equal error rate can be reduced by 15.35% when combining the SGMM-based confidence measures with the above phoneme decoder-based confidence measures. In the utterance-level PV task, the equal error rate can be reduced by 12.94%. This equal error rate reduction is believed to result from an efficient characterization of pronunciation variation for each phoneme by the SGMM.

Note: Richard Rose and Shou-Chun Yin are now at Google, New York City, USA and Nuance Communications Canada Inc, respectively. The work was done while the authors were at McGill University, Montreal, Canada, and it does not contain any Google and Nuance Communications proprietary information.

Shou-Chun Yin, Nuance Communications Canada Inc., Montreal, Quebec, Canada
Richard Rose, Google, New York, USA

https://doi.org/10.1515/9781501513138-004

Keywords: speech therapy, pronunciation verification, subspace Gaussian mixture model

4.1 Automatic intelligibility assessment of speech disorders

A major application of the work presented in this article is in the area of computer-aided speech and language therapy (CASLT). The specific task addressed here is verifying the quality of pronunciations in utterances from a population of impaired children and young adults with neuromuscular disorders. However, the area of CASLT is quite broad. A wide range of speech technologies have been applied to treating speech disorders in speaker populations whose disabilities range from mild to profound [1–5].

This chapter has three goals. First, a brief overview of work in the area of CASLT will be presented. Second, the CASLT system that was used in this project is described. This system was originally developed as part of the "Comunica" project and evaluated at the Public School for Special Education (CPEE), "Alborada," in Zaragiza, Spain [6]. Third, the data collection scenario, speaker population and speech annotation strategy associated with the pronunciation verification (PV) task domain considered in this article are presented.

4.1.1 Computer-aided speech and language therapy

Speech therapy for disabled individuals involves interaction between the patient and a highly trained speech therapist. This interaction is considered to be important for providing personalized diagnoses and assessment of patients' progress. Tools for CASLT have been developed to enhance the efficiency and the quality of services provided by the human therapist. Diagnosing and assessing performance can be made more efficient for the therapist by automating portions of the interactive component of therapy. This reduces the time and the level of expertise required for providing the interactive component of therapy. As a result, these services can be made available to a large population of disabled individuals. Another potential result of this automation is the quality of the therapists' diagnoses can be made more consistent. This can be achieved through the use of more accurate and objective measures of patients' performance.

Speech technology applications to CASLT

There are a number of CASLT applications that have been developed over the last decade in an effort to address specific kinds of speech disabilities. Most of these applications are based on statistical modeling approaches that were developed for speech and language processing. These approaches include HMM-based ASR, Gaussian mixture model (GMM)-based speaker modeling and machine learning based pronunciation modeling.

Automatic speech recognition for severely disabled speakers: One of the most compelling applications for ASR in the disabled community is for communication aids for individuals with severe dysarthria [7]. These communication aids are called for in cases where speech impairments are severe. The disabled speaker is equipped with a personalized device which is used to transcribe the speakers' utterances and then synthesize an equivalent spoken message. Research in this area has been in developing acoustic modeling algorithms that can provide adequate speech recognition performance for this population of disabled speakers when interacting with these communication aids.

Speaker modeling for assessing pathological speech: Statistical speaker modeling techniques have been used for assessment of acoustic properties of pathological voices that result from laryngeal cancer [4]. GMM models were trained using utterances from individuals after having gone through reconstructive surgery and then again after receiving speech therapy. Statistical measures were derived from these model parameters to determine the degree to which the speakers' voices differed from normal speakers after these procedures were performed. The goal is to use this approach as an automated means for measuring patients' progress in response to speech therapy.

Analyzing prosodic contours to assess intonation: Computer-assisted pronunciation training (CAPT) involves providing impaired or unimpaired speakers with training on the use of proper rhythm, intonation and stress patterns in speech [5]. Research in machine learning and feature analysis has been performed to develop automated measures of the accuracy of prosodic contours with respect to utterances from normal speakers.

The goal of the pronunciation verification techniques presented here is to provide automated assessment of speakers' performance as part of a CASLT scenario. The task domain for this scenario is presented in more detail in Section 4.1.3.

Human–computer interaction in CASLT

Automatic speech and language therapy requires consideration of the interaction between the patient and the automatic system. The applications described above must be implemented in association with a well-designed human–computer interaction (HCI) component in order to elicit natural utterances. There have been several projects involving the design of CASLT user environments and their evaluation of an actual user population. These projects are briefly summarized here.

An audio-visual pronunciation teaching software system, the INCO-Copernicus program of European Commission has been developed to help speech handicapped people to train their speech pronunciation for vowels and misarticulated fricative phones [8]. This system records the speech from the children. Then a series of acoustic features, such as loudness, pitch contours and spectral distribution along the time axis, are obtained through the acoustic speech processing. Based on these acoustic features, several amusing pictures are created and displayed to the children. These pictures attempt to draw the children's attention and encourage the children to improve their speech pronunciation.

Voice-input voice-output communication aid (VIVOCA) is another CASLT example proposed by the University of Sheffield, UK [7]. VIVOCA is a personalized and portable device which can be equipped by a disabled speaker. It serves as a communication aid to transcribe the speech from individuals with severe dysarthria, based on the ASR.

The project "Comunica" is another CASLT developed in Zaragiza, Spain [9]. The objective "Comunica" project is to develop a semi-automated system for providing interactive speech therapy to a larger population of impaired individuals and help professional speech therapists. There are three components involved in "Comunica": "Prelingua," "Vocaliza" and "Cuéntame." "PreLingua" teaches basic phonation skills to children with neuromuscular disorders. "Vocaliza" aims to train mainly the articulatory level of language. Finally, "Cuéntame" attempts to introduce impaired children population to language understanding. The pronunciation verification (PV) scenario presented in this article gives a contribution toward "Vocaliza" development.

4.1.2 A system for articulatory speech therapy

"Vocaliza" is a human–computer interactive tool designed for helping impaired children with neuromuscular disorders acquire articulation abilities in isolated words or short sentences [9]. From the human–computer interface (HCI) point

of view, “Vocaliza” provides an easy user interface that motivates children to enjoy the application while practicing their speech. This involves the use of text, images and sounds, which can reinforce concepts and correct pronunciations of words or sentences in utterances spoken by the user.

“Vocaliza” encourages the impaired child to utter a set of words preselected by the speech therapist or educator to focus on the special needs of the child. After receiving the speech utterances from the children, it segments the utterance and evaluates the accepted utterances using a word-level PV algorithm to display a grade as the final outcome of the game.

This article focuses on developing the PV scenario, which is used for helping children with neuromuscular disorders improving the phonological level of the communication skills. In [1], a word-level likelihood ratio-based utterance verification (UV) procedure is used as a measure of confidence to each hypothesized word in an utterance. This method obtains the ratio between the likelihood of the input utterance with respect to two models: one generated from nonimpaired speech and one adapted to impaired speech. The PV scenario proposed in Section 4.2 involves the same user interface as described above for verifying word-level pronunciations, but is concerned with phoneme-level pronunciation verification.

4.1.3 Task domain

The pronunciation verification scenario in Section 4.2.1 is evaluated using a novel Spanish speech corpus obtained from a population of children and young adults speakers enrolled in a special education program. This corpus was collected from the University of Zaragoza [10]. The speech uttered by children and young adults is recorded through the HCI of “Vocaliza” described in Section 4.1.2. The speech utterances involved in this novel speech corpus are isolated word based, where a small vocabulary set involved in these isolated words is designed for speech therapy purpose. The recorded utterances are separated into two categories: an impaired children corpus designed for evaluation purpose and an unimpaired children corpus for model development purpose.

Data collection scenario

The set of words involved in our PV scenario is given by the induced phonological register (RFI) [11]. RFI, while containing only 57 words is a powerful set of words for speech therapy as it contains examples of all the 25 phonemes, which

represent a reduced phone set from the 51 allophones described traditionally in the Spanish language [12], as shown in Table 4.1. The total number of syllables in the 57 words is 129 which gives an average of 2.26 syllables per word, with 90 different syllables. The total number of phonemes is 292 which gives an average of 5.13 phonemes per word. The isolated word utterances based on RFI and recorded from "Vocaliza" are sampled at 16 kHz and stored in 16 bits format. A wireless close-talking microphone is used in the recording. This way, comfortability of the speakers is guaranteed as they are not directly attached to the computer while obtaining the best speech quality possible with a high signal-to-noise ratio (SNR). While recording the utterances, the speaker will be told to

Table 4.1: 57 words in the induced phonological register (RFI) and their phonetic transcription.

Word	Transcription	Word	Transcription	Word	Transcription
árbol	/a-r-B-o-l/	boca	/B-o-k-a/	bruja	/B-r-u-x-a/
cabra	/k-a-B-r-a/	campana	/k-a-m-p-a-n-a/	caramelo	/k-a-r-a-m-e-l-o/
casa	/k-a-s-a/	clavo	/k-l-a-B-o/	cuchara	/k-u-tS-a-r-a/
dedo	/D-e-D-o/	ducha	/D-u-tS-a/	escoba	/e-s-k-o-B-a/
flan	/f-l-a-n/	fresa	/f-r-e-s-a/	fuma	/f-u-m-a/
gafas	/G-a-f-a-s/	globo	/G-l-o-b-o/	gorro	/G-o-rr-o/
grifo	/G-r-i-f-o/	indio	/i-n-d-j-o/	jarra	/x-a-rr-a/
jaula	/x-a-w-l-a/	lápiz	/l-a-p-i-T/	lavadora	/l-a-B-a-D-o-r-a/
luna	/l-u-n-a/	llave	/L-a-B-e/	mariposa	/m-a-r-i-p-o-s-a/
moto	/m-o-t-o/	niño	/n-i-J-o/	ojo	/o-x-o/
pala	/p-a-l-a/	palmera	/p-a-l-m-e-r-a/	pan	/p-a-n/
peine	/p-e-j-n-e/	periódico	/p-e-r-j-o-d-i-k-o/	pez	/p-e-T/
piano	/p-j-a-n-o/	pie	/p-j-e/	piña	/p-i-J-a/
pistola	/p-i-s-t-o-l-a/	plátano	/p-l-a-t-a-n-o/	playa	/p-l-a-L-a/
preso	/p-r-e-s-o/	pueblo	/p-w-e-B-l-o/	puerta	/p-w-e-r-t-a/
ratón	/rr-a-t-o-n/	semáforo	/s-e-m-a-f-o-r-o/	silla	/s-i-L-a/
sol	/s-o-l/	tambor	/t-a-m-B-o-r/	taza	/t-a-T-a/
teléfono	/t-e-l-e-f-o-n-o/	toalla	/t-o-a-L-a/	toro	/t-o-r-o/
tortuga	/t-o-r-t-u-g-a/	tren	/t-r-e-n/	zapato	/T-a-p-a-t-o/

repeat the same utterance if an excess of noise from the environment is captured. The speaker population providing these isolated word utterances from "Vocaliza" is described in the section "Speaker population." The corresponding speech corpus is used for our acoustic model training and PV evaluation.

Speaker population

The impaired children speakers involved in our PV scenario suffered developmental disabilities of different origins and degrees that affected their language abilities, especially at the phonological level. It was believed that all speakers suffered from a neuromuscular disorder so that all of them can be characterized as having dysarthria. Before recording the word utterances from "Vocaliza," none of the speakers were known to be hearing impaired or to have suffered from any abnormality or pathology in the articulatory or phonatory organs. All the impaired children were students at the Public School for Special Education "Alborada" in Zaragoza, Spain [1]. There were 14 impaired speakers, 7 males and 7 females, that participate in the recording. These 14 speakers were distributed in age from 11 to 21 years as shown in Table 4.2. All of these 14 speakers recorded 4 sessions of the 57 isolated words to make a total of 3,192 isolated word utterances. These provide 2 hours and 56 seconds of speech including silence. In order to reflect intra-speaker variability, every session was recorded on different days.

Table 4.2: Information about 14 impaired children.

Impaired speech corpus								
Code	**Gender**	**Age**	**Code**	**Gender**	**Age**	**Code**	**Gender**	**Age**
Spk01	F	13	Spk06	M	16	Spk11	F	19
Spk02	M	11	Spk07	M	18	Spk12	M	18
Spk03	M	21	Spk08	M	19	Spk13	F	13
Spk04	F	20	Spk09	F	11	Spk14	F	11
Spk05	M	18	Spk10	F	14	–	–	–

A reference corpus containing speech from unimpaired children speakers was collected in parallel with recording the impaired speech corpus. This corpus was intended to contain speech from children within the same age range as the children in the impaired population. Utterances from the unimpaired

children were collected from the same task domain. In order to reduce the mismatch between the impaired speech corpus and the unimpaired speech corpus, the unimpaired corpus was collected under the same acquisition scenario via "Vocaliza" as the impaired speech corpus. The same RFI vocabulary set and the same type of isolated word sessions were chosen for this unimpaired children recording scenario. The amount of speakers in this unimpaired corpus is 168, 73 males and 95 females ranging in age from 10 to 18 years. Every unimpaired speaker utters a single session of the 57 words in the RFI, which makes a total number of 9,576 isolated-word utterances in the unimpaired speech corpus. This includes 6 hours, 17 minutes and 43 seconds of speech including silence, which gives an average of only 2.37 seconds in length for each word utterances.

Pronunciation labeling by nonexpert human labelers

In order to evaluate the PV algorithm, a simple manual system for phoneme-level pronunciation labeling is devised. In this process, every phoneme in the impaired children corpus is labeled by three independent nonexpert labelers. The phonemes in the isolated word utterances produced by impaired children speakers are labeled as having been either deleted by the speaker, mispronounced and therefore substituted with another phoneme, or correctly pronounced. In the end, the final label for the phoneme was chosen by consensus among the labelers.

Pairwise inter-labeler agreement for this manual labeling task is 85.81%. This agreement raises to 89.7% when considering only a binary decision: correct versus incorrect (deletions plus substitutions). This consistent labeling avoids the problems of a subjective speech quality measurement that would have required very experienced labelers. The percentage of phoneme occurrences that were labeled as correct is 82.4%, while 10.3% of the phoneme occurrences are substitutions and 7.3% are deletions. The label distributions for each speaker are shown in Table 4.3. The total number of labelers was 10, all of them with expertise in the fields of speech technologies or phonetics.

4.1.4 Summary

This chapter has provided a brief overview of CASLT technology and examples of interactive environments for speech therapy. The "Vocaliza" environment used for eliciting utterances from children as part of articulation training was

Table 4.3: Labeling results per speaker: Rate of deletions, substitutions and correct phonemes.

				Human labeling			
Speaker	**Del.**	**Subs.**	**Corr.**	**Speaker**	**Del.**	**Subs.**	**Corr.**
#01	0.2%	0.9%	98.9%	#08	13.1%	17.7%	69.2%
#02	9.2%	12.4%	78.4%	#09	2.9%	5.3%	91.8%
#03	0.7%	4.6%	94.8%	#10	8.4%	13.1%	78.5%
#04	1.1%	2.1%	96.8%	#11	2.1%	4.7%	93.2%
#05	17.4%	26.1%	56.5%	#12	11.7%	14.0%	74.3%
#06	0.2%	0.5%	99.3%	#13	25.9%	30.5%	43.6%
#07	5.6%	7.3%	87.1%	#14	3.9%	5.1%	91.0%

described. Finally, the speech corpus collected under the "Vocaliza" scenario was described along with the annotation strategy that was used to provide the reference labels for the experimental studies given in Section 4.2 and 4.4.

4.2 Automatic pronunciation verification

This chapter presents a multiple pass approach to verifying the correctness of pronunciations in utterances from disabled speakers. Pronunciation verification (PV) will be presented as a problem of verifying the claim that a particular word or subword unit in an utterance has been correctly pronounced. PV will be performed by detecting phoneme-level mispronunciations in utterances from an impaired speaker population.

The PV scenario presented here is similar to the problem of PV for automated language learning and language skills evaluation applications [13]. The test corpus involved in our PV scenario and described in Section 4.1.3 is recorded from members of young impaired speaker population suffering from neuromuscular disorders of varying severity. This distinguishes the speaker population from language learners who are assumed to be nonproficient in the given language but at the same time are assumed to not suffer from any speaking impairments. Hence, speech obtained from the impaired population of speakers is more likely to be significantly affected at multiple levels than

speech from unimpaired speakers. The disorder effects can be observed in frame-level spectral characteristics, segment-level coarticulation, lexical-level pronunciation rules and super-segmental prosodic contours [14]. While there has been considerable effort made to model how these disorders are reflected in the underlying articulatory dynamics of speech production [15], the techniques described here are based on a posteriori probabilities derived from HMM-based ASR. Phone-level measures of confidence are derived from the acoustic speech utterance and are used to define a decision rule for accepting or rejecting the hypothesis that a phoneme was mispronounced.

This chapter is organized into two parts. First, Section 4.2.1 presents a procedure for deriving phone-level confidence measures based on posterior probabilities derived from phone lattices. In order to effectively obtain a confidence measure in detecting variability arising from the mispronunciations produced by the impaired speaker population, it is necessary to reduce the influence that other sources of variability have on the confidence measure. Section 4.2.1 also presents various ASR techniques to limit the effects of inter-speaker variability and task variability is described. Finally, Section 4.2.2 presents the experimental study.

4.2.1 Pronunciation verification scenario

This section describes the pronunciation verification (PV) scenario which verifies the phone-level pronunciation accuracy for a given utterance pronounced by an impaired speaker [16, 17]. The impaired speaker population has been summarized in the section "Speaker population." First, the section "Phoneme-level confidence measure" describes how to generate the phone-level confidence scores for making the PV decision, based on the ASR and confusion network (CN) techniques. Second, the section "Reducing variability through model adaptation" shows how to apply the acoustic model adaptation techniques using maximum a posteriori (MAP) approach and maximum likelihood linear regression (MLLR) approach [18, 19]. The objective of MAP/MLLR adaptation is to construct a more robust acoustic model from a task-dependent speech corpus, the unimpaired speech corpus described in the section "Speaker population." Finally, the section "Nonlinear mapping of posterior probabilities" introduces a nonlinear mapping idea which maps the lattice-based posterior confidence score to a more robust confidence measure.

Phoneme-level confidence measure

In the phoneme PV scenario, it is assumed that the "target" word sequence and its baseform phonetic expansion are known. For the experimental study described in Section 4.2.2, it is assumed that the input test utterance corresponds to an isolated word from the RFI described in the section "Data collection scenario." The corresponding baseform phone string, q_n, $n = 1, \ldots, N$, is assumed to be known. PV in this context simply refers to obtaining confidence measures for each phoneme in the baseform expansion and applying a decision rule for accepting or rejecting the hypothesis that a given phone was correctly pronounced. This process, as depicted in Figure 4.1, is performed in two steps.

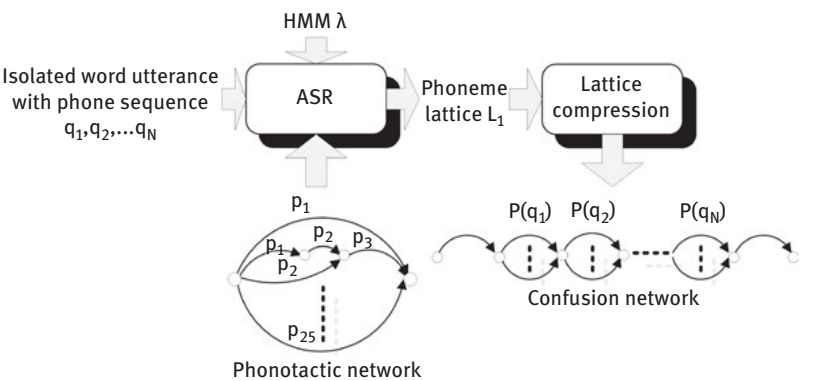

Figure 4.1: Confusion network-based posterior probability estimation.

First, phoneme recognition based on the ASR scenario is performed on the given isolated word utterance. The ASR search is constrained using a network that describes the potential pronunciations that might be expected from an unimpaired speaker. This network could potentially be created from the syllabification rules of the language or be trained from observed pronunciations decoded from the population of unimpaired speakers. While rule-based constraints are currently being investigated, simple N-gram phonotactic constraints are applied here. Specifically, a bigram phonotactic model is trained from baseform phonetic expansions obtained from an 8 million word subset of the Spanish-language section of the Europarl speech corpus [20], containing transcriptions from several sessions in the European Parliament translated to different European languages. This phonotactic bigram model is also used for constraining the search as outlined in Figure 4.1. In order to observe how the phonotactic model would affect the PV performance, a

simple unconstrained, zerogram-based, phone network is also used in the experimental study presented in Section 4.2.2

Second, a confusion network, as depicted in Figure 4.1, is created using a lattice compression algorithm. A phoneme lattice L_1 that contains phone labels and their associated acoustic and language probabilities on the arcs is generated by the ASR decoder. The posterior phone probabilities $P(q_n)$, $n = 1, \ldots, N$, appearing on the transitions of the CN are obtained by forming the sum of probabilities of all the paths passing through the target phoneme arc in the lattice and normalizing by the sum of probabilities of all the paths in the lattice L_1. Compared with the original phoneme lattice L_1, the compact structure and the ordering properties of the CN facilitate efficient evaluation of posterior-based confidence measures for verifying phone pronunciations. The posterior probabilities for the baseform phones in the target phone string are obtained by aligning the target phone string with the phoneme lattice L_1. These posterior phone probabilities are used as phone-dependent confidence scores. Comparing these confidence scores with a decision threshold can be served as a decision criterion for verifying whether a given target phoneme has been correctly pronounced.

Reducing variability through model adaptation

Acoustic model adaptation techniques such as MAP or MLLR adaptation can be applied in the PV scenario in order to reduce the effects of other sources of variability. These may include all sources of variability outside of those introduced by the speech disorders existing among the disabled speaker population. For example, physiological and dialect differences among speakers, differences in microphones, and differing acoustic environments can all influence the ability to detect mispronunciations in the PV scenario.

The baseline ASR system and the adaptation scenarios included in our experimental study are introduced here. Baseline HMM models are trained from the Spanish language Albayzin speech corpus [21], which includes 6,800 sentences with 63,193 words. This corpus contains 6 hours of speech including silence; however, only 700 unique sentences are contained in the corpus. Because of this lack of phonetic diversity, it is difficult to train context-dependent acoustic models that will generalize across task domains. For this reason and because of the simplicity of this small vocabulary task, context-independent monophone models are used here. In all experiments, 25 monophone-based context-independent HMMs are used which consist of 3 states per phone and 16 Gaussians per state.

MFCC observation vectors along with their first and second difference coefficients are used as 39-dimensional acoustic features.

As mentioned in the section "Phoneme-level confidence measure," phoneme-level PV is performed on isolated word utterances from the RFI 57 word vocabulary were each utterance is recorded from an impaired children speaker. The speaker property and vocabulary involved in this impaired task is different from the Albayzin speech corpus used for the baseline HMM training. In order to obtain a more robust task-dependent acoustic model, the unimpaired corpus described in the section "Speaker population" is used to perform an MAP-based [18] and MLLR transform-based [19] adaptation of the Gaussian mean vectors. The MLLR adaptation involved two regression classes, one for the silence and the other one for all the nonsilence 25 phonemes. The reason for combining both MAP and MLLR adaptation is based on their complementary behavior [22]. Simply stated, MAP adaptation is performed independently on the means associated with distributions assigned to each phoneme classes. If the phonemes are well represented in the adaptation data, improved acoustic models can be obtained using MAP. On the other hand, MLLR adaptation is applied as a linear transformation to the mean vectors of the distributions. MLLR has the ability to benefit from observation vectors belonging to all phoneme classes to adapt those models that are not well represented in the adaptation data. As a result, simply combining the two adaptation procedures can result in complementary performance increases.

The MAP/MLLR task-dependent adaptation corpus includes 6,840 adaptation utterances spoken by 120 unimpaired speakers from the 168 children and young adults in the section "Speaker population." Each unimpaired speaker provides 57 RFI isolated word utterances where all the words are assumed to be accurately pronounced. The adaptation corpus contains 4.5 hours of speech including silence.

Supervised speaker-dependent adaptation for each of the 14 test speakers summarized in Table 4.3 is also performed using an MLLR-based transform applied to the Gaussian means of the task-dependent HMM. For each speaker, a single MLLR transform matrix is estimated and applied for speaker adaptation. The speaker-dependent MLLR adaptation data consists of 57 isolated word utterances or 2.2 minutes of speech for each of the test speaker. The remaining 2,394 impaired speaker utterances, three sessions of 57 isolated word utterances for each impaired speaker, are used for evaluation. The supervised speaker-dependent MLLR transformation is then applied prior to verifying the phoneme-level pronunciation of the impaired speech utterances.

Even a supervised speaker adaptation paradigm is problematic for the impaired children population since the utterances contain many phonemes that

are known to be mispronounced or deleted. It is possible, however, to modify the adaptation procedure to incorporate the pronunciation labels obtained from the human labelers. This was done for MLLR adaptation to the impaired speakers by creating two regression matrices. One regression matrix was estimated from occurrences of phonemes in the adaptation data that were labeled as being correctly pronounced and another matrix was estimated from occurrences of phonemes that were labeled as being incorrectly pronounced. During recognition, only the first matrix was applied to transforming the mean vectors of all model distributions. Phonemes in the adaptation data that were labeled by the human labelers as having been deleted by the speaker were simply deleted from the reference transcription during adaptation. This procedure, referred to later as "Label Supervised MLLR," is similar in spirit to unsupervised adaptation procedures that rely on acoustic confidence measures [23]. These procedures apply varying weight to regions of an adaptation utterance to reflect the relevance of the region to the distributions being adapted. It will be shown in Section 4.2.2 that significant performance improvement can be obtained by exploiting the supervision provided by the human labelers.

Nonlinear mapping of posterior probabilities

A nonlinear transformation can be performed to map the lattice posterior probabilities to phone-level confidence measures. There are two motivations for this: The first motivation stems from the fact that all of the PV techniques presented here are evaluated in terms of their ability to predict the labels defined by the labeling scheme defined in the section "Pronunciation labeling by nonexpert human labelers." The decision made by an expert as to whether a given occurrence of a phone is classified as being "mispronounced" rather than as a "pronunciation variant" will always have a subjective component. The labeling scheme presented in the section "Pronunciation labeling by nonexpert human labelers" is important because it addresses the trade-off between the need for a consistent, repeatable and easily implemented labeling strategy against the need for an accurate characterization of the quality of pronunciation of a given phoneme. There is no guarantee, however, that the posterior probabilities estimated as shown in Figure 4.1 will always be accurate predictors of these labels.

The second motivation is the fact that there is a great deal of prior information available in this PV scenario. This includes knowledge of the target word, the target phone and the position of the phone within the word. This prior information can be combined with the phone-level posterior probability using

one of many possible fusion strategies to better predict the human derived labels.

In the experimental study described in Section 4.2.2, the parameters of a multilayer perceptron with the above parameters as input are trained to implement a nonlinear transformation. Backpropagation training is performed for a network with 47 hidden nodes and with input activations which include the phone-level posterior probabilities, indicator variables corresponding to each of the phone labels, and indicator variables corresponding to the word labels. The network is trained with the human derived pronunciation labels serving as targets. PV is performed using the output activations obtained from this network on test utterances with the same kinds of input parameters as the ones used in the training phase.

4.2.2 Experimental study of PV

This section presents an experimental study performed to evaluate the ability of the PV techniques presented in Section 4.2.1 to detect mispronunciations in utterances obtained from impaired speakers. Verification performance is measured using utterances from the 14 speaker population of impaired speakers as a test corpus. For each phoneme in the baseform phonetic expansion of a word, the task is to verify the claim that the pronunciation of that phone is correct according to the human labels assigned using the labeling scheme described in the section "Pronunciation labeling by nonexpert human labelers." Since this is in fact a detection problem, the performance is presented using detection error trade off (DET) curves and the equal error rate (EER) measure. The EER is computed by applying a decision threshold to the phone-level confidence scores and identifying the threshold setting where the probability of false acceptance is equal to the probability of false rejection. The phone-level confidence scores are computed based on the scenario described in the section "Phoneme-level confidence measure."

The performance relating to several issues will be considered. First, the effect of the adaptation strategies for reducing task-dependent (TDEP) and speaker-dependent (SDEP) variability will be considered. Second, the effect of the applied phonotactic bigram constraints in decoding will be evaluated with respect to an unconstrained (zerogram) decoding. Third, the performance of the nonlinear neuron network (NN)-based mapping procedure will be presented.

First, The PV verification performance is found to vary across phoneme classes. For example, when the results in Table 4.4 are reported separately for phonemes classified as vowels and nonvowels, the performance for the vowel

Table 4.4: Phone detection performance measured as the equal error rate (EER) for task-independent (TIND) baseline, task-dependent (TDEP) MAP/MLLR adaptation, speaker-dependent (SDEP) MLLR adaptation with and without label supervision and SDEP neural network (NN)-based nonlinear mapping.

Phone-level verification performance (EER)		
Adaptation scenario	**zerogram**	**bigram**
TIND HMM (baseline)	25.3%	22.2%
TDEP MAP/MLLR adaptation	19.7%	18.4%
SDEP MLLR adaptation	18.3%	17.1%
SDEP label supervised MLLR adaptation	17.2%	16.2%
SDEP NN mapping	14.9%	N/A

class is considerably worse than the nonvowel class. For the TDEP MAP adaptation case using the zerogram network, the vowel class EER is approximately 18% higher than the EER obtained for the nonvowel class. Vowels represent 44% of the total phoneme occurrences in the corpus. This surprising difference in EER is partly due to the human labeling strategy. Rather forgiving subjective judgments were made by the labelers when deciding whether a given utterance contained a "pronunciation variant" of a phoneme as opposed to a labeled mispronunciation error. This results in many cases where the decision threshold defines a phoneme instance to be mispronounced when the reference label indicates the phoneme was correctly pronounced. There is less ambiguity in human labelers' decisions for the labeling of deletion errors. The higher EER observed for vowels results from the fact that mispronunciation errors are more common for vowels and deletion errors are more common for nonvowels.

Table 4.4 presents the global PV performance under different experimental conditions where each is delineated in the first column of the table. The second and third columns of Table 4.4 display the performance in EER for the zerogram and bigram recognition networks respectively. The results in Table 4.4 are obtained on a test set consisting of 2,394 utterances and 12,264 monophone test trials. These include 10,083 phonemes labeled as being correctly pronounced and 2,128 labeled as incorrectly pronounced. The 2,128 "incorrect" test trials correspond to phoneme instances that have been either mispronounced by the test speaker (substituted for another phoneme) or deleted altogether.

There are several observations that can be made from the results given in Table 4.4. First, from the first row of the table, it is clear that the EER for verifying phone-level pronunciation task-independent HMM models trained from the Albayzin speech corpus is fairly high. An EER of over 25% is obtained when no phonotactic constraints are applied in decoding. An EER of 22% is obtained when the bigram network is used. The second row of the table shows that MAP/MLLR adaptation of the HMM to the corpus of unimpaired children and young adults speaking utterances of the same vocabulary words results in approximately 20% decrease in EER. This rather significant improvement is due largely to the significant mismatch in speaker characteristics that exists between the largely adult speaker population in the Albayzin corpus and the unimpaired younger speaker population in the adaptation corpus.

The age of the children and young adults in the corpora used here ranged from 11 to 21 years old. The age of members of the speaker population in the Albayzin training corpus ranged from 19 to 64 years old. The ages of one third of the speakers in the Albayzin corpus were between 39 and 64, approximately two thirds of the speakers were between the ages of 23 and 38, and less than 1% of the speakers in the Albayzin corpus were less than 22. Hence, the degree of overlap between the ages of the two speaker populations was extremely small.

The third row of Table 4.4 shows that speaker-dependent MLLR adaptation of the TDEP HMM models using 57 utterances from each test speaker results in approximately 7% decrease in EER. Note that the speaker-dependent adaptation data includes both correctly pronounced phonemes and phonemes that were mispronounced by the impaired speakers. Including the mispronounced phonemes in the adaptation data may limit the potential performance improvements that are achievable in this scenario. The fourth row of Table 4.4 displays the result after performing SDEP adaptation using the "label supervised" MLLR adaptation described in Chapter 4.2.1. The corresponding results show that when the MLLR regression matrix is trained only from phoneme segments that have been labeled as being correctly pronounced, the relative reduction in EER increases from 7% to 12% with respect to the TDEP EER.

The fifth row of the table shows the effect on performance when the same utterances used for MLLR adaptation are instead used to train the NN based mapping described in Chapter 4.2.1. This results in a substantial 18% reduction in EER with respect to the TDEP case. Finally, comparing the EER displayed in the second and third columns of Table 4.4, the bigram phonotactic constraints result in a reduction in EER rate between 7% and 12%.

Recall that the performance of phoneme-level pronunciation verification presented in Table 4.4 is measured for two different subsets of the incorrectly

pronounced test trials. These included instances where, first, the target phoneme was deleted by the impaired speaker and, second, where the target phoneme was mispronounced by the target speaker and substituted with another phoneme. The performance measured for these two subsets of the incorrect utterances are shown in Table 4.5 and Figure 4.2. These results show that for all conditions, verifying the hypothesis that a phoneme was deleted by the speaker is easier than verifying that a phoneme was mispronounced. There are many potential explanations for this behavior. One explanation may relate to the strategy followed by the human labelers in assigning correct and incorrect pronunciation labels to the phonemic expansions of words in the test corpus. Rather forgiving subjective judgements were made when deciding whether a given utterance contained a "pronunciation variant" of a phoneme as opposed to a mispronunciation. This may result in many cases where the decision threshold defines a phoneme instance to be mispronounced when the reference label indicates the phoneme was correctly pronounced.

Table 4.5: Phone detection performance comparison for baseline, task-dependent MAP adaptation, and speaker-dependent MLLR adaptation with and without label supervision performed on different subsets of the test data. The "All" case includes all the 12,264 test trials (10,083 correct, 2,181 incorrect). "Deletion Errors" includes only the 943 incorrect test trials that correspond to deleted phonemes. "Mispronunciation Errors" includes only the 1,238 incorrect test trials corresponding to mispronounced phonemes. Zerogram, or unconstrained, network used in ASR.

Comparison in equal error rate (EER)			
Adaptation	**All**	**Deletion**	**Mispron.**
Scenario	Error	Errors	Errors
TIND (baseline)	25.3%	23.4%	26.8%
TDEP MAP/MLLR adaptation	19.7%	17.6%	21.7%
SDEP MLLR adaptation	18.3%	15.4%	20.1%
SDEP Label supervised MLLR adaptation	17.2%	14.4%	18.8%

The issue of the statistical significance of differences between measures based on false accept rates and false reject rates has been addressed in the literature [24, 25]. However, there is no significant test for these applications that has become widely accepted in the speech and language community, so the results of these significance tests can be difficult to interpret. By any test, one would assume that

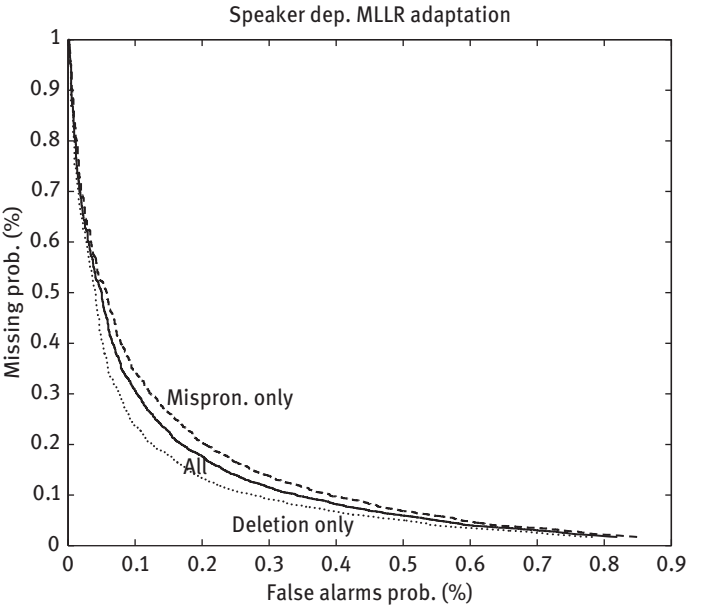

Figure 4.2: DET curve comparison for different test cases. Zerogram, or unconstrained, network used in ASR.

the EER difference of 1.3% shown for rows two and three of Table 4.4 are at best barely statistically significant. This equal error rate point corresponds to a difference of 130 false rejection trials out of 10,083 correctly pronounced phonemes and 27 false acceptance trials out of 2,128 incorrectly pronounced phonemes. Without computing confidence intervals on these outcomes, one cannot conclude with any certainty that the resulting estimate of the difference in error rates for this case is significant.

Figure 4.3 displays the pronunciation verification performance over the 2,394 utterance test set in the form of DET curves. The DET curves labeled TIND, TDEP and SDEP in Figure 4.3 correspond to the systems whose zerogram EER results are given in rows two through four of Table 4.4. Note that the performance characteristics are well behaved in that the same rank order of performance is achieved by the three systems at all operating points.

While the NN-based nonlinear mapping was shown in Table 4.4 to provide a substantial reduction in EER, the scenario followed for the system in Table 4.4 involved using speaker-dependent data in training the NN. In order to investigate the effect of this mapping in a speaker-independent scenario, the 14 speaker training set was divided in half. Utterances from the first seven speakers were

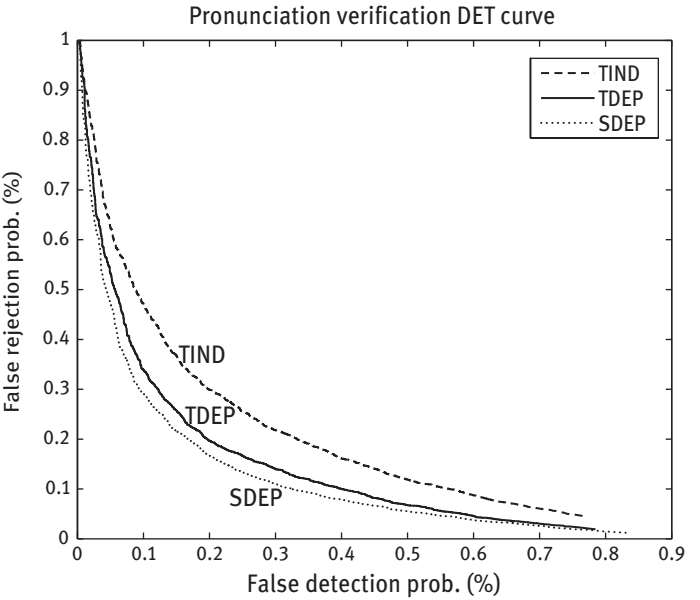

Figure 4.3: DET curves displaying phone-level verification using baseline, task adapted and speaker adapted HMM models. Zerogram, or unconstrained, network used in ASR.

used for training the NN-based mapping and utterances from the second set of seven speakers were used as a test set. The results for this revised training and testing scenario on the reduced test set are displayed in Table 4.6. It is clear from Table 4.6 that the impact of the NN-based mapping in reducing the EER relative to the TDEP performance is far less for the speaker-independent training of the NN than it is for the speaker-dependent case reported in Table 4.4.

Table 4.6: Task-dependent (TDEP) phone detection performance using unconstrained zerogram network obtained with and without speaker-independent neural network (NN)-based nonlinear mapping.

TDEP verification performance using reduced test set (EER)	
TDEP MAP/MLLR adaptation	20.5%
TDEP + NN mapping	19.7%

4.2.3 Summary

A simple phoneme-level confidence measures based on CN posterior probabilities were found to provide reasonable performance in detecting mispronunciations in utterances taken from the impaired children corpus described in Section 4.1.3. However, after adapting acoustic models and performing nonlinear mapping of the CN posteriors as described in the section 4.2.1, "Reducing variability through model adaptation," a relative 40% improvement in detection performance was obtained. The absolute EER can be reduced from 25.3% to 14.9%. In additional, comparing with the unconstraint network, Table 4.4 shows that the bigram phonotactic constraints can result in a reduction in EER rate between 7% and 12%. The results obtained here demonstrate that the ability to detect mispronunciations resulting from neuromuscular disorders can be significantly improved by reducing the effects of other sources of variability in speech. It is believed that the confidence measures used in this system achieve a performance that is close to that necessary to provide useful feedback to impaired speakers in language learning and speech therapy applications.

4.3 Subspace Gaussian mixture models

One of the important issues resulting from the experimental study of pronunciation verification (PV) in Section 4.2. is the importance of distinguishing between phonetic variation arising from speech impairments and variation arising from natural coarticulation in speech from unimpaired speakers. To address this issue, a modeling formalism is investigated in [26, 27] and described in this chapter that provides an efficient subspace decomposition of the acoustic space. In doing so, the goal is to improve the PV performance on the tasks presented in Section 4.2. In this chapter, a subspace-based Gaussian mixture model (SGMM) is introduced. The important aspect of the SGMM is its representation of state-level acoustics as simple projections in multiple subspaces. An experimental study will be presented in Section 4.4, where the SGMM acoustic model is applied to the phone-level PV task that was described in Section 4.2.

In this chapter, the SGMM will be evaluated in terms of the ASR word error rate on a standard read speech task. A brief description of the SGMM will be presented in Section 4.3.1. In Section 4.3.2, a comparison between the conventional continuous density hidden Markov model (CDHMM) acoustic model described in [28, 29] and SGMM performance will be evaluated in terms of word

recognition performance. In addition, phone recognition performance is measured for the CDHMM and SGMM and presented in Section 4.3.3.

4.3.1 A description of subspace-based models

A brief description of the SGMM acoustic modeling framework and its application in the ASR decoder is presented here. First, the section "The single substate per state-based SGMM structure" presents the simplest form of the SGMM structure, where each state in the SGMM is represented by a single projection vector. Second, the section "The multiple substate per state-based SGMM structure" presents a more general parameterization of the model and discusses practical issues involved in parameter estimation. Third, the section "The observation likelihood computation with the SGMM" describes a method for efficient computation of local likelihood in the SGMM. Finally, the section "SGMM initialization and training" discusses additional practical issues including SGMM parameter initialization and issues associated with updating parameters in training.

The single substate per state-based SGMM structure

The CDHMM observation density $P(x_t|s_t = j)$ for an F-dimensional feature vector, x_t, and given state j is formed from a mixture of state-dependent diagonal covariance Gaussians.

Assume the state index at time t, $s_t = j$, is given for an F-dimensional feature vector x_t, the probability density function for an I^j component Gaussian mixture model λ_j is defined as

$$b_j(x_t) = P(x_t|s_t) = \sum_{i=1}^{I^{(j)}} w_i^{(j)} P_i^{(j)}(x_t)$$

where I^j represents the number of Gaussian mixtures associated with the state j. eq. ( ) corresponds to a weighted linear combination of I^j state-dependent unimodal Gaussian densities, $P_i^{(j)}(x_t)$, where i represents the mixture index, $i = 1, \ldots, I^{(j)}$. Each Gaussian density $P_i^{(j)}(x_t)$ is parameterized by a F-dimensional mean vector μ_i^j and a $F \times F$ dimensional covariance matrix $\Sigma_i^{(j)}$. The mixture weights, $w_i^{(j)}$, satisfy the constraint $\Sigma_{i=1}^{I^j} w_i^j = 1$ The likelihood $P_i^j(x_t)$ is given by

$$P_i^{(j)}(x_t) = 1/(2\pi)^{F/2} |\Sigma_i^{(j)}|^{1/2} e^{-\frac{1}{2}(x-\mu_i^{(j)})^* \Sigma_i^{(j)} - 1(x-\mu_i^{(j)})},$$

where the notation $\{.\}^*$ indicates the transpose operation.

$$b_j(x_t) = P(x_t|s_t = j) = \sum_{i=1}^{I^{(j)}} w_i^{(j)} P_i^{(j)}(x_t) \tag{4.1}$$

The observation density in eq. (4.1) is based on a set of I shared full covariance Gaussian densities with mean $\mathbf{m}_i$ and full covariance matrix Σ_i, where $i = 1, \ldots, I$. Typically, I in eq. (4.1) may be on the order of 100 to 1,000. The state-dependent mean vector, $\boldsymbol{\mu}_{ji}$, for state j is a projection into the ith subspace defined by linear subspace projection matrix $\mathbf{M}_i$,

$$\boldsymbol{\mu}_{ji} = \mathbf{m}_i + \mathbf{M}_i \mathbf{v}_j, \tag{4.2}$$

where $\mathbf{v}_j$ is a state-dependent projection vector. In eq. (4.2), $\mathbf{m}_i$ corresponds to the mean vector for the ith shared Gaussian density [26, 30]. The parameterization of $\mathbf{m}_i$ as a state-independent offset of the mean vectors provides a minor departure from the formalization given by [26, 27]. The impact of introducing $\mathbf{m}_i$ in eq. (4.2) for the recognition performance will be addressed in Sections 4.3.2 and 4.3.3. The term $\mathbf{v}_j$ is the projection vector associated with the state j. The global mean projection matrices $\mathbf{M}_i$ in eq. (4.2) is of dimension $F \times S$ where S is the dimension of the subspace associated with the global mean vectors $\boldsymbol{\mu}_{ji}$. The state-specific weights, w_{ji}, in Equation (4.1) are obtained from the state projection vector, $\mathbf{v}_j$, using a log-linear model,

$$w_{ji} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_j}{\sum_{k=1}^{I} \exp \mathbf{w}_k^T \mathbf{v}_j}. \tag{4.3}$$

Using the exponential function in eq. (4.3) provides a nondecreasing auxiliary function in training [31]. Similar to the case for the CDHMM, there is no closed form solution for estimating parameters in the SGMM. An EM-based training procedure, is summarized from [27] for the SGMM in the section "SGMM initialization and training."

As shown in eqs. (4.2) and (4.3), most of the SGMM parameters are shared across the states. Intuitively, the state-independent parameters $\{\mathbf{M}_i, \mathbf{w}_i\}$, $i = 1, \ldots, I$, correspond to the shared subspace parameters. The state-dependent projection vectors $\{\mathbf{v}_j\}$, $j = 1, \ldots, J$, represent projections within these subspaces. The dimension, S, of the state-dependent vectors $\{\mathbf{v}_j\}$ is chosen to be $S = F$ in this work. Thus, the observation probability for a given state can be described by mapping the state projection vector $\mathbf{v}_j$ to the GMM means and weights. The state-level covariances correspond to the global state-independent covariances $\{\Sigma_i\}$, $i = 1, \ldots, I$. The SGMM parameterization relies on a large number of state-independent shared parameters $\{\mathbf{M}_i, \mathbf{w}_i, \Sigma_i\}$, and a relatively small amount of

state-dependent parameters $\{\mathbf{v}_j\}$. As a result, a robust acoustic model can be expected from the SGMM structure, even when the amount of speech training data is limited.

The shared I full covariance Gaussian densities with mean $\mathbf{m}_i$ and full covariance matrix Σ_i form a global GMM which is similar to the universal background model (UBM) described in [32]. In speaker recognition, the UBM is commonly used as the prior model for running the speaker model adaptation [33].

The multiple substate per state-based SGMM structure

The notion of a substate can be used to provide additional flexibility in the parameterization of the SGMM [26, 31]. In this case, the distribution for the observation in state j is a weighted combination of densities

$$P(x_t|s_t = j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} N(x_t; \boldsymbol{\mu}_{ji}, \Sigma_i), \tag{4.4}$$

where c_{jm} is the substate weight associated with the substate index m in state j, and M_j is the number of substates in state j. The means and mixture weights are now obtained using substate projection vectors, $\mathbf{v}_{jm}$,

$$\boldsymbol{\mu}_{jmi} = \mathbf{m}_i + \mathbf{M}_i \mathbf{v}_{jm} \tag{4.5}$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{k=1}^{I} \exp \mathbf{w}_k^T \mathbf{v}_{jm}}. \tag{4.6}$$

The basic motivation for parameterizing the state j with multiple substates is to increase the model resolution for those states having sufficient occurrences in the training data. Additional substates for those states are created based on the observed accumulated zero-order statistics for that state. In the section "SGMM initialization and training," a procedure of obtaining these substate projection vectors $\mathbf{v}_{jm}$ will be described in more detail.

The observation likelihood computation with the SGMM

Comparing the observation likelihoods in the SGMM and the CDHMM, the computational complexity associated with the SGMM is higher than that associated with the CDHMM. There are two reasons for this. First, all the Gaussians involved in a

SGMM have full covariance rather than diagonal covariance matrices. Second, the total number of Gaussians associated with each state is much higher for the SGMM since all Gaussians are shared amongst the states. The technique used for solving this issue is Gaussian preselection. Since summing over all the I Gaussian probabilities as shown in eq. (4.1) is time-consuming, one can save computation by assuming that the observation likelihood computation is dominated by a relatively small number of Gaussians. In the SGMM experimental studies presented in this article, an appropriate value for this number was empirically determined to be 10.

Gaussian preselection is performed for each observation vector, x_t, and involves finding a set of Gaussians that are likely to dominate the computation of eq. (4.1). This set of Gaussians is identified in the following steps. First, all the Gaussian full covariances in the SGMM are converted from full covariances into diagonal covariances, $\Sigma_i^{(diag)}$. Second, an observation likelihood computation is performed based on these diagonal covariances. Because all the Gaussian covariances become diagonal, this likelihood can be computed quickly. Among all the Gaussian mixture $i = 1, \ldots, I$, the Top-10 Gaussian mixtures providing the Top-10 highest $N(x_t; \mathbf{m}_{ji}, \Sigma_i^{(diag)})$ can be obtained for each state j and each input feature vector x_t. Finally, instead of using all the I full covariance-based Gaussians, only the Top-10 preselected full covariance-based Gaussians, for each state j and each input feature vector x_t, will be used for computing $P(x_t|s_t = j)$ in eq. (4.1). The observation likelihood computation based on the Gaussian preselection is obviously just an approximation of eq. (4.1). The experimental studies show that, when using the Top-10 preselected Gaussian mixtures instead of using all the 256 Gaussian mixtures for computing the observation likelihood, a less than 0.15% relatively word error rate increase is obtained on a standard read speech task. However, this resulted in close to an order of magnitude reduction in decoding time.

SGMM initialization and training

The SGMM formalization and training procedure were originally proposed by Povey et al. [26, 27]. The contributions made in this work toward the practical implementation of SGMMs are discussed in this section. The first contribution is a method for parameter initialization based on the use of joint HMM state and GMM mixture posterior probabilities. The second contribution is a method for initializing SGMM substate projection vectors through a method of binary splitting of substates. This section will describe SGMM training, SGMM sub-state splitting and the joint state/mixture posteriors method for SGMM initialization.

Training

The SGMM structure is inherited from an initial CDHMM and the shared Gaussians are obtained from a universal background model (UBM). The same training corpus used for training the CDHMM and UBM is used for training the SGMM. The SGMM shared Gaussian parameters $\{\mathbf{m}_i, \Sigma_i\}$ are initialized from the speech UBM directly. The state list and state transition probabilities are inherited from the baseline CDHMM, and are not updated during the SGMM training. The rest of the subspace model parameters, $\mathbf{M}_i$, $\mathbf{w}_i$ and $\mathbf{v}_j$, are initialized from a flat start initialization, where $\mathbf{M}_i$ are initialized as identity matrices, and $\mathbf{w}_i$ and $\mathbf{v}_j$ are initialized as the zero vectors. The initial SGMM parameters, $\{\mathbf{m}_i, \Sigma_i, \mathbf{M}_i, \mathbf{w}_i\}$ for each mixture i, and $\{\mathbf{v}_j\}$ for each state j, can be updated through the EM training iterations [31].

There is no closed form solution for the maximum likelihood (ML) estimation of SGMM parameters. The same EM approach described in [18, 34, 35] is required to progressively obtain a ML estimate of SGMM parameters, given an initial estimate of the parameters. The auxiliary function for optimizing the SGMM parameters can be expressed as

$$Q(\Phi, \bar{\Phi}) = constant + \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_{ji}(t) log P(x_t, i|s_t = j). \tag{4.7}$$

The term $P(x_t, i|s_t = j)$ represents the contribution of the observation likelihood from state j and Gaussian mixture i at time t, which is shown in eq. (4.1).

Except for $\{\mathbf{M}_i, \mathbf{w}_i, \mathbf{v}_j\}$, the derivation of EM procedure for updating the SGMM parameters is straightforward. Different parameters will be updated separately on different EM iterations, following the order: $\{\mathbf{v}_j\}$, $\{\mathbf{w}_i\}$, $\{\mathbf{M}_i\}$, $\{\mathbf{m}_i\}$ and $\{\Sigma_i\}$. The details of the training procedure for the SGMM parameters $\{\mathbf{M}_i, \mathbf{w}_i, \mathbf{v}_j\}$ are described in [26, 27], which is also summarized in [27].

Substate splitting

When the SGMM is parameterized using multiple substates per state, the substates for a given state j are created based on the binary splitting of the initial state projection vector $\mathbf{v}_{jm}$:

$$\mathbf{v}_{jm}^{(1)} = \mathbf{v}_{jm} + 0.1\mathbf{H}^{-0.5}\mathbf{r}, \tag{4.8}$$

and

$$\mathbf{v}_{jm}^{(2)} = \mathbf{v}_{jm} - 0.1\mathbf{H}^{-0.5}\mathbf{r}. \tag{4.9}$$

The vector **r** is a random vector where each element is uniformly distributed between -1 and 1. $\mathbf{H}^{-0.5}$ represents the inverse Cholesky decomposition of the matrix **H**, which is given by

$$\mathbf{H} = \frac{1}{\sum_i \gamma_i} \sum_i \gamma_i \mathbf{M}_i^T \Sigma_i^{-1} \mathbf{M}_i. \tag{4.10}$$

The mixture-dependent zero-order statistics γ_i is obtained by accumulating $\gamma_{ji}(t)$ over all the time indices and state indices.

$$\begin{aligned} \gamma_{ji}(t) &= P(s_t = j, m_t = i|X, \Phi) \\ &= P(s_t = j|X, \Phi) P(m_t = i|s_t = j, X, \Phi) \\ &= \left\{ \frac{\alpha_j(t)\beta_j(t)}{\sum_j \alpha_j(T)} \right\} \left\{ \frac{w_{ji} N(x_t; \mathbf{m}_{ji}, \mathbf{S}_i)}{\sum_i w_{ji} N(x_t; \mathbf{m}_{ji}, \mathbf{S}_i)} \right\}. \end{aligned} \tag{4.11}$$

The state posterior probability $P(s_t = j|X, \Phi)$ in eq. (4.11) is obtained using the conventional forward-backward algorithm. The mixture-specific probability for a given state, $P(m_t = i|s_t = j, X, \Phi)$, can be derived from eq. (4.1), with a fixed mixture index *i*.

Initialization

In the early training iterations, the SGMM parameters initialized from a flat start are not well estimated. As a result, the posterior $\gamma_{ji}(t)$ estimated based on these SGMM parameters can not be robustly computed. In order to deal with this issue, an alternative method which is so called the joint state/mixture posteriors-based SGMM parameters initialization is investigated and presented here. The goal of the joint state/mixture posteriors method is to exploit the alignment between HMM states and Gaussian mixture indices as follows.

$$\gamma_{ji}(t) \approx P(s_t = j|X, \Phi_1) P(m_t = i|X, \Phi_2), \tag{4.12}$$

where the state posterior probability $P(s_t = j|X, \Phi_1)$ is obtained from a well-trained CDHMM, Φ_1. The mixture posterior probability $P(m_t = i|X, \Phi_2)$ is obtained from a well-trained UBM, Φ_2. In this case, $P(s_t = j|X, \Phi_1)$ and $P(m_t = i|X, \Phi_2)$ are computed independently from two well-trained models using the forward-backward algorithm. In all the experimental studies presented in this article, when using the joint state/mixture posteriors-based initialization scenario, the

posteriors, $\gamma_{ji}(t)$, are obtained from the joint posteriors in the first seven SGMM training iterations. After the seventh iteration, $\gamma_{ji}(t)$ is computed using the conventional forward-backward algorithm shown in eq. (4.11). The impact of using this joint state/mixture posteriors initialization strategy will be considered in Section 4.3.2.

4.3.2 Word recognition task for the subspace-based models

This section describes the experimental study performed to evaluate the word recognition performance of the SGMM system on the resource management (RM) task domain [30]. The original objective of implementing the SGMM is to apply this new acoustic formalization in the PV task shown in Section 4.2. However, it is also important to see how ASR performance on standard ASR tasks is impacted by the SGMM.

The configuration of this RM task is given as follows. Acoustic speaker-independent (SI) CDHMMs and SGMMs are trained using 3,990 utterances from 109 speakers taken from the standard RM SI-109 training set. Mel frequency cepstrum coefficient (MFCC) feature analysis described in [36] is used. Feature vectors include 12 MFCC coefficients, normalized energy, and their first and second difference coefficients for a 39-dimensional feature vector. Baseline speaker-independent (SI) CDHMM's contains 1,700 left-to-right 3-state state clustered triphones with 6 diagonal Gaussian mixtures per state for a total of 10,224 Gaussians. The SGMM structure is inherited from the baseline CDHMM and the shared Gaussians are initialized using $I = 256$ component UBM.

The CDHMM acoustic model used for obtaining the experimental results relied on the HTK Toolkit for model training and recognition [37]. On the other hand, SGMM training and recognition was implemented by updating HTK training and recognition tools.

Table 4.7 displays the word error rates obtained from CDHMM and SGMM systems configured with a range of parameter allocations for model states and substates. The first four rows of Table 4.7 show the WERs obtained using a baseline CDHMM with 1,700 states and the SGMM models configured using the same number of states. The third row of Table 4.7, labeled "SGMM-FSinit," displays the WER for the flat start initialization, without using the joint state/mixture posteriors. All other SGMM results shown in this work are initialized from the joint state/mixture posteriors described in the section "SGMM initialization and training" and [30].

Table 4.7: WERs for multiple parameter allocations of SGMM and CDHMM.

Acoustic model	States	Subst.	WER
CDHMM	1,700	–	4.91%
SGMM-FSinit	1,700	1,700	4.48%
SGMM	1,700	1,700	4.26%
SGMM	1,700	5,000	3.99%
CDHMM	5,005	–	6.24%
SGMM	5,005	5,005	4.24%

There are several observations that can be made from the top portion of Table 4.7. The first observation is that SGMM configurations with $J = 1700$ states obtain a WER reduction ranging from 8% to 18%. Second, comparing rows two and three, there is a small 5% WER reduction obtained by initializing SGMM training from joint posteriors relative to flat start initialization. Finally, by comparing the third and fourth rows of Table 4.7, it clearly shows that increasing from a single substate per state to approximately three substates per state, the SGMM WER is reduced by 6% relatively.

With only a small number of projection vectors representing state-level information, one might expect that it would be efficient to train these state-level parameters with a relatively small number of effective observations per state. To investigate this conjecture, CDHMM and SGMM models with a much larger number of context clustered states were trained on the same data set used to train the original models. This was thought to be a better means for evaluating training efficiency than simply reducing the overall number of training utterances. There is less of a chance in this case of introducing artifacts that can arise from the highly skewed distribution of phonetic contexts that can occur with a very small corpus size. The efficiency of the SGMM models is demonstrated by comparing the WER for the 5,005 state systems in rows four and five of Table 4.7. The WER obtained for the SGMM system represents a 32% reduction compared to the 5,005 state baseline CDHMM model. This is a much greater reduction than was obtained for the 1,700 state case and illustrates the robustness of the SGMM model with respect to sparseness in training data.

4.3.3 Evaluation on a Spanish language phone recognition task

Given the SGMM formalization introduced in this chapter, this section investigates the SGMM to the PV task by replacing the CDHMM decoder shown in Figure 4.1 with the SGMM-based decoder. This section describes the experimental study performed to evaluate the phone recognition performance of an SGMM-based system on the Spanish unimpaired children speech corpus. In addition to providing another recognition performance comparison between CDHMM and SGMM, the study is extended in Section 4.4.1 to measure how the phone-level PV performance is affected by replacing the CDHMM acoustic model with the SGMM in Figure 4.1.

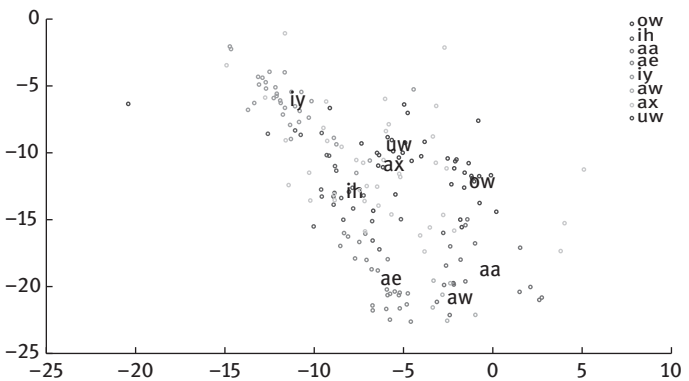

Figure 4.4: State projection plot for vowels in the RM corpus, without mean vector $\mathbf{m}_j$.

The speech corpus involved in this phone recognition task is described in the section "Speaker population," which contains Spanish language utterances from 168 unimpaired children. Each unimpaired speaker provides a set of isolated word recordings for 57 RFI words, shown in Table 4.1. A subset of this corpus containing 120 unimpaired children serves as the acoustic model training corpus. Both CDHMM and SGMM acoustic models are trained from this corpus. Another subset of the speech corpus containing speech from the remaining 48 unimpaired children serves as the phone recognition evaluation data, which gives a total of 14,016 phoneme instances. The impact of acoustic model training from this corpus is compared with the impact of training from the adult speaker Albayzin corpus described in Section 4.2.

The CDHMM consists of 25 phonemes with three state, left-to-right models for each phoneme resulting in a total of 75 nonsilence states. Each state in the

CDHMM consists of 128 diagonal Gaussian mixtures. The SGMM has 25 phonemes, $J = 75$ nonsilence states and $I = 256$ shared Gaussians. The same bigram phonotactic model used in Section 4.2.2 is also used in this phone recognition task. The 39-dimensional MFCC feature analysis is used here.

The phone recognition performance comparison between CDHMM and SGMM is presented in Table 4.8 in terms of the phone error rate (PER) on the above corpus. The second column of the table indicates what combination of "Adult" and "Children" speech data is used for acoustic model training. There are several observations that can be made from the results shown in Table 4.8. First, by comparing rows one and three, it is clear that incorporating children's speech in acoustic model training results in nearly a factor of two reduction in PER. In this case, CDHMM + MAP/MLLR indicates that a combination of MAP adaptation and MLLR adaptation is performed using children speech data on the CDHMM acoustic model trained from adult speech data, as described in the section "Reducing variability through model adaptation." Second, by comparing the best performing CDHMM and SGMM systems in rows four and six respectively, a nearly 25% reduction in PER is achieved. This trend is similar to that observed for WER on the RM task as shown in Table 4.7.

Table 4.8: Phone Recognition performance measured as the phone error rate (%) for CDHMM and SGMM acoustic models, on a 48 speakers subset of the unimpaired children speech corpus.

Phone recognition performance (phone error rate)		
Acoustic model	**Training corpus**	**PER**
CDHMM (16 Gaussians per state)	Adult	31.11%
CDHMM (16 Gaussians per state)	Children	16.8%
CDHMM (16 Gaussians per state) + MAP/MLLR	Adult + Children	16.2%
CDHMM (128 Gaussians per state)	Children	11.7%
SGMM (256 shared Gaussians, without $\mathbf{m}_j$)	Children	9.6%
SGMM (256 shared Gaussians, with $\mathbf{m}_j$)	Children	8.8%

Phone recognition performance as measured on the impaired children corpus for the CDHMM and SGMM acoustic models is presented in Table 4.9. This impaired children corpus is described in section "Speaker population," which involves 14 impaired children. Each impaired speaker provides four set of isolated word

Table 4.9: Phone Recognition performance measured as the phone error rate (%) for CDHMM and SGMM acoustic models, impaired children speech corpus.

Phone recognition performance (phone error rate)		
Acoustic model	**Training corpus**	**PER**
CDHMM (16 Gaussians per state)	Adult	51.9%
CDHMM (16 Gaussians per state)	Children	43.1%
CDHMM (16 Gaussians per state) + MAP/MLLR	Adult + Children	42.4%
CDHMM (128 Gaussians per state)	Children	41.6%
SGMM (256 shared Gaussians, without $\mathbf{m}_i$)	Children	38.4%
SGMM (256 shared Gaussians, with $\mathbf{m}_i$)	Children	41.0%

recordings for 57 RFI words. The phone error rates evaluated on this corpus may be misleading because, as described in the section "Pronunciation labeling by nonexpert human labelers," many of the phone occurrences in the utterances from disabled speakers are known to be deleted or badly mispronounced.

There are several observations that can be made from Table 4.9 about ASR performance on impaired children speech utterances. First, by comparing all PERs in Tables 4.8 and 4.9, it is obvious that the PERs for impaired speaker utterances are dramatically higher than they are for unimpaired speakers, Second, by comparing rows one and three in Table 4.9, incorporating training data from children speakers has shown to significantly reduce PER. However, the impact is far less than that obtained from evaluating on unimpaired speaker utterances. Third, by comparing rows four and five, it is clear that the best SGMM PER obtained for the impaired speaker corpus is significantly lower than that obtained for unimpaired speakers.

In comparing the PERs displayed for the unimpaired and impaired utterances in Tables 4.8 and 4.9 respectively, it is clear that the general trends in performance are similar. However, the absolute difference in PERs observed for the acoustic modeling approaches are not the same. This is due to two reasons. First, the difference between unimpaired children speech used in training and the impaired speech evaluation utterances is significant. Second, as stated above, there is a fundamental issue with reporting PER on impaired speakers' utterances due to the large number of phone deletions and mispronunciations.

4.3.4 Summary

The SGMM acoustic model formalism has been introduced in this chapter. Based on the ASR experimental study in Sections 4.3.2 and 4.3.3, SGMM provides a lower word error rate on the RM task and a lower phone error rate on the unimpaired children corpus. It is believed that the recognition performance improvement given by the SGMM structure comes from the efficient decomposition of acoustic space to modeling the pronunciation variation, and also given by efficiently reducing the number of model parameters required to be estimated. The impact on the phone-level PV performance when replacing the CDHMM by the SGMM in the CN-based PV scenario shown in Section 4.2 will be addressed in Section 4.4.1. Besides, in Section 4.4, the SGMM state projections in subspaces will be applied in a new pronunciation verification scenario.

4.4 Applying subspace-based pronunciation modeling in verifying pronunciation accuracy

This chapter introduces a new PV approach [38, 39], which is based on SGMM described in Section 4.3. It extends the CN-based phoneme-level pronunciation verification (PV) scenario which has been presented in Section 4.2. First, the performance of two CN-based PV systems are compared in Section 4.4.1. The first is based on the HMM-based phonetic decoder, and the second is based on the SGMM-based phonetic decoder. Second, a discussion of an articulatory interpretation of SGMM subspace projection vectors is presented in Section 4.4.2. This interpretation motivates a new approach for detecting phoneme-level mispronunciations from utterances obtained from impaired children with neuromuscular disorders. The new phoneme-level PV scenario will be described as follows. Section 4.4.3 describes the SGMM parameters used for this new PV scenario. Then, a distance measure between two state projection vectors within the same subspaces is investigated in Section 4.4.4 as a decision criterion for detecting phoneme-level mispronunciations. Finally, Section 4.4.5 presents the experimental study for this new PV approach.

4.4.1 Applying SGMM into the CN-based PV scenario

In Section 4.2, a CN-based approach has been described for verifying the phoneme-level pronunciation accuracy. Given a well-trained monophone-based

continuous density hidden Markov model (CDHMM), the approach begins by performing a phonemic decoding on the given testing isolated word utterance. A phone lattice containing phone labels and their associated acoustic probabilities is then generated through the phonetic decoder. Finally, a CN is created from the phone lattice using a lattice compression algorithm. The CN is a linear network where all arcs that emanate from the same start node terminate in the same end node. The ordering properties of the original lattice are maintained in the confusion network. The posterior phone probability corresponding to the given target phoneme from the baseform expansion of the given testing word will appear on the arcs of the confusion network. This posterior phone probability is used as the utterance-specific phoneme-based confidence score.

An alternative acoustic modeling technique, the SGMM, is introduced in Section 4.3. As discussed in Sections 4.3.2 and 4.3.3, both word recognition and phone recognition performance given by the SGMM are significantly higher than the recognition performance given by the conventional CDHMM. Therefore, it is reasonable to consider the use of the SGMM acoustic model in place of the CDHMM model in the phonetic decoder for the CN-based PV approach presented in Section 4.2. The corresponding phone-level PV performance comparison between the CDHMM phonetic decoder and the SGMM phonetic decoder will be presented in this section. The main difficulty of applying the SGMM in the CN-based PV framework arises from the fact that no implementation of SGMM-based speaker adaptation existed at the time this work was being performed. As a result, none of the adaptation scenarios used for the CDHMM acoustic model is investigated here. Thus, the CDHMM or SGMM training corpus used for the experimental study shown in this section includes only the unimpaired children speakers described in the section "Speaker population."

The same training recipe presented in Section 4.3.3 is used in this CN-based PV experimental study. The speech obtained from the 120 unimpaired children speakers serve as the CDHMM and SGMM acoustic model training corpus. The CDHMM consists of 25 monophones modeled by three state left-to-right HMMs, resulting in a total of 75 nonsilence states. The SGMM has $J = 75$ nonsilence states, and $I = 256$ shared Gaussians. The same bigram phonotactic model used in Section 4.2.2 is used in this PV task. All the utterances from the 14 impaired children speakers are used as the PV evaluation set. This involves a total of 16,352 phoneme trials, which are given by four recording sessions of 57 RFI words per impaired children speakers. Each session of 57 RFI words consists of 292 phoneme trials.

The PV performance given by various acoustic models are summarized in Table 4.10. The same acoustic modeling approaches shown in Table 4.9 are involved in this Table 4.10. The EERs given by rows one and three are similar to

Table 4.10: PV performance comparison between CDHMM and SGMM acoustic models, impaired children speech corpus.

Pronunciation verification performance (equal error rate)		
Acoustic model	**Training corpus**	**EER**
CDHMM (16 Gaussians per state)	Adult	22.4%
CDHMM (16 Gaussians per state)	Children	20.5%
CDHMM (16 Gaussians per state) + MAP/MLLR	Adult + Children	18.3%
CDHMM (128 Gaussians per state)	Children	20.1%
SGMM (256 shared Gaussians, without $\mathbf{m}_i$)	Children	19.3%
SGMM (256 shared Gaussians, with $\mathbf{m}_i$)	Children	19.2%

the EERs presented in the first two rows in Table 4.4, The slight EER difference is due to the fact that this PV experimental study involves four instead of three recordings for each of 57 RFI words for each of the impaired children.

There are several observations that can be made from the EERs shown in Table 4.10. First, when comparing the PERs shown in Table 4.9 and the EERs shown in Table 4.10, it indicates that the acoustic model training scenario will have the same impact on the phone recognition performance and on the CN-based phoneme-level pronunciation verification performance. A more robust acoustic model which gives a lower recognition error rate, a better CN-based PV performance can be expected when the CN is obtained from that acoustic model. Second, by comparing rows one and three in Table 4.10, incorporating training data from children speakers has shown to significantly reduce EER. This observation has also been made in the experimental study shown in Section 4.2.2. Third, by comparing rows four and six, it is clear that the best SGMM outperforms the best CDHMM in terms of EER, when the same unimpaired speaker corpus is involved in the acoustic model training. A 4.48% EER relative reduction can be achieved. In order to further improve the PV performance using the SGMM acoustic model formalism, a new phone-level pronunciation verification scenario will be investigated in the following sections.

4.4.2 Subspace interpretation

Multiple subspace matrices provided by SGMM, which is introduced in Section 4.3, describe the allowable variation associated with individual ASR acoustic model

distributions. Each state associated with a phonetic context event in the CDHMM is represented as one or more low dimensional projection vectors within these subspaces in the SGMM. One obvious advantage of using SGMM is to obtain a more robust acoustic model from a limit amount of training data. Since each state in the SGMM is only parameterized by one or few low dimensional projection vectors, it provides the potential for modeling phones associated with individual states using only a relatively small amount of occurrences of that phone.

Another advantage of using SGMM is SGMM can be loosely interpreted as a subspace representation of phonetic-level variation in speech recognition [40]. It has been found by Lukas Burget et al. that the state projection vectors in the SGMM can loosely represent vowel sounds by two-dimensional plots. Simply stated, if the first two elements of each state projection vectors associated with the center state of each context-dependent vowels in the SGMM are displayed in the two dimensional plot, there is a well-behaved clustering property associated for each vowels.

An example of such state projection vectors plot is given by Figure 4.4. The SGMM involved in this figure is trained using the resource management (RM) corpus, as described in Section 4.3.2. Each dot in Figure 4.4 is associated with the center state of one context-dependent vowel shown in the RM training corpus. The location of each vowel label is given by the center of each vowel cluster. The distribution of vowel symbols is close to the English-based vowel triangle plot, as shown in Figure 4.5. Vowels are distinct from each other based on their acoustic form, or spectral properties. Spectral properties consist of the speech sound's fundamental frequency and its formants. Each vowel in the vowel triangle diagram has a unique first and second formant, denoted as F1 and F2 respectively. In Figure 4.5, F1 values are shown in the y-axis, and the F2 values are shown in the x-axis.

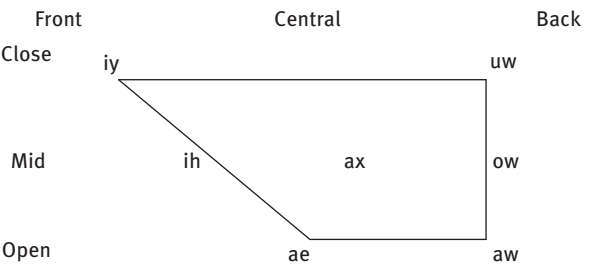

Figure 4.5: Vowel Triangle for English, y-axis means the lip status, x-axis means the tongue position.

The state projection vectors plot shown in Figure 4.4 is obtained from the SGMM structure without using the mean vector $\mathbf{m}_i$. In other words, $\mathbf{m}_i$ in eq. (4.2) is a zero vector. It has been found that the state projection vectors clustering is not so significant when the mean vector $\mathbf{m}_i$ is involved in eq. (4.2). In this case, the state-dependent means $\boldsymbol{\mu}_{ji}$ is dominated by $\mathbf{m}_i$, and only the residual can be modeled by the product of $\mathbf{M}_i$ and $\mathbf{v}_j$.

The state projection vectors have also been found to form well-behaved clusters for the SGMM trained using multiple unimpaired children speaker described in the section "Speaker population." A similar state projection vectors plot for Spanish vowels, $\{a, e, i, o, u\}$, is shown in Figure 4.6. The SGMM involved in Figure 4.6 is trained using both unimpaired and impaired children speech corpus. There is one set of speaker-independent state projection vectors obtained from the speech provided by 60 unimpaired children, which are labeled as $\{a_{U1}, e_{U1}, i_{U1}, o_{U1}, u_{U1}\}$ in the plot. There is another set of speaker-independent state projection vectors obtained from the speech provided by other 60 disjointed unimpaired children, which are labeled as $\{a_{U2}, e_{U2}, i_{U2}, o_{U2}, u_{U2}\}$ in the plot. On the other hand, the SGMM also involves 14 sets of speaker-dependent state projection vectors, and each set is obtained from the speech provided by one impaired children. The labels $\{a_{Imp}, e_{Imp}, i_{Imp}, o_{Imp}, u_{Imp}\}$ shown in the plot are associated with one of the 14 impaired children speaker.

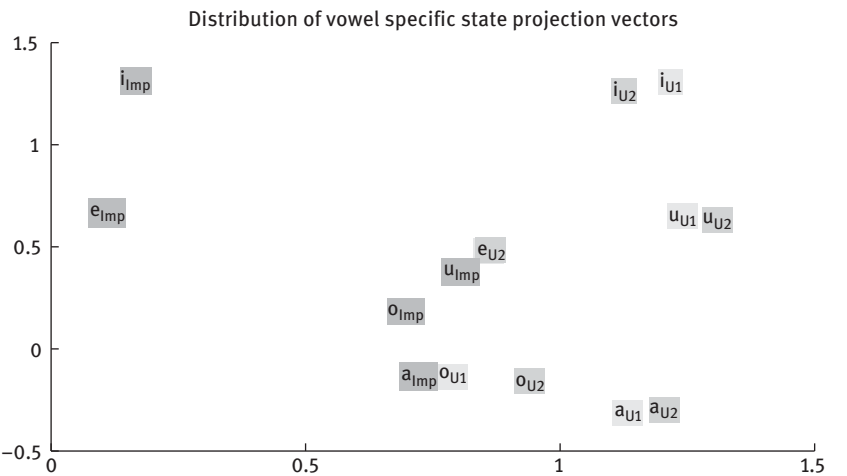

Figure 4.6: Distribution of vowel-specific state projection vectors. U_1 represents the first set of unimpaired speaker population. U_2 represents the second set of unimpaired speaker population. *Imp* represents one of the 14 impaired speaker. e_{U_1} is overlapped with e_{U_2}.

In Figure 4.6, the location of each label is given by the center of each vowel clusters. The distribution of five centers of $\{a, e, i, o, u\}$ obtained from one set of 60 unimpaired children is similar to the distribution of five centers obtained from another set of 60 unimpaired children. On the other hand, the distribution of $\{a_{Imp}, e_{Imp}, i_{Imp}, o_{Imp}, u_{Imp}\}$ obtained from an impaired children speaker is much different from the distributions obtained from unimpaired speaker population. This clustering behavior provides a motivation to compare a phoneme pronounced by an impaired speaker and by a set of unimpaired speaker population, which will be described in the following sections.

4.4.3 SGMM parameterization

The SGMM configuration involved in the new PV scenario, which will be described in Section 4.4.4, is shown in Figure 4.7.

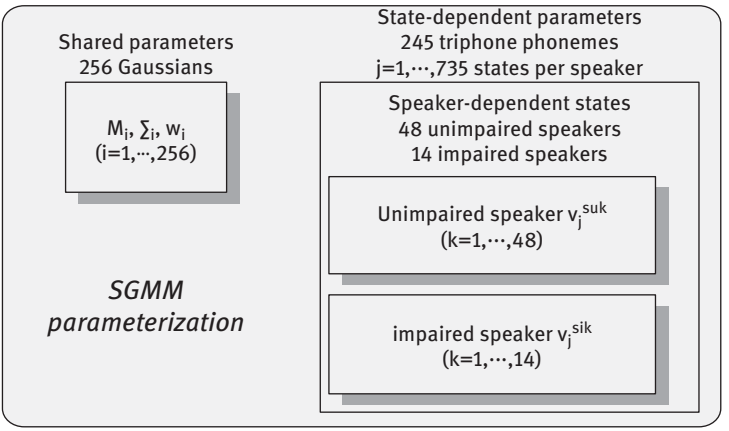

Figure 4.7: SGMM: speaker-dependent states.

The shared parameters in the model are trained from the entire population of training speakers. However, a separate set of states is allocated for each speaker. Note in Figure 4.7, speech from both unimpaired and impaired children speakers is used for training this SGMM model.

Specifically, the SGMM parameterization is described as follows. The dimension, F, of feature vectors is 39, which includes 12 MFCC coefficients, normalized energy, and their first and second difference coefficients. The

subspace dimension, S, is also set to 39. The SGMM structure is inherited from an initial CDHMM and speech UBM. Both CDHMM and speech UBM are trained using the unimpaired speech corpus described in the section "Speaker population." The CDHMM states are part of three state unclustered context-dependent triphones obtained from the 57 words from the RFI. The total number of context-dependent triphone units is 245, which corresponds to 108 context-dependent vowels and 137 context-dependent consonants. The speech UBM consists of 256 full covariance Gaussians.

There are several speaker-dependent sets of SGMM state projection vectors. Each set contains 735 state projection vectors, and is trained using the speech corpus for one unimpaired or impaired individual speaker. The 735 state projection vectors in each set are associated with the 245 three state context-dependent triphones. A subset of the unimpaired children speech corpus, which consists of 48 unimpaired children, is used for training 48 sets of unimpaired speaker-dependent state projection vectors, $\left\{\mathbf{v}_j^{su_k}\right\}$. Each of the 48 sets is represented by one unimpaired child, su_k, $k=1,\ldots,48$. Each of the impaired children, si_k, $k=1,\ldots,14$, from the impaired children speech corpus is used for training one set of impaired speaker-dependent state projection vectors, $\left\{\mathbf{v}_j^{si_k}\right\}$. This gives 14 sets of impaired speaker-dependent state projection vectors. Note that each of the impaired speaker-specific state projection vector are trained using multiple instances in the SGMM training corpus. Some of them are correctly pronounced, and some of them are mispronounced.

The motivation for configuring the model in Figure 7 is to define a measure of phonetic variation directly in the state projection vector space. State projection vectors associated with the states of individual phones from multiple unimpaired speakers have been found to form well-behaved clusters, as described in Sections 4.4.2. 4.4.4 defines a PV distance measurement that exploits this behavior by measuring the deviation of state projection vectors obtained from an impaired speaker from the state projection vectors obtained from an unimpaired speaker population.

4.4.4 State projection-based PV scenario

Pronunciation verification (PV) refers to obtaining confidence measures for each phoneme in the baseform expansion and applying a decision rule for accepting or rejecting the hypothesis that a given phone was correctly pronounced. First, the section "Distance between two state projection vectors" presents a new SGMM-based approach of obtaining the utterance-independent confidence score for each context-dependent triphone unit from each of the impaired speakers. Second, the

section "Distance between two state projection supervectors" proposes a supervector approach to construct the confidence scores. Third, the section "Linear discriminant analysis" discusses the use of linear discriminant analysis (LDA) for suppressing inter-speaker variation in the phoneme-level PV scenario.

Distance between two state projection vectors

It is assumed that the pronunciation of a context-dependent phoneme, q, is characterized by the state projection vector associated with its center state, j. Given the state index, j, as the center state of the phoneme q, the PV decision of rejecting or accepting that phoneme q has been correctly pronounced by an impaired speaker si_k can be obtained from the distance between two state projection vectors within the same SGMM subspace. One state projection vector $\mathbf{v}_j^{si_k}$ is trained from an impaired speaker si_k. Another reference state projection $\mathbf{v}_j^{ref}$ is obtained using utterances from the unimpaired speaker population. Specifically, $\mathbf{v}_j^{ref}$ is obtained by clustering state projection vectors, $\mathbf{v}_j^{su_k}$, $k = 1, \ldots, 48$, where $\mathbf{v}_j^{su_k}$ is the state projection vector for the kth unimpaired speaker.

Euclidean distance and the cosine distance were investigated for evaluating the distance between two state projection vectors. It was found that the cosine distance measurement provided more robust PV performance than the Euclidean distance. The cosine distance is given by

$$D\left(\mathbf{v}_j^{ref}, \mathbf{v}_j^{si_k}\right) = \frac{\mathbf{v}_j^{ref} \cdot \mathbf{v}_j^{si_k}}{||\mathbf{v}_j^{ref}||||\mathbf{v}_j^{si_k}||}, \tag{4.13}$$

where the notation $||\mathbf{v}_j||$ indicates the magnitude of the vector $\mathbf{v}_j$. A phonetic subspace normalization is also implemented, in order to avoid the numerical issue where some components of a state projection vector may have a very high dynamic range [27]. This normalization over phonetic subspaces will improve the robustness of constructing the distance measurement between $\mathbf{v}_j^{ref}$ and $\mathbf{v}_j^{si_k}$, when a unique PV decision threshold is applied across all states.

Distance between two state projection supervectors

Equation (4.13) proposes a mechanism to run the phoneme-level PV task by comparing the pronunciation of the phoneme q between an impaired speaker and a cluster of unimpaired speakers by measuring the cosine distance between two state projection vectors within the same subspace of the SGMM. One can also

assume that the pronunciation of a given phoneme q is not just characterized by its center state projection vector, but is instead characterized by the three state projection vectors associated with states $\{j-1, j, j+1\}$. A state projection supervector associated with the phoneme q with dimension of $3S$, where $S = 39$, can be constructed by concatenating the three state projection vectors as follows,

$$\mathbf{V}_q = (\mathbf{v}_{j-1}, \mathbf{v}_j, \mathbf{v}_{j+1}). \tag{4.14}$$

The cosine distance can then be computed in the supervector domain,

$$D(\mathbf{V}_q^{ref}, \mathbf{V}_q^{si_k}) = \frac{\mathbf{V}_q^{ref} \cdot \mathbf{V}_q^{si_k}}{||\mathbf{V}_q^{ref}||||\mathbf{V}_q^{si_k}||}. \tag{4.15}$$

The reference $\mathbf{V}_q^{ref}$ in the supervector domain is given by clustering the 48 unimpaired speaker-specific state projection supervectors, $\mathbf{V}_q^{su_k}$. The idea of obtaining $\mathbf{V}_q^{ref}$ is similar to obtaining the $\mathbf{v}_j^{ref}$, as described in the section "Distance between two state projection vectors."

Linear discriminant analysis

Linear discriminant analysis (LDA) is a well-known technique in pattern recognition and machine learning [41]. The goal in this work is to apply LDA to reduce the impact of speaker variability in the PV task by performing a dimensionality reducing linear transformation on the super vector $\mathbf{V}_q$ in eqs. (4.14). The new state projection supervector $\hat{\mathbf{V}}_q$ for a given phoneme, q, can be obtained by applying the LDA transform as follows,

$$\hat{\mathbf{V}}_q = \mathbf{L}^T \mathbf{V}_q. \tag{4.16}$$

The LDA transform $\mathbf{L}$ is of dimension $3S \times S'$, where $S' \leq 3S$. Intuitively, LDA rotates the state projection supervectors to a new direction that better discriminates between state projection supervectors belonging to different phoneme classes. Thus, a better phoneme-level PV performance based on eqs. (4.15) could be achieved after applying the LDA to state projection supervectors.

The column vectors of the LDA transform $\mathbf{L}$ are given by the eigenvectors with the S' highest eigenvalues of the generalized eigenvalue problem,

$$\sum\nolimits_B \mathbf{E} = \sum\nolimits_W \mathbf{E}\mathbf{D}. \tag{4.17}$$

The eigenvectors and the corresponding eigenvalues are obtained from the columns of matrix **E** and the diagonal elements of diagonal matrix **D**.

It is assumed that each context-dependent phoneme q, $q = 1, \ldots, Q$, where $Q = 245$, represents one class in the LDA, and each class contains 48 unimpaired speaker-specific supervectors, $\mathbf{V}_q^{su_k}$, provided by a subset of the unimpaired speech corpus as described in Section 4.4.3. The between classes covariance Σ_B and the within classes covariance Σ_W can be computed as follows:

$$\Sigma_{\mathbf{W}} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{48}\sum_{k=1}^{48}(\mathbf{V}_q^{su_k} - \mu_q)(\mathbf{V}_q^{su_k} - \mu_q)^T, \tag{4.18}$$

and

$$\Sigma_{\mathbf{B}} = \frac{1}{Q}\sum_{q=1}^{Q}(\mu_q - \mu)(\mu_q - \mu)^T. \tag{4.19}$$

The phoneme-dependent mean μ_q is given by the average of all the 48 supervectors $\mathbf{V}_q^{su_k}$ for each phoneme q. The mean μ is given by the average of all the phoneme-dependent mean vectors, μ_q, over all the phoneme classes. In this way, LDA attempts to increase the class separability by maximizing the between classes covariance capturing the intra-speaker variability, and minimizing the within classes covariance capturing the inter-speaker variability. The effect of applying the LDA on state projection supervectors will be discussed in Section 4.4.5.

4.4.5 Experimental study

The PV evaluation shown in this experimental study is based on the SGMM cosine distance measurement described in Section 4.4.4. First, the baseline system is described in the section "Baseline system." Second, the session-level PV evaluation is presented in the section "Session-level PV results." This is presented as an average equal error rate (EER) which describes detection performance across an entire session. Third, the utterance-level PV evaluation is presented in section "Utterance-level PV results." This provides a measure of performance for verifying the occurrence of individual instances of phoneme-level mispronunciation.

Baseline system

The baseline PV scenario is based on the CN-based confidence scores, which is presented in Section 4.2. For the baseline system presented in this work, the acoustic CDHMM training is initialized using Albayzín corpus [21]. Then it is adapted to the unimpaired children speech corpus using both maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR) adaptation, as described in the section "Reducing variability through model adaptation." A zerogram-based unconstrained phonotactic network is used for decoding.

Session-level PV results

The session-level PV performance is reported as the average mispronunciation detection performance across all context-dependent phonemes in the test corpus. There are four utterances of each phoneme in an impaired speaker's session. If a majority of the individual instances of a phoneme in a speaker's session are labeled as being correctly pronounced, then that phoneme is labeled as correctly pronounced for the session. There are 245 context-dependent phonemes from the 57 RFI words and provided by each of 14 impaired speakers. This gives total of 3,430 session-level test trials. This involves 2,756 test trials labeled as "correctly pronounced," 543 test trials labeled as "incorrectly pronounced," and 131 test trials which are excluded from the evaluation.

For the baseline PV scenario, the overall session-level CN confidence score for a phoneme q is obtained by averaging all the CN-based confidence scores corresponding to q, among all the four instances of that phoneme. This gives an equal error rate (EER) of 15.83%.

In the section "Distance between two state projection vectors," a cosine distance measure between two state projection vectors sharing the same linear subspaces in a SGMM is defined. This provides a model-based approach for measuring the similarity between phoneme models trained from an impaired speaker and a cluster of unimpaired speakers. It gives a session-level EER of 21.73%. When the cosine distance is measured between two state projection supervectors, as described in eq. (4.14), the EER reduces to 19.85%. If LDA transformed supervectors described in eq. (4.16) are used for computing the cosine distances, the EER can be further reduced to 18.44%. These results can be shown in Figure 8 with a fusion weight equal to zero.

It is reasonable to assume that there would be some advantage to combining the scores from the two systems. The SGMM cosine distances can provide context information which is potentially complementary to the baseline CN

scores, which are obtained from a context-independent phonetic decoder. A simple session-level combined score, $S_{ui}(q, si_k)$, for a given impaired speaker si_k and context-dependent phoneme q can be expressed as follows,

$$S_{ui}(q, si_k) = \alpha S^{CN}(q, si_k) + (1 - \alpha) D(\mathbf{v}_j^{ref}, \mathbf{v}_j^{si_k}). \tag{4.20}$$

The first additive term in eq. (4.20), $S^{CN}(q, si_k)$, represents the overall session-level CN posterior score obtained from the baseline system. The second additive term is the SGMM-based cosine distance measurement between two state projection vectors, as defined in eq. (4.13). The state index j represents the center state of the phoneme q. This cosine distance can also be computed between two state projection supervectors, as defined in eq. (4.14), or between two LDA transformed supervectors, as defined in eq. (4.16). The fusion weight is controlled by the factor α, varying from zero to one. The session-level PV performance from the combined scores computed from eq. (4.20) is shown in Figure 4.8. These results show that the CN baseline PV performance can be improved by incorporating any of three kinds of SGMM cosine distance scores, if a proper fusion weight is selected. Comparing the best results from the combined scores with the baseline results, the EER drops from 15.83% to 13.40%.

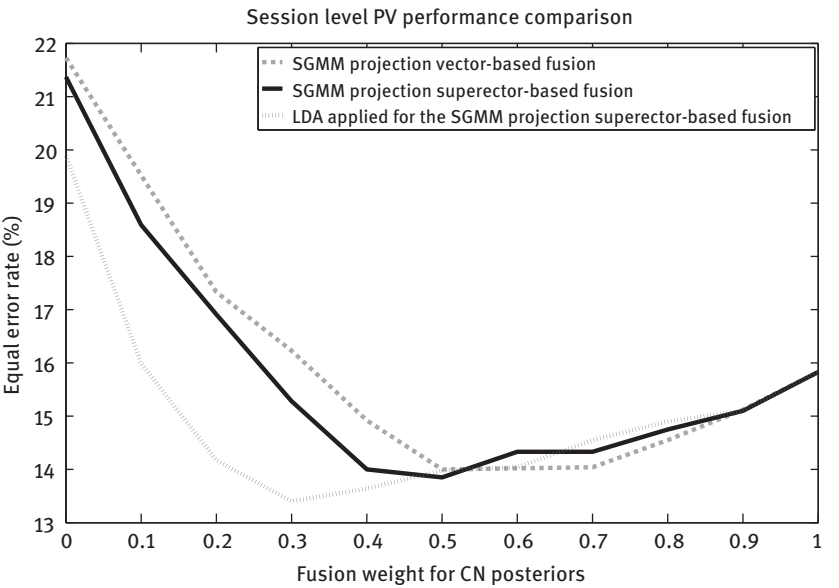

Figure 4.8: EER comparison, session-level PV evaluation.

Utterance-level PV results

The utterance-level PV is evaluated directly based on the transcription provided by human labelers, as described in the section "Pronunciation labeling by non-expert human labelers," There are 16,352 phoneme-level test trials, which involve 13,472 "correctly pronounced" trials and 2,880 "incorrectly pronounced" trials. The "incorrectly pronounced" trials include all the "mispronounced" and "deleted" trials.

The baseline CN confidence scores provide an EER of 19.86%, which is shown in Figure 4.9 with a fusion weight equal to one. This result is slightly different from the result reported in the second row of Table 4.4. It is because the evaluation shown in this PV experimental study involves four instead of three recordings for each of 57 RFI words for each of the impaired children.

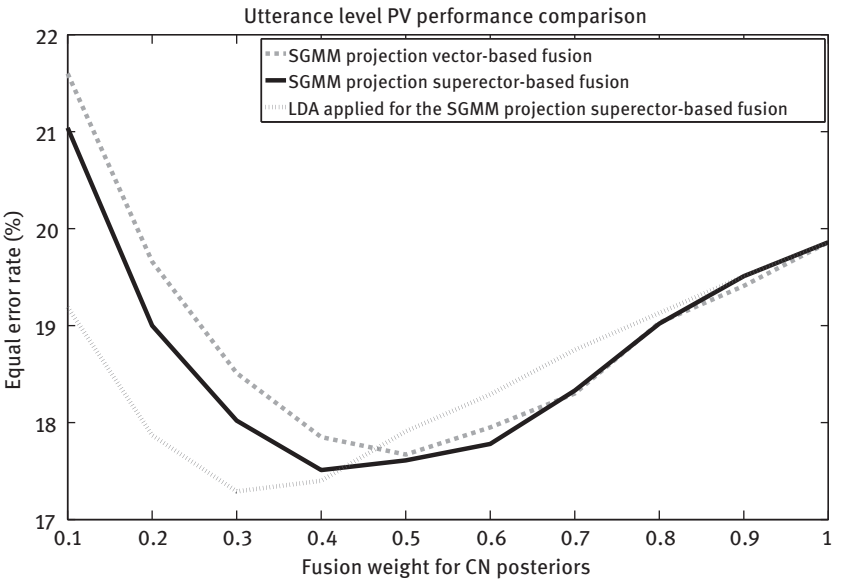

Figure 4.9: EER comparison, utterance-level PV evaluation.

For a given impaired test speaker, SGMM state projection vectors have been trained using all occurrences of each phone from the test speaker utterances. So the model-based distance given in eq. (4.13) is a score that implicitly incorporates all of these occurrences. However, it is still believed that the context information captured by SGMM cosine distances can be complementary to the

utterance-level CN posterior scores. The combined score $S_{ud}(q, si_k, u)$ for a given impaired speaker si_k and phoneme q in the baseform expansion of the word in the utterance u is proposed as follows,

$$S_{ud}(q, si_k, u) = \alpha S^{CN}(q, si_k, u) + (1 - \alpha) D(\mathbf{v}_j^{ref}, \mathbf{v}_j^{si_k}). \tag{4.21}$$

The first additive term in eq. (4.21), $S^{CN}(q, si_k, u)$, represents the utterance-specific CN posterior score obtained from the baseline system. The second additive term is the SGMM-based cosine distance measurement between two utterance-independent state projection vectors, which has exactly the same form as the second additive term in eq. (4.20). Similarly, the cosine distance involved in the second additive term can also be computed between two state projection supervectors, or between two LDA transformed state projection supervectors.

The fusion weight is controlled by the factor α, varying from zero to one. The utterance-level PV performance obtained from the combined scores given by eq. (4.21) is shown in Figure 4.9. These results show that, even in the utterance-level PV task, the context information captured by the SGMM cosine distances can still be useful for improving the CN baseline performance. Comparing the best results from the combined scores with the baseline results, there is a reduction in EER from 19.86% to 17.29%.

4.4.6 Summary

An SGMM-based measure of performance accuracy has been presented and evaluated on a pronunciation verification task for impaired children speakers. It was that, when combined with a lattice-based method for deriving phone-level confidence measures, a PV EER of as low as 13.4% was obtained. The best performance was obtained by forming supervectors by concatenating the SGMM state projection vectors and performing discriminative dimensionality reduction in this space. These performance improvements are believed to result from an efficient characterization of context information for each phoneme by SGMM parameterization.

4.5 Conclusion

This article has addressed the issues of phonetic variability in speech arising from speech impairments. One of the main challenges in addressing this issue has been characterizing the variations in pronunciation associated with disabled

speakers differs from what might be considered normal inter-speaker pronunciation variation within populations of impaired speakers. The main contributions of this work are the statistical modeling techniques for characterizing phonetic variation in speech and multiple pass pronunciation verification (PV) approaches relying on CDHMM and SGMM modeling formalisms. The development and evaluation of these approaches has been enabled by the existence of an annotated corpus of speech utterances from impaired and unimpaired children speakers. This Chapter summarizes the contributions and proposes topics for future work.

4.5.1 Summary of contributions

The speech therapy application reviews are introduced in Section 4.1. The main contributions of this thesis work are presented in Sections 4.2, 4.3 and 4.4. A summary of these contributions will be briefly presented as follows.

Multiple pass ASR approach to pronunciation verification

In Section 4.2, a posterior probability CN-based PV scenario has presented. It was evaluated on a task where instances of mispronounced phonemes occurred in a predefined isolated word-based speech corpus. Given a well-trained CDHMM acoustic model, a decoded phonemic lattice is produced on the given isolated word utterance. The phone lattice structure contains phone labels and their associated acoustic probabilities. Then, a CN is created from the phone lattice. The posterior phone probability corresponding to the given target phoneme from the baseform expansion of the given testing word will appear on the arcs of the confusion network. This posterior phone probability is used as the utterance-specific phoneme-based decision criteria for making the mispronunciation decision. Various different speaker adaptation strategies were investigated for reducing the mismatch between the CDHMM acoustic model and the impaired younger children speaker population. As a result, the phoneme-level PV performance with respect to the impaired children test corpus was improved. The equal error rate (EER) was reduced from 25.3% to 17.2% through the acoustic model adaptation, and an EER of 14.9% was achieved when applying a nonlinear mapping for the CN posterior scores. This CN-based PV approach provides the baseline verification performance which was needed as a point of comparison for the approaches investigated in Section 4.3 and 4.4.

Subspace Gaussian mixture model implementation in ASR

There were two major sets of contributions in the thesis relating to acoustic modeling using the SGMM. The SGMM was originally developed in [26, 27]. The first set of contributions in this thesis addressed a number of practical problems associated with SGMM implementation for ASR tasks. The second set of contributions involved the application of the SGMM acoustic model to the pronunciation verification task. The SGMM infrastructure used in these works was developed at McGill.

SGMM Implementation in ASR: The contributions related to SGMM implementation in ASR are presented in Section 4.3 and briefly described as follows. First, an efficient strategy for parameter initialization in EM-based training was presented. Second, a robust approach for identifying substate projection vectors was implemented. Third, an efficient likelihood computation approach based on Guassian preselection in ASR decoding was empirically established. Fourth, it was demonstrated that a 18.74% relative reduction in word error rate with respect to the well-known CDHMM acoustic model on a medium vocabulary ASR task. Finally, it was demonstrated that a 24.79% relative reduction in phone error rate with respect to the CDHMM for an unimpaired children speech corpus.

Applying SGMM in PV: The SGMM acoustic model formalism introduced in Section 4.3 provides a state-level acoustics representation in subspaces. Section 4.4.2 presents some two-dimensional plots obtained from the vowel-specific SGMM state projection vectors. The state projection vectors associated with the same phone are clustered in the two-dimensional plots. This clustering property illustrates that SGMM can be loosely interpreted as a subspace representation of phonetic-level variation. This behavior provides a motivation to build a phone-level mispronunciation decision criteria based on the distance of two state projection vectors within a SGMM. One state projection vector is obtained from the reference unimpaired speaker population; another state projection vector is obtained from an impaired speaker. In other words, the PV confidence scores are not constructed from the ASR decoder as presented in Section 4.2, but constructed directly from the SGMM formalism. A cosine distance between two state projection vectors serves as the confidence score used for making the phoneme-level PV decision. In order to characterize the phonetic variation, the SGMM states are based on the context-dependent triphone units instead of monophone units. For each of the impaired or unimpaired speaker, a set of state projection vector are obtained from all the training observation vectors provided by that speaker. Therefore, the SGMM cosine distance scores are utterance-independent and context-dependent. Both session-level and utterance-level PV scenarios are proposed and the corresponding

experimental studies are presented in Section 4.4. In both PV scenarios, in order to achieve the best PV performance, the CN-based confidence scores proposed in Section 4.2 and the SGMM-based cosine distance score are combined together through a simple linear equation. Consider the CN-based PV scenario as a baseline system. It is shown that, in the session-level PV task, the equal error rate can be reduced from 15.83% to 13.40% when combining the CN-based confidence scores and the SGMM cosine distance scores. On the other hand, in the utterance-level PV task, the equal error rate can be reduced from 19.86% to 17.29%. These equal error rate performance improvements are believed to result from an efficient characterization of context information for each phoneme by SGMM parameterization. It is reasonable to assume that there would be some advantage to combining the scores from the two systems. The SGMM cosine distances can provide context information which is potentially complementary to the baseline CN scores, which are obtained from a context-independent phonetic decoder.

4.5.2 Investigation in the future

Other ASR techniques

There are two ASR techniques which have potential to improve the PV performance proposed this thesis work.

Adaptation in acoustic feature domain

Replacing the CDHMM by the SGMM as the phonotactic decoder in the CN-based PV scenario proposed in Section 4.2 is trivial. However, the PV performance can not easily be improved due to the lack of efficient SGMM adaptation techniques. One solution is to perform the adaptation in the MFCC acoustic feature domain, such as the vocal track length normalization (VTLN), a well-known speaker adaptation technique used to improve the speech recognition accuracy. The standard filterbank-based mel-frequency cepstrum coefficient (MFCC) is introduced [36]. The MFCC features are constructed based on the following steps: sliding a data window along the speech signal, doing FFT analysis to obtain the discrete spectrum magnitude, passing a set of the mel-warped filterbanks to get the filter bank energies, take a logarithmic compression of the filter bank outputs, finally doing the discrete cosine transform (DCT) to get the MFCC features. The idea of VTLN speaker adaptation is to estimate a set of speaker-specific warping factors, which are used to warp

the center frequencies of the filterbanks. One easy way to estimate the warping factors is based on the maximum likelihood (ML) criterion over the training data [42]. One can do a grid search over a set of warping factors and find the optimized warping factor which maximizes the ASR likelihood score over the speaker-specific data. Once the warping factor is found and applied for warping the filterbanks, the speaker adaptation can be performed in the MFCC domain. It will be interesting to see if the additional speaker adaptation performed in the acoustic feature domain can further improve the PV performance or not.

Applying LDA in acoustic feature domain

In the section "Linear discriminant analysis," the LDA transform is performed on the state projection supervector domain in order to reduce the impact of speaker variability in the PV task. The experimental studies presented in Section 4.4.5 shows that the LDA transformed state projection supervector-based cosine distance scores gives the better PV performance than the raw state projection supervector-based cosine distance scores. The success of applying the LDA in the model parameter domain gives a hint to try the LDA directly in the acoustic feature domain. There are some research papers show how to apply the LDA transform to the acoustic features in order to carry out the discriminant information for the phoneme classification [43, 44]. First, the acoustic feature vectors from the training data are aligned with the states of the baseline acoustic model. Second, for each time frame in a training utterance, a supervector is constructed by concatenating the 2N adjacent feature vectors. Suppose the feature vector is given by the 13-dimensional MFCC vectors, and $N = 5$, then the supervector for each time index will have a dimension of 143. The state indices determine the LDA classes. The mean vector for each LDA classes are computed from all the supervectors aligned to that state. The between classes covariance and the within classes covariance can then be computed, as shown in the section "Linear discriminant analysis." It will be interesting to see if the LDA transform applied in the acoustic feature domain can achieve a similar PV performance as the LDA applied in the model parameter domain.

Restrictions and other applications

For the PV task, due to the restriction of the labeling scheme, there is no phoneme-level pronunciation diagnosis-based experimental study. Another issue is how to let the speech and language therapist assess the patient's ability based on

the confidence scores provided by the proposed PV scenario. The current experimental studies are based on an overall PV performance. However, it is certain that the pronunciation of some phonemes is more difficult for the assessment. Or, from a therapist point of view, some phonemes might be more critical than other phonemes for evaluating the patient's pronunciation ability. Thus, a phoneme-specific PV performance would be required. Besides, the PV experimental study is based on a small Spanish impaired children corpus. It will be interesting to see if the proposed PV scenario can provide useful feedbacks to other languages, or to the nonnative speaker corpus.

References

[1] C. Vaquero, O. Saz, E. Lleida, and W.-R. Rodrǵuez. E-inclusion technologies for the speech handicapped, Proc. ICASSP, Las Vegas, USA, Apr 2008.

[2] N. Schiavetti. Scaling procedures for the measurement of speech intelligibility, Kent RD, ed., Intelligibility in Speech Disorders: Theory, Measurement, and Management, John Benjamins Publishing Company, 1992.

[3] K. Vicsi, P. Roach, A. Öster, Z. Kacic, P. Barczikay, A. Tantos, F. Csatári, Z. Bakcsi, and A. Sfakianaki. A multimedia, multilingual teaching and training system for children with speech disorders, International Journal of Speech Technology, 3(3–4), 289–300, 2000.

[4] T. Bocklet, K. Riedhammer, E. Nöth, U. Eysholdt, and T. Haderlein. Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling, Journal of Voice, 26, 390–397, 2012.

[5] F. Hönig, A. Batliner, and E. Nöth. "Automatic assessment of non-native prosody – annotation, modelling and evaluation," in Proc. ISADEPT, 2012.

[6] A. Escartïn, O. Saz, C. Vaquero, W.-R. Rodrïguez, and E. Lleida. "Comunica framework web site," http://www.vocaliza.es, 2008.

[7] M. Hawley, S. Cunningham, F. Cardinaux, A. Coy, P. O'Neill, S. Seghal, and P. Enderby. "Challenges in developing a voice input voice output communication aid for people with severe dysarthria," in Proceedings of 9th European Conference for the Advancement of Assistive Technoligy in Europe, (San Sebastian, Spain), Oct. 2007.

[8] K. Vicsi, P. Roach, A. Öster, Z. Kacic, P. Barczikay, and I. Sinka. "Speco: A multimedia multilingual teaching and training system for speech handicapped children," in Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-Interspeech), (Budapest, Hungary), September 1999.

[9] O. Saz, S.-C. Yin, E. Lleida, R. Rose, and C. Vaquero. Tools and technologies for computer-aided speech and language therapy, ISCA Speech Communication Journal, 51, 948–967, 2009.

[10] O. Saz, W.-R. Rodrïguez, E. Lleida, and C. Vaquero. "A novel corpus of children disordered speech," in Proceedings of the First Workshop on Child, Computer and Interaction, (Chania, Greece), Oct. 2008.

[11] M. Monfort, and A. Juárez-Sánchez. Registro fonológico inducido (tarjetas gráficas), Ed., Cepe, Madrid, 1989.

[12] E. Alarcos Fonología española. Ed., Gredos, Madrid, 1950.
[13] F. Zhang, C. Huang, F.K. Soong, M. Chu, and R. Wang. Automatic mispronunciation detection for Mandarin, Proc. ICASSP, Las Vegas, USA, Apr 2008.
[14] R. Patel. Phonatory control in adults with cerebral palsy and severe dysarthria, AAC Augmentative and Alternative Communication, 18, 2–11, March 2002.
[15] J.-R. Deller, D. Hsu, and L.-J. Ferrier. On the use of hidden Markov modelling for recognition of dysarthric speech, Computer Methods and Programs in Biomedicine, 35, 125–139, 1991.
[16] S.-C. Yin, and R. Rose. Verifying pronunciation accuracy from speakers with neuromuscular disorders, Proc. InterSpeech, Brisbane, Australia, July 2008, 2008.
[17] S.-C. Yin, R. Rose, O. Saz, and E. Lleida. A study of pronunciation verification in a speech therapy application, Proc. ICASSP 2009, Taipei,, Taiwan, May 2009.
[18] J.-L. Gauvain, and C.-H. Lee. "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," vol. 2, pp. 291–298, Apr. 1994.
[19] C. Leggetter, and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech and Language, 9, 171–185, 1995.
[20] P. Koehn. "Europarl: A parallel corpus for statistical machine translation," in Proceedings of the 10th Machine Translation Summit, (Phuket,Thailand), September 2005.
[21] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.-B.M. no, and C. Nadeu. Albayzin speech database: Design of the phonetic corpus, Proc. Eurospeech, Berlin, Germany, Sept 1993.
[22] S. Goronzy, and R. Kompe. A combined MAP + MLLR approach for speaker adaptation, Proceedings of Sony Research Forum 99, 1, 9–14, 1999.
[23] G. Jang, S. Woo, M. Jin, and C. Yoo. Improvements in speaker adaptation using weighted training, Proc. ICASSP, Hong Kong, China, Apr 2003.
[24] S. Bengio, and J. Mariethoz. A statistical significance test for person authentication, Proc. Odyssey, Toledo, Spain, June 2004, 2004.
[25] E.-R. Delong, D.-M. Delong, and D.-L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, Biometrics, 44(3), 837–845, 1988.
[26] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N.K. Goel, M. Karafiat, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas. Subspace Gaussian mixture models for speech recognition, Proc. ICASSP 2010, Dallas, Texas, USA, Mar 2010.
[27] D. Povey. A tutorial-style introduction to subspace Gaussian mixture models for speech recognition, Tech. Rep. MSR-TR-2009-111, 2009.
[28] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition, Proc. of IEEE, 77(2), 1989.
[29] X. Huang, A. Acero, H. Hon. Spoken Language Processing: A Guide to Theory, System and Algorithm Development. Englewood Cliffs, New Jersey: Prentice Hall, 2001.
[30] R. Rose, S.-C. Yin, and Y. Tang. An investigation of subspace modeling for phonetic and speaker variability in automatic speech recognition, Proc. ICASSP 2011, Prague, Czech Republic, May, 2011.
[31] D. Povey, L. Burget, A. Ghoshal, N. Goel, R.C. Rose, P. Schwartz, and S. Thomas. The subspace Gaussian mixture model – A structured model for speech recognition, Computer Speech and Language, 25(2), 2010.

[32] D. Reynolds, R. Rose. Robust text-independent speaker identification using Gaussian mixture models, IEEE Trans on SAP, 3, 72–83, 1995.
[33] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted Gaussian mixture models, Digital Signal Processing, 10, 19–41, 2000.
[34] A. Dempster, N. Laird, D. Robin. Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B, 1–38, 1977.
[35] R. Redner, H. Walker. Mixture densities, maximum likelihood and the em algorithm, SIAM Review, 26(2), 195–239, 1984.
[36] S. Davis, P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans on ASSP, 28, 357–366, 1980.
[37] S. Young. "The HTK hidden Markov model toolkit: Design and philosophy," tech. rep., Cambridge University Engineering Department, Speech Group, Cambridge, 1993.
[38] S.-C. Yin, R. Rose, and Y. Tang. A study of applying subspace based pronunciation modeling in verifying pronunciation accuracy, Proc. ISSPA, Montreal, Canada, July 2012, 2012.
[39] S.-C. Yin, R. Rose, and Y. Tang. Verifying session level pronunciation accuracy in a speech therapy application, Proc. InterSpeech 2012, Portland, USA, Sept 2012.
[40] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, D. Povey, A. Rastrow, R. Rose, and S. Thomas. "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in Proc. ICASSP 2010, (Dallas, Texas, USA), Mar. 2010.
[41] R.A. Fisher. The use of multiple measurements in taxonomic problems, Annals of Eugenics, 7(7), 179–188, 1936.
[42] L. Lee, and R. Rose. Speaker normalization using efficient frequency warping procedures, Proc. Int. Conf. on Acoust., Speech, and Sig. Proc., Atlanta, GA, May 2003.
[43] K. Beulen, L. Welling, and H. Ney. Experiments with linear feature extraction in speech recognition, in in Proc. Europ. Conf. on Speech Communication and Technology, 1415–1418, 1995.
[44] G. Saon, G. Zweig, and M. Padmanabhan. Linear feature space projections for speaker adaptation, in Proc. ICASSP, 325–328, 2001.

Heejin Kim and Mark Hasegawa-Johnson

5 Communication improves when human or computer listeners adapt to dysarthria

Abstract: This chapter reviews methods that improve the ability of human and machine listeners to understand dysarthric speech. Traditionally, a speaker-oriented approach has been the dominant technique to improve the intelligibility of dysarthric speech. Recent work demonstrates the potential of listeners' role in enhancing intelligibility. For human listeners, a training method called familiarization is evidenced to be effective, especially when the training is structured in a way to maximize perceptual learning. For machine listeners, the accuracy in understanding dysarthric speech can be improved by adaptive machine learning methods, with an initial model that already incorporates information about speakers' characteristic speech patterns. Future direction to optimize the training results for human and machine listeners is discussed.

Keywords: dysarthria, intelligibility, familiarization, ASR, speech perception

5.1 Introduction

Speech and language disorders result from many types of congenital or traumatic disorders of the brain, nerves and muscles [1]. Dysarthria refers to the set of disorders in which unintelligible or perceptually abnormal speech results from impaired control of the oral, pharyngeal or laryngeal articulators. The specific type of speech impairment is often an indication of the neuromotor deficit causing it, therefore speech language pathologists have developed a system of dysarthria categories reflecting both genesis and symptoms of the disorder [1]. The most common category of dysarthria among children and young adults is spastic dysarthria [2], typically characterized by strained phonation, imprecise placement of the articulators, incomplete consonant closure, reduced voice onset time distinctions between voiced and unvoiced stops, distorted vowels, and monotonic or excessive variation of loudness and pitch.

We are interested in spastic dysarthria because it is the most common type of severe, chronic speech disorder experienced by students at the University of

Heejin Kim, Department Linguistics, University of Illinois at Urbana-Champaign, IL
Mark Hasegawa-Johnson, Beckman Institute and Department Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL

https://doi.org/10.1515/9781501513138-005

Illinois, as well as being one of the most common types of dysarthria generally. Spastic dysarthria is associated with a variety of disabilities such as, but not limited to, cerebral palsy and traumatic brain injury [3, 4]. 0.26% of all 7-year-old children in the United States have moderate or severe cerebral palsy, and an additional 0.2% are reported to have mild cerebral palsy [5]. Adults with cerebral palsy are able to perform most of the tasks required of a college student, including reading, listening, thinking, talking and composing text: in our experience, their greatest handicap is their relative inability to control personal computers. Typing typically requires painstaking selection of individual keys. Some students are unable to type with their hands (or find it too tiring), and therefore choose to type using a head-mounted pointer. Many students with noticeable dysarthria are less impaired by their dysarthria, in daily life, than by their inability to use computers.

The speech impairments resulting from spastic dysarthria are neither arbitrary nor unpredictable; indeed, van Santen [6] demonstrated a dynamic systems model of vowel distortion under dysarthria. Table 5.1 lists a number of specific phoneme substitutions errors attested in the literature [7]. As emphasized by the organization of the table, most of the specific impairments reported in the literature can be characterized as imprecision in the implementation of one or two distinctive features; for example, /t/→/k/ is a mistake in the place of articulation of the stop. In order to provide more evidence, the authors of this chapter phonetically transcribed four long recordings from Aronson [8]: a phonetically rich read paragraph (the "grandfather passage"), and three diadokinesis sequences, all read by one male talker with moderate spastic dysarthria. All words in the grandfather passage with nonstandard pronunciation [9] were marked as "errors"; likewise, as were all diadokinesis syllables containing a consonant other than the target consonant

Table 5.1: Phoneme production errors in dysarthria, as reported in [7], listed with the distinctive feature(s) changed by the error [10].

Articulatory deficit	Distinctive feature(s)	Examples
Tongue positioning	[blade]	/t/ vs. /k/
Tongue blade positioning	[anterior]	/ʃ/ vs. /s/
Oral-laryngeal timing	[spread glottis]	/t/ vs. /d/
Degree of closure	[continuant]	/t/ vs. /s/
Manner of closure	[sonorant]	/p/ vs. /m/
Vowel articulation	[advanced tongue root]	/u/ vs. /ʊ/
Lexical stress	[reduced,front]	/æ/ vs. /ə/

(/p/, /t/ or /k/). Table 5.2 lists all substitution errors found in this corpus; deletion errors are not listed, and there were no insertion errors. Tables 5.1 and 5.2 suggest that speech production errors in dysarthria, though random, are not uniformly random: almost all errors are errors of a single distinctive feature.

Table 5.2: Pronunciation errors found in paragraph reading and diadokinesis, one male talker from [8], phonemically transcribed at the University of Illinois.

Phonemes	Count	Distinctive features	Phonemes	Count	Distinctive features
/p/→/b/	31	[spread glottis]	/ŋ/→/n/	1	[blade]
/t/→/d/	24	[spread glottis]	/z/→/n/	1	[sonorant, continuant]
/k/→/g/	19	[spread glottis]	/k/→/ŋ/	1	[sonorant, continuant]
/p/→/m/	2	[sonorant]	/f/→/h/	1	[lips]
/s/→/z/	1	[spread glottis]	/ɑ/→/ə/	1	[reduced]
/z/→/d/	1	[continuant]	/d/→/ɾ/	1	[reduced]
/ʒ/→/z/	1	[anterior]	/ɪ/→/ə/	1	[reduced]

Errors of a single distinctive feature may be enough to confuse a listener who is unfamiliar with dysarthria, but considerable experimental evidence suggests that a familiar listener is able to take advantage of the regularities in the error pattern in order to understand dysarthric speech: higher word and segment identification accuracies were found in familiarized listeners compared to non-familiarized listeners, even for individuals with severe dysarthria.

This chapter is about methods that improve the ability of a listener to understand dysarthric speech. We consider both human and machine listeners. The ability of human listeners to understand dysarthric speech can be improved by *familiarization*, a listener training method in which listeners receive brief, yet structured, exposure to dysarthria. The ability of machine listeners to understand dysarthric speech can be improved by adaptive machine learning methods; the efficacy of such methods can be boosted by starting with an initial model that already incorporates some information about the talker's speech patterns.

5.2 Perceptual adaptation of human listeners

There is an array of factors that can influence listeners' perception of dysarthric speech. Paralinguistic cues, which are independent of the acoustic signal of

dysarthria, can help listeners to decipher dysarthric speech: for example, facial expressions, hand and body gestures, and situational context including knowing the topic of the conversation [11–13]. Of particular interest to us, there is a signal-dependent factor that influences perception: that is, the human listener's adaptability to the acoustic-phonetic content of dysarthric speech. Familiarity can be defined as the listener's previous contact or exposure with an acoustic signal [14], and has been shown to improve the listener's ability to understand the atypical acoustic signal in dysarthria. In order to examine listeners' familiarity effects on the perception of dysarthria, studies have tested a training method, *familiarization*, in which listeners are provided brief exposure to dysarthric speech. The major finding of these studies is that when naïve listeners receive brief exposure to dysarthric speech, their ability to perceive the speech improves [14–20]. Familiarization effects are evidenced not only in dysarthria but also in a variety of atypical speech including both natural and synthetic speech: for example, nonnative accented speech [21–23], speech produced by individuals with hearing impairment [12, 24, 25] and artificially created or altered speech such as synthesized voice, noise-vocoded speech and time-compressed speech [26–33]. These convergent findings of familiarity in the field of speech perception highlight a cognitive-perceptual process called *perceptual learning* in which listeners seem to automatically engage when they encounter atypical speech. By this process, listeners are capable of recognizing speech that sounds deviant from what they know as normal, and are capable furthermore of recalibrating existing speech categories. The result of perceptual adaptation is an improved ability to understand atypical speech that was initially difficult to perceive [34, 35]. Studies on familiarization effects in dysarthria underscore the listeners' flexibility of speech perception and strongly support the potential of familiarization training as an intervention method in dysarthria.

Concerning improved communication through familiarity-induced perceptual learning in dysarthria, two issues are particularly relevant: (1) what is learnable through perceptual adaptation and (2) how to facilitate perceptual learning.

What is learnable: Prior work has striven to understand perceptual learning mechanisms, especially by identifying learning sources [36]. Evidence exists in support of both suprasegmental and segmental learning. For example, when listeners are exposed to sentential prosody in dysarthria during familiarization, their ability to understand the speech improves [15, 18]. Segmental learning through familiarization has been discussed in many studies. The majority of studies reported familiarization benefits based on word intelligibility measures, with the hypothesis that the source of improved word intelligibility would be primarily at the segmental level of perceptual reorganization by which listeners adjust the mapping between acoustic-phonetic information and the phonemic

representation in the language [27, 37–42]. Familiarized listeners in [17] exhibited improved word intelligibility compared to the nonfamiliarized, but two groups did not differ in terms of the lexical boundary error types, indicating that the benefit could be attributed predominantly to the segmental level than prosodic level of strong/weak syllables. Extended support was found in Borrie et al. [37] and Spitzer et al. [14]: that is, more word substitution errors that bore phonemic resemblance to the target for familiarized listeners versus nonfamiliarized listeners. The results drawn from word transcription tasks need to be interpreted with caution since it is not certain how much effect is attributable to the acoustic processing versus high-level lexical or semantic knowledge [34, 36]. In order to find more direct evidence of segmental learning, our work [43, 44] employed consonant identification tasks instead of word transcription tasks. Four American-English native speakers diagnosed with CP (Cerebral Palsy) provided speech data. Their age, gender and intelligibility were as follows: Speaker 1 (58; male, 28%), Speaker 2 (18; male, 39%), Speaker 3 (21; male, 59%) and Speaker 4 (18; female, 62%). A total of 120 listeners (30 listeners/speaker), who had no more than incidental experience with persons with speech disorders, participated in perception experiments. The main finding was that consonant identification scores were higher in familiarized listeners, compared to nonfamiliarized listeners. The effect size (the magnitude of difference between pre- vs. post-familiarization) was medium to large for all speakers.

Researchers have investigated acoustic correlates of speech perception in dysarthria to identify acoustic metrics that can predict listeners' performance in understanding dysarthria as well as to better understand the production deficits in dysarthria. A substantial number of acoustic studies in dysarthria including our work [45–47] have shown that there is a strong association between acoustic measures of segmental clarity, on the one hand, and reduced speech intelligibility on the other hand, supporting the perceptual consequences of acoustic deviances in dysarthria. However, to our knowledge, no study to date has investigated acoustic correlates of perceptual learning effects in dysarthria. Perceptual learning is influenced by the nature of acoustic deviances [48], but most familiarization studies in dysarthria have examined only word intelligibility, failing to address acoustic sources for the familiarization benefit. On the other hand, a study on synthesized speech [27] examined the training process related to acoustic properties of the speech signal and reported evidence for perceptual learning of acoustic features. It is unknown whether the finding can be generalized to dysarthria. Broadmore [49] a study on the familiarization effect in dysarthria associated with Parkinson's disease (PD), found that familiarization training improved intelligibility for two out of three speakers. The author speculated on the possibility of an acoustic explanation for the

reduced perceptual learning that occurred for one of the speakers. In other words, the degree of perceptual learning success might be dependent on the idiosyncratic acoustic characteristics of the speech. We are currently investigating acoustic deviances in voiceless sibilant fricative productions in dysarthria and their relevance to a listener's perceptual accuracy. Results show that regression functions relating moment measures and identification scores of /s/ in post familiarization are significant: in fact, the ability of listeners to understand dysarthric speech is almost completely explained (over 80%) by the way in which dysarthria changes the acoustic characteristics of speech segments. The least amount of learning after familiarization was found for /ʃ/ (the first sound in the word "ship") when its spectral measures were overlapped with those for /s/ (the first sound in the word "sip"). This preliminary finding suggests that perceptual learning might be impeded when acoustic deviances are in the form of phonemic substitution, less so in the form of nonprototypical, noisy version of normal production.

How to facilitate perceptual learning: While most studies reported a familiarization benefit in dysarthria, the size of intelligibility improvement varied across studies. Methodological differences such as dysarthria severity and the listener's experience level to dysarthric speech may be responsible for the different degrees of the benefit, but in particular, differences in familiarization conditions suggest that a certain training method can be more effective in facilitating perceptual learning for listeners. Three different training conditions appear in prior work: an *active* (or sometimes called e*xplicit*) condition, a *passive* condition and no separate familiarization with sequential repetition of experimental tasks. Several studies examined an active condition, in which listeners were familiarized with both audio signals and a written transcript, and reported a significant effect of familiarization [14, 15, 17–19]. A passive familiarization method, in which listeners were presented with only an audio signal, resulted in no significant effect. The third type, that is, no separate familiarization but only sequential repetition of experimental tasks, was tested in Hustad and Cahill [20], and resulted in a significant improvement in intelligibility. Cross-study comparisons between active versus passive conditions in Borrie et al. [37] reported a greater benefit of an active condition over a passive condition in terms of word transcription, phonemic resemblance and syllabic stress perception.

We have conducted integrated studies that compared the efficacy of all three conditions (the active- vs. passive- vs. no separate familiarization conditions) for both word and consonant intelligibility [40, 41, 43, 44]. Thirty listeners were recruited per speaker, and were randomly assigned to one of the three conditions:

passive versus active versus control. Listeners in the control condition received no exposure to the audio signal prior to identification tasks, but performed multiple identification tasks. Thus, the control condition in our experiments was similar to the experimental condition in Hustad and Cahill [20]. The main finding was that an active condition was superior to other conditions in both word and consonant intelligibility. The advantage was manifested in terms of two aspects: the magnitude and the rapidity of improvement. First, for all speakers, listeners in the active familiarization condition exhibited higher intelligibility scores compared to the other two conditions: for word identification scores, the improvement amount ranged from 21% to 28%, and for consonant identification scores, it ranged from 5% to 19%. Furthermore, a larger degree of listener's improvement was found for speakers with severe dysarthria. Second, only the listeners in an active condition improved consonantal perception as early as immediately after the first familiarization training while the other groups reached significant increases at a later session or no improvement at all. Rapid learning of consonants in the active condition demonstrates that the use of orthographic transcripts facilitate listeners' segmental learning. We note that echoing the finding in Hustad and Cahill [20], even without separate familiarization, the repetition of test material resulted in intelligibility improvements, highlighting the plasticity of speech perception; however, the extra gain and the rapid improvement in the active condition highlighted the efficacy of the active condition over others. For the purpose of dysarthria management, an optimal intervention method should be sought that considers not only the magnitude but also the rapidity of improvement. Thus, an active training condition that expedites perceptual learning would be more desirable compared to other conditions that require more training time.

As discussed in Hustad and Cahill [20], caution is needed when interpreting familiarization effects. It is difficult to know how much of the observed improvement reflects a true learning effect, because we cannot exclude the possibility that practice improves one's ability to simply perform the experimental task, without any specific perceptual learning benefit. Our work included different familiarization conditions in the experimental design, thus the key finding in our work is that the additional effect in the active condition compared to other conditions can be attributed to the availability of written material in the active condition. Importantly, the active condition showed a long-term effect as well: 1-month delayed test scores were higher than pre-familiarization scores. This finding extends support for the claim that perceptual learning is not a temporary adjustment but rather a long-lasting effect [36], similar to other work in dysarthria [37], in synthetic speech [31, 50] and in nonnative speech [51].

5.3 Adaptive automatic speech recognition

It is counter-intuitive to imagine that a person with a speech pathology might be able to use a spoken computer interface more effectively than a keyboard, but it is often true. Many types of speech pathology accompany severe motor disorders, and the large-scale motor disorder often makes the use of a keyboard extremely difficult. Several studies have demonstrated that adults with dysarthria are capable of using automatic speech recognition (ASR), and that in some cases, human–computer interaction using ASR is faster than interaction using a keyboard [52–55].

The utility of ASR for dysarthric speech is limited by the average characteristics of dysarthric speech, and by its variability. For example, Fager [56] investigated the durations of single words and phonemes as produced by 10 participants with dysarthria, and 10 control participants. The study also examined the relationships between word intelligibility and word duration, and between word intelligibility and variability for the participants with dysarthria. Results showed statistically significant differences of word and phoneme durations between the participants with versus without dysarthria. Because of these significant population differences, it is difficult for speakers with dysarthria and speakers without dysarthria to use the same ASR: an ASR trained without dysarthria fails to correctly recognize the longer-duration phonemes produced by a speaker with dysarthria.

Many speakers with dysarthria have solved the problem of population difference by using speaker-dependent or speaker-adaptive ASR. A speaker-dependent ASR is trained entirely using speech read by the intended user; a speaker-adaptive ASR is initialized using a database of many speakers, but then adapted to the speech of the intended user. Unfortunately, the utility of speaker-dependent and speaker-adaptive ASR for speakers with dysarthria is limited by variability, from one sentence to the next, in the speech produced by a speaker with dysarthria. Parker et al. [57] found the consistency of phonetic representation over time to be crucial for accurate ASR. Blaney and Wilson [58] noted that intra-speaker variability is a correlate of dysarthria, especially with regard to voice onset time of stop consonants, vowel duration, and fricative duration. Speech from speakers with moderate dysarthria exhibited greater variability across all acoustic measures, compared to the speakers with mild dysarthria and the speakers without speech pathology. A “minimal pair” can be defined as a pair of words that differ in only one distinctive feature; Blaney and Wilson [58] document several cases in which dysarthria erased the acoustic distinction between minimal pairs.

Raghavendra et al. [59] compared recognition accuracy of a speaker-adaptive system and a speaker-dependent system. They found that the speaker-adaptive system adapted well to speech with mild or moderate dysarthria, but the recognition scores were lower than for a speaker without dysarthria. The subject with severe dysarthria was able to achieve better performance with the speaker-dependent system than with the speaker-adaptive system. These findings were also supported by Rudzicz [60], who compared the performance of speaker-dependent and semi-adaptive systems on the Nemours database [61] by varying independently the amount of data for training and the number of Gaussian components used for modeling the output probability distributions. Doyle et al. [62] asked six speakers with dysarthria, and six without, to read a list of 70 words once in each of five training sessions. They found that the word recognition accuracy of a speaker-adaptive ASR increased rapidly after the first training session, then increased more gradually during training sessions two through five.

Speakers with dysarthria may have trouble training a speaker-dependent or speaker-adaptive ASR because of the great amount of training data required. Reading a long training passage can be very tiring for a speaker with dysarthria. Speaker adaptive ASR may require less training data than speaker-dependent ASR, and is therefore a useful method to provide ASR without over-tiring the user. However, even if one applied such adaptation methods, there exists a second obstacle: a speaker-adaptive system is initialized using recordings of other speakers, who usually do not have dysarthria. Speaker-adaptive systems may therefore converge more slowly to the voice of a speaker with dysarthria than to the voice of one without dysarthria. Conventional adaptation techniques such as maximum likelihood linear regression [63, 64], maximum a posteriori (MAP) adaptation [65, 66], or structured MAP adaptation [67] do not explicitly model the mismatch between the speech characteristics of the target speaker population and those of the population used to train the to-be-adapted acoustic model.

The MAP adaptation algorithm of [65] was used by Sharma et al. [68–70] to create a series of increasingly refined speech recognizers for speakers with dysarthria. First, [70] proposed a relatively standard MAP adaptation algorithm, which was later dubbed SI-MAP (MAP initialized using a speaker independent = SI speech recognizer). The SI-MAP algorithm is initialized with a hidden Markov model trained using the TIMIT speaker-independent automatic speech recognition database [71]. After being so initialized, the model is adapted to the speech of a speaker with dysarthria. For the in-domain speech, Sharma et al. used the UA-Speech corpus [72], which contains recordings of 16 speakers informally diagnosed with spastic dysarthria. Each speaker recorded three blocks of isolated words: each block contained the same 155 core words, plus 100 “uncommon

words" that differed across blocks. The core words included the 10 digits ("zero" through "nine"), the 26 letters of the international radio alphabet ("alpha, bravo, charlie, ... "), 19 computer commands ("command, enter, paragraph, ... ") and the 100 most common words in the Brown corpus of written English ("is, it, ... "). The uncommon words were selected from children's novels digitized by Project Gutenberg (e.g., Wizard of Oz, Peter Pan) to maximize phoneme-sequence diversity. Digits and common words were primarily composed of monosyllables, computer commands and radio alphabet letters of bisyllables, and uncommon words of polysyllabic words (more than half of the uncommon words were trisyllabic or longer). Each speaker recorded a total of 765 words, including 455 distinct words.

A second paper [68] proposed initializing MAP using a speaker-dependent background model (SDB). The SDB is trained entirely using data recorded by the speaker with dysarthria. None of the speakers in the UA-Speech corpus recorded enough data to train a complete speaker-dependent automatic speech recognizer, however, so the SDB used a completely new approach. Rather than trying to distinguish between the different phonemes produced by the intended speaker, the SDB learns, instead, a Gaussian mixture model (GMM: a kind of smoothed histogram) representing the set of all speech sounds produced by the intended speaker. The SDB is a model of the general characteristics of the target population speaker: it does not learn any patterns that can discriminate between phones/words but is intended to capture aspects of time-frequency variation that depend on the speaker rather than on what was spoken by him/her. Because it does not try to distinguish among the different phonemes produced by the intended speaker, the SDB can be learned using a much smaller training dataset. After being trained in this way, the SDB is then cloned, and each clone is MAP-adapted to the examples of one particular phoneme. In this way, relatively rich models of each phoneme can be trained, using an extremely small amount of training data per talker.

In general, because of the population mismatch, the SI and SDB models will have very different parameters. A third paper [69] proposed using the SI and SDB models to define a continuum of different initial models. The "background-interpolated" (BI) model is formulated as a linear interpolation between the SI and SDB models. Each of the parameters of the speech recognizer is linearly interpolated between the corresponding parameter of the SI model, and that of the SDB model. Three ASR configurations were studied. SI-MAP is initialized using an SI model, then MAP-adapted to a speaker with dysarthria using the MAP adaptation algorithm of [65]. SDB-MAP is initialized using the speaker-dependent background model, then cloned to generate initial models of each phoneme, which are then MAP-adapted to the phonemes of the speaker with dysarthria. BI-MAP is initialized using the background-interpolated model, then MAP-adapted. These

systems were developed for each of the 16 UA-Speech speakers and employed word-internal, context dependent triphone hidden Markov models, with 3 hidden states and observations modeled using a mixture of 32 Gaussians.

Table 5.3 lists the word recognition accuracy (WRA: higher is better) scores for each UA-Speech speaker, for the three system configurations, in increasing order of the speakers' average intelligibility. SDB-MAP performs quite poorly: apparently it is not possible to initialize a speech recognizer with an initial model that completely ignores the differences between phonemes. On the other hand, BI-MAP does quite well: it outperforms the SI-MAP system for 12 of the 16 speakers (all except F03, M05, M11, M08).

Table 5.3: Word recognition accuracy (WRA) of three ASR systems, after adaptation to each speaker in the UA-Speech corpus. Intelligibility = % of words correctly transcribed by human listeners who had no more than incidental experience with persons having speech disorders. SI-MAP = adapted from a TIMIT-based speaker independent model. SDB-MAP = adapted from a GMM that models speaker characteristics but does not differentiate phonemes. BI-MAP = adapted from a model linearly interpolated between SI and SDB.

Speaker	Intelligibility (%)	WRA (%), SDB-MAP	WRA (%), SI-MAP	WRA (%), BI-MAP
M04	2	0.6	3.0	3.2
F03	6	7.1	21.4	19.8
M12	7	4.6	14.8	16.4
M01	17	4.4	12.6	14.1
M07	28	17.7	39.0	42.5
F02	29	17.5	29.0	31.1
M06	39	13.5	36.8	39.3
M16	43	5.2	26.5	32.1
M05	58	15.4	38.1	36.8
M11	62	10.2	29.8	28.9
F04	62	10.6	32.9	34.8
M09	86	25.5	63.9	70.0
M14	90	29.1	60.7	64.1
M10	93	52.3	73.1	74.2
M08	95	21.2	69.6	66.9
F05	95	57.9	78.7	80.7

5.4 Summary: Adaptation by humans and machines

Findings in our familiarization studies support a listener's capacity to learn the atypical characteristics of dysarthric speech. Our work demonstrates the potential of using familiarization as a listener-oriented intervention technique for dysarthria management. Research efforts should continue to further exploit perceptual learning mechanisms and to fully utilize the potential of listener training for developing an optimal protocol of familiarization. Given the ample evidence of segmental benefits, further investigation is warranted on how to structure training materials to promote segmental learning. More work is also needed to investigate what factors influence the longevity of perceptual learning, and the ways in which the detailed acoustic characteristics of dysarthric speech are related to perceptual learning effects induced by familiarization. We note that the active condition did not use any explicit method that guides listeners to attune to specific phonetic or phonological features. A minimal-pair approach, in which word pairs that contrast by a single phoneme are presented during training, has been used in listener training for synthesized speech and is a common model for intervention techniques in second language learning and children's phonological delay [73–75]. Further investigation is necessary, to test whether words in minimal pairs will expedite segmental learning in dysarthria, by permitting listeners to have explicit experience in differentiating distinctive features and to capitalize on that experience more efficiently [48]. Perceptual learning results should be evaluated in terms of the amount of communication improvement (effectiveness), rapidness of the improvement (efficiency) and longevity of learning (robustness).

Speakers with dysarthria can sometimes use spoken language human–computer interfaces more effectively than they can use a keyboard. Just like human listeners, machines are more effectively able to understand dysarthric speech if given some training materials that include dysarthria. If it were possible to train an ASR using a large quantity of speech produced by the intended user (speaker-dependent training), then the ASR would probably perform pretty well; unfortunately, speaker-dependent training is usually not possible, because speaker-dependent requires a great deal of speech, and speakers with dysarthria (like the rest of us) get tired before they complete speaker-dependent training regimen. Speaker-adaptive training is possible, but the resulting accuracy depends on the way in which the speaker-adaptive system is initialized. An ASR initialized using the speech of individuals (SI-MAP) without dysarthria performs poorly. As an alternative, we proposed initializing the ASR using a

model of the target speaker's voice, without any specified phoneme distinctions (SDB-MAP), but SDB-MAP method yields even worse results than SI-MAP. The best results are achieved using background-interpolated MAP (BI-MAP), which is initialized using a linear interpolation between the parameters of a speaker-independent ASR, and the parameters of a speaker-dependent but phoneme-independent ASR.

References

[1] J. Duffy. Motor Speech Disorders, Boston, MA, Elsevier Mosby, 1995.
[2] R.J. Love. Childhood Motor Speech Disability, Allyn and Bacon, Boston, 1992.
[3] F. Darley, and A. Aronson. Differential diagnostic patterns of dysarthria, Journal of Speech and Hearing Research, 12, 246–269, 1969.
[4] F. Darley, and A. Aronson. Motor Speech Disorders, Philadelphia, PA, W.B. Saunders Co, 1975.
[5] M.C. Leske. Prevalence estimates of communicative disorders in the U.S., Speech disorders," ASHA Leader, 23, 3, 1981.
[6] J. Van Santen. Applying speech / language technologies to communication disorders: New challenges for basic research, Unpublished presentation delivered at Johns Hopkins University, Baltimore, MD, USA, 2004.
[7] R.D. Kent, G. Weismer, J.F. Kent, H.K. Vorperian, and J.R. Duffy. Acoustic Studies of dysarthric speech: methods, progress, and potential, Journal of Communication Disorder, 32, 141–186, 1999.
[8] A. Aronson. Dysarthria Differential Diagnosis, Rochester, MN, Mentor Seminars S.L.P, 1999.
[9] P. Kingsbury, S. Strassel, C. McLemore, and R. MacIntyre. LDC97L20: CALLHOME American English Lexicon (PRONLEX), Philadelphia, PA, Linguistic Data Consortium, 1997.
[10] K.N. Stevens. Acoustic Phonetics, Cambridge, MA, MIT Press, 1999.
[11] K.C. Hustad. Contribution of two sources of listener knowledge to intelligibility of speakers with cerebral palsy, Journal of Speech, Language, and Hearing Research, 50, 1228–1240, 2007.
[12] R. Monsen. The oral speech intelligibility of hearing impaired talkers, Journal of Speech and Hearing Disorders, 48, 286–296, 1983.
[13] K.M. Yorkston, P.A. Dowden, and D.R. Beukelman. Intelligibility measurement as a tool in the clinical management of dysarthric speakers, R. Kent, Ed., Intelligibility in speech disorders: Theory, measurement, and management, 265–285, Philadelphia, John Benjamins, 1992.
[14] S.M. Spitzer, J.M. Liss, J.N. Caviness, and C. Adler. An exploration of familiarization effects in the perception of hypokinetic and ataxic dysarthric speech, Journal of Medical Speech-Language Pathology, 8, 285–293, 2000.
[15] J. D'Innocenzo, K. Tjaden, and G. Greenman. Intelligibility in dysarthria: Effects of listener familiarity and speaking condition, Clinical Linguistics and Phonetics, 20(9), 659–675, 2006.

[16] J.M. Garcia, and M.P. Cannito. Influence of verbal and nonverbal contexts on the sentence intelligibility of a speaker with dysarthria, Journal of Speech and Hearing Research, 39(4), 1996.
[17] J.M. Liss, S.M. Spitzer, J.N. Caviness, and C. Adler. The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. Journal of the Acoustical Society of America., 11:2(6), 3022–3030, 2002.
[18] K. Tjaden, and J.M. Liss. The influence of familiarity on judgments of treated speech, American Journal of Speech Language Pathology, 4(1), 39–48, 1995a.
[19] K. Tjaden, and J.M. Liss. The role of listener familiarity in the perception of dysarthric speech, Clinical Linguistics Phonetics, 9(2), 139–154, 1995b.
[20] K.C. Hustad, and M.A. Cahill. Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech, American Journal of Speech-Language Pathology, 12, 198–208, 2003.
[21] A.R. Bradlow, and T. Bent. Perceptual adaptation to non-native speech, Cognition, 106(2), 707–729, 2008.
[22] S. Gass, and E.M. Varonis. The effect of familiarity on the comprehensibility of nonnative speech, Language Learning, 34, 65–89, 1984.
[23] S.A. Weill. Foreign accented speech: Encoding and generalization, Journal of the Acoustical Society of America, 109, 2473–2473, 2001.
[24] L.W. Ellis, and S.A. Beltyukova. Effects of training on naïve listeners' judgments of the speech intelligibility of children with severe-to-profound hearing loss, Journal of Speech Language and Hearing Research, 51, 1114–1123, 2008.
[25] J.L. Loebach, D.B. Pisoni, and M.A. Svirsky. Effects of semantic context and feedback on perceptual learning of speech processed through an acoustic simulation of a cochlear implant, Journal of Experimental Psychology Human Perception Performance, 36(1), 224–234, 2010.
[26] M.H. Davis, I.S. Johnsrude, A. Hervais-Adelman, K. Taylor, and C. McGettigan. Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences, Journal of Experimental Psychology General, 134(2), 222–241, 2005.
[27] A.L. Francis, H.C. Nusbaum, and K. Fenn. Effects of training on the acoustic phonetic representation of synthetic speech, Journal of Speech, Language, and Hearing Research, 50, 1445–1465, 2007.
[28] J.D. Golomb, J.E. Peelle, and A. Wingfield Effects of stimulus variability and adult aging on adaption to time-compressed speech, Journal of the Acoustical Society of America, 121(3), 1701–1708, 33, 2007.
[29] D. McNaughton, K. Fallon, J. Tod, F. Weiner, and J. Neisworth. Effect of repeated listening experiences on the intelligibility of synthesized speech, Augmentative and Alternative Communication, 10, 161–168, 1994.
[30] C. Pallier, N. Sebastian-Gallés, E. Dupoux, A. Christophe, and J. Mehler. Perceptual adjustment to time-compressed speech: A cross-linguistic study, Memory & Cognition, 26(4), 844–851, 1998.
[31] E.C. Schwab, H.C. Nusbaum, and D.B. Pisoni. Some effects of training on the perception of synthetic speech, Human Factors, 27, 395–408, 1985.
[32] N. Sebastian-Galles, E. Dupoux, A. Costa, and J. Mehler. Adaptation to time compressed speech: phonological determinants, Percept Psychophys, 62, 834–842, 2000.

[33] H. Venkatagiri. Effect of sentence length and exposure on the intelligibility of synthesized speech, Augmentative and Alternative Communication, 10(2), 96–104, 1994.
[34] S.L. Mattys. Speech perception, Daniel R, ed, The Oxford Handbook of Cognitive Psychology, Oxford, Oxford University Press, 2011, 2011.
[35] N. Miller. Measuring up to speech intelligibility, International Jouranal of Language Communication Disorder, 48, 601–612, 2013.
[36] S.A. Borrie, M.J. McAuliffe, and J.M. Liss. Perceptual learning of dysarthric speech: A review of experimental studies, Journal of Speech, Language, and Hearing Research, 290–305, 2012a.
[37] S.A. Borrie, M.J. McAuliffe, J.M. Liss, G.A. O'Beirne, and T.J. Anderson. Familiarisation conditions and the mechanisms that underlie improved recognition of dysarthric speech, Language and Cognitive Processes, 27(7–8), 1039–1055, 2012b.
[38] E. Dupoux, and K. Green. Perceptual adjustment to highly compressed speech: Effects of talker and rate changes, Journal of experimental psychology. Human perception and performance, 23(3), 914–927, 1997.
[39] S.L. Greenspan, H.C. Nusbaum, and D.B. Pisoni. Perceptual learning of synthetic speech, Journal of Experimental Psychology, Learning, Memory and Cognition, 14(3), 421–433, 1988.
[40] H. Kim, and S. Nanney. Familiarization effects on word and phoneme transcriptions of dysarthric speech, ASHA Convention, Chicago, IL, November 14–16, 2013, 2013.
[41] H. Kim, and S. Nanney. Familiarization effects on word intelligibility in dysarthric speech, Folia Phoniatrica et Logopaedica, 66(5), 258–264, 2014a. PMC4341980 .
[42] D.B. Pisoni, S.E. Lively, and J.S. Logan. Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception, Nusbaum HC, Goodman J, eds., The development of speech perception: The transition from speech sounds to spoken words, Cambridge, MA MIT Press, 121–166, 1994.
[43] H. Kim, and S. Nanney. (2014b). Familiarization effects on dysarthric speech perception: Evidence of enhanced segmental perception. Motor Speech Conference, Sarasota, FL, February 27-March 2, 2014.
[44] H. Kim. Familiarization effects on consonant intelligibility in dysarthric speech, Folia Phoniatrica et Logopaedica, 67(5), 245–252, 2015.
[45] H. Kim, M. Hasegawa-Johnson, and A. Perlman. Vowel contrast and speech intelligibility in Dysarthria, Folia Phoniatrica et Logopaedica, 63(4), 187–194, 2008a.
[46] H. Kim, and M. Hasegawa-Johnson. Temporal and spectral characteristics of fricatives in dysarthria, The162nd Meeting of the Acoustical Society of America, San Diego, California, October 31 – November 4 2011.
[47] H. Kim, and S. Nanney. Relationship between the spectral characteristics of fricatives and familiarization-induced intelligibility enhancement, ASHA Convention, Orlando, FL, November 20–22, 2014, 2014c.
[48] E.J. Gipson, Principles of perceptual learning and development, New York, Prentice Hall College Div, 1969.
[49] S. Broadmore. (2011). Listener Strategies in the Perception of Dysarthric Speech: A thesis submitted in partial fulfilment of the requirements for the Degree of Master of Speech Language Therapy, Department of Communication Disorders, University of Canterbury, 2011.
[50] F. Eisner, and J.M. McQueen. The specificity of perceptual learning in speech processing, Perception & Psychophysics, 67(2), 224–238, 2005.

[51] K. Nishi K, D. Kewley-Port, Training Japanese listeners to perceive American English vowels: influence of training sets, *Journal of Speech, Language, and Hearing Research*, 50, 1496–1509, 2007.
[52] H.P. Chang Speech input for dysarthric users, Journal of the Acoustical Society of America, 2aSP7, 1993.
[53] K. Hux, J. Rankin-Erickson, N. Manasse, and E. Lauritzen. Accuracy of three speech recognition systems: Case study of dysarthric speech, Augmentative and Alternative Communication, 16(3), 186–196, 2000.
[54] E. Sanders, M. Ruiter, L. Beijer, and H. Strik. Automatic Recognition of Dutch Dysarthric Speech: A Pilot Study, ICSLP, 2002.
[55] N. Thomas-Stonell, A.-L. Kotler, H.A. Leeper, and P.C. Doyle. Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition, Augmentative and Alternative Communication, 14(1), 51–56, 1998.
[56] S.K. Fager. Duration and variability in dysarthric speakers with traumatic brain injury. Ph.D. thesis, University of Nebraska, Lincoln, 2008.
[57] M. Parker, S. Cunningham, P. Enderby, M. Hawley, and P. Green. Automatic speech recognition and training for severely dysarthric users of assistive technology: The STARDUST project, Clinical Linguistics & Phonetics, 20, 149–156, 2006.
[58] B. Blaney, and J. Wilson Acoustic variability in dysarthria and computer speech recognition, Clinical Linguistics & Phonetics, 14, 307–327, 2000.
[59] P. Raghavendra, E. Rosengren, and S. Hunnicutt. An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems, Augmentative and Alternative Communication, 17(4), 265–275, 2001.
[60] F. Rudzicz. (2007). Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech, in: Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility, ACM. p. 256.
[61] X. Menendez-Pidal, J.B. Poliko, S.M. Peters, J.E. Leonzio, and H.T. Bunnell. 1996. The Nemours database of Dysarthric Speech, in: Proceedings of the Fourth International Conference on Spoken Language Processing, pp. 1962–1965.
[62] P.C. Doyle, and H.A. Leeper. Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility, Journal of Rehabilitation Research and Development, 34(3), 309–316, 1997.
[63] V. Digalakis, D. Rtischev, and L. Neumeyer. Speaker adaptation using constrained estimation of Gaussian mixtures, IEEE Transactions on Speech and Audio Processing, 3(5), 357–366, 1995.
[64] C. Leggetter, and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech and Language, 9(2), 171–185, 1995.
[65] J. Gauvain, and C. Lee. (1991). Bayesian learning of Gaussian mixture densities for hidden Markov models. In Proceedings of the DARPA Speech and Natural Language Workshop 272–277.
[66] J. Gauvain, and C. Lee. (1992). MAP estimation of continuous density HMM: Theory and applications. In Proceedings of the DARPA Speech and Natural Language Workshop, 185–190.
[67] K. Shinoda, and C. Lee. (1997). Structural MAP speaker adaptation using hierarchical priors. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 381–388.

[68] H.V. Sharma, and M. Hasegawa-Johnson. Adapting Acoustic Models to New Speaker Populations using in-domain Background Models, Interspeech, 2011.
[69] H.V. Sharma, and M. Hasegawa-Johnson. Acoustic model adaptation using in-domain background models for dysarthric speech recognition, Computer Speech and Language, 27(6), 1147–1162, 2013.
[70] H.V. Sharma, M. Hasegawa-Johnson, J. Gunderson, and A. Perlman. (2009). Universal Access: Speech Recognition for Talkers with Spastic Dysarthria. Interspeech 2009, paper 42862, 1–4.
[71] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, and V. Zue. TIMIT Acoustic-Phonetic Continuous Speech Corpus, Philadelphia, PA, Linguistic Data Consortium, 1993.
[72] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame. (2008b). Dysarthric speech database for universal access research. In Proceedings of Interspeech, Brisbane, Australia.
[73] J.A. Barlow, and J.A. Gierut. Minimal Pair Approaches to Phonological Remediation, Updates in Phonological Intervention Ph.D. Seminars in Speech and Language, 23(1), 57–68, 2002.
[74] J.A. Gierut. Enhancement of learning for children with phonological disorders, Slifka J, Manuel S, Matthies M, eds, From Sound to Sense: 50+ Years of Discoveries in Speech Communication, Cambridge, Mass, MIT Press, 164–172, 2004.
[75] K. Nishi, and D. Kewley-Port. Non-native speech perception training using vowel subsets: Effects of vowels in sets and order of training. Journal of Speech, Language, and Hearing Research, 51(6), 1480–1493, 2008.

Kirtana Sunil Phatnani and Hemant A. Patil

6 Role of music on infant developments

Abstract: Sound plays a crucial role in the development and evolution of *nature*, where animals protect their species from the other animals via alarming sounds and learn to identify their species. In human beings, the linguistic development takes place after the infant is 2 years old, before which soothing music forms a part of their early arrival in this world. It helps calm them and put them to sleep. This depicts the close connection of infants with music. We explore how the neural development of the brain and the healthy growth of the body are improved by a piece of music played in the neonatal intensive care unit (NICU). Analyzing previous studies on voice lullabies played to infants on the NICU, we found a greater average weight gain of 79 g over a 3 day time period for the preterm infants subjected to music and a 62 g weight gain of infants without music. We also observe that the total stay in the NICU of the preterm infants reduced by 5 days, who were subjected to music when compared with the preterm infants not subjected to music. Conducting the study on 40 adults, the blood pressure, heart rate, and oxygen saturation were measured, which stabilizes with the onset of music. We observe studies compiling all the studies over the decades of music therapy incorporated in the NICUs, showing significantly the positive effects of music therapy. Furthermore, we discern that the dopamine-based mechanism present in our brain is crucial in the early development years of the infant, and in case of not receiving enough love, care, and attention from the mother or the family, the child develops the neurobehavioral disorder, such as attention-deficit hyperactivity disorder. Music plays its role in activating the same dopamine-based learning behavior while listening to music, thereby allowing the child to be treated for such diseases. Furthermore, the neuroplasticity of the brain is improved. We also construct an Upper Confidence Bound Reinforcement Learning algorithm to model the dopamine-based reward system in our brain, through which we observe that simpler the note repetition structure, smaller the learning curve of the song. All these aspects form a scientific base in using music for the cure of medical illnesses related to the brain, and behavior in the form of music therapy.

Keywords: Infant brain development, Maslow's Pyramid, Music Therapy, Dopamine-based Learning, Reinforcement Learning

Kirtana Sunil Phatnani, Hemant A. Patil, Speech Research Lab, DA-IICT, Gujarat, India.

https://doi.org/10.1515/9781501513138-006

6.1 Introduction

A sound can be of many types but what exactly it is can be either defined technically in terms of the audible range of wave frequencies perceived by a human ear, or it can be said to characterize particular information known and unknown to us, each characterizing a stimulus for our brain calling for attention. Our response to a sound stimulus is among the fastest in our brain. We also observe that every visual is guided by a trailing background sound in advertisements, directed dramas, and horror movies, which attract our attention. This influences us to pay attention and connect the sounds to emotions we recognize. Recently, in machine learning literature of speech applications, it has been observed that the models trained using attention-based layers perform better [1]. In the study of animal behavior, a bird's young one is born without its eyes fully developed and identifies the mother's presence solely by her chirp to which it responds by its chirp when inside the egg, and when outside the egg, it responds by opening its beak for the mother bird to feed the young one with food. It responds in this manner only to the specific bird's chirp; if it does not recognize the chirp, it does not react in this manner. This allows the mother bird to ensure that the baby bird is developing well inside the egg and ensure its safety by its reaction to specific sounds when the young one is outside the egg [2]. Similarly, a mother cobra has a very sensitive hearing mechanism to any sound in the forest up to a certain distance; in case it hears any sound approaching its eggs, it wraps the eggs by curling around them and attacks on any animal that approaches it. In this way, the safety of the child is ensured by the mother. In human beings, when the infant is in the mother's womb, the inner ear is the first sense organ that develops fully in the womb (i.e., immediately after 6 weeks of conception, first sensory mechanism that is active for the fetus is hearing mechanism). This does not indicate that the infant can hear the sound as we do. Due to the coverings over it and fluids around it, it hears *muffled* voices. Furthermore, the infant in the 18th week can hear high-pitched voices at first and then develops the mechanism to recognize voices by the 26th week [3]. This allows the infants to recognize their mother's voice much before they are born. We discuss in detail the cognitive abilities of the brain in Section 6.3, to understand in depth why do certain diseases occur in infants, and which portion of the brain is responsible for what kind of response. For example, the positive effect of the mother's voice on the infant is produced by the limbic system in the brain. Furthermore, the voice of the mother also improves the health of a preterm born infant, reduces its cortisol levels, the stress hormone, and improves its oxytocin levels, social bonding hormone [4]. The mother's voice surrounds the infants in the form of conversations or lullabies that are soothing. We investigate in Section 6.4 as to how do this soothing lullabies provide a musical stimulus to the

hearing mechanism of the infant, and help increase the vital signs of preterm infants kept in the neonatal intensive care units (NICU).

During the development of the neural connections in the brain in the third trimester, the infant can react to certain stimuli in the form of sound and touch. The infant responds to the loud noises in the womb. This signifies the development of the amygdala responsible for the feeling of fear. All of these developments correspond to the building of a dense connection between the environment with its active sensory organs as stimulus and the infant's response in return to that stimulus. The neonatal and postnatal development of the brain is responsible for these early cognitive abilities [5]. Due to the advances in the images of the development occurring in the fetus, the underlying abnormalities if present are also detected before birth. The development of the dopaminergic neurons in the brain can be observed and analyzed for detection and cure of Parkinson's disease [6]. This forms the basis for computing the underlying learning mechanism in our brain (Section 6.5). Hence, we can summarize a trifold approach by studying the impact of music (as shown in Figure 6.1).

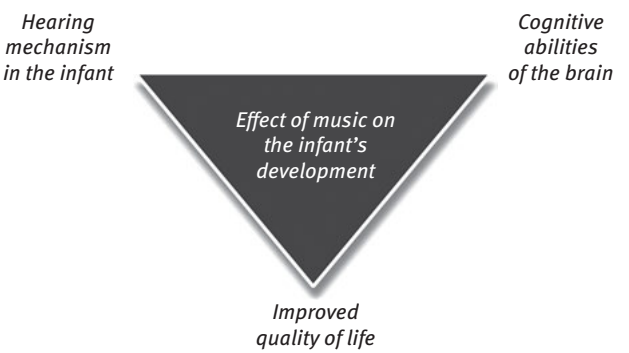

Figure 6.1: Trifold approach of proposed study.

6.2 Cognitive abilities of the brain

The development of the brain plays a crucial role in forming the life of the infant. The layered parts of the brain function to control our autonomous nervous system via the brain stem, controlling our breathing, heart rate, and blood pressure. The cerebellum plays its crucial role in learning and development of an activity, for example, learning how to balance on uneven terrain by corrective learning [7]. The limbic system provides a mechanism to interpret the environment around

us and localize it to our experience of the environment, providing us with a motivation in learning and behavioral patterns that are important for social interaction [8]. Above which lies the cerebral cortex, which works as a relay station for receiving and sending all the sensory information. Figure 6.2 represents the schematic structure of a human brain.

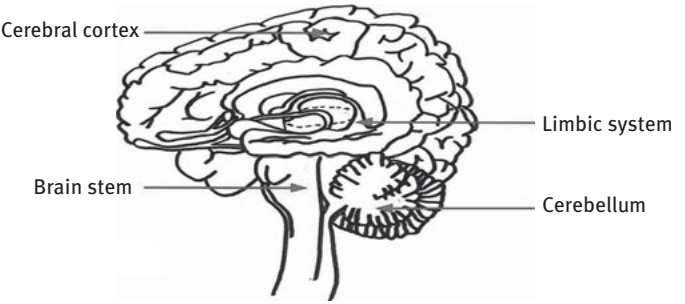

Figure 6.2: Schematic structure of the human brain.

For infants, it has been found to indicate autism spectrum disorder in the first year [9]. Furthermore, as per a study that is sponsored by the National Institute of Health (NIH), USA, the medulla oblongata (a region of the brain stem, which is known to control the breathing functions) affected in the fetus causes sudden infant death syndrome [10, 11]. The magnetic resonance imaging of the brain has been shown to indicate evidence supporting mental illnesses in adults, such as reduced gray matter in certain areas of the brain in diseases, such as dementia, depression, bipolar disorder, and schizophrenia [12–16].

The development of neurons in the brain occurs depending on the frequency of neuron firing in that region (i.e., forms a denser network of neurons, and hence, grows larger in size). For example, the amygdala (a part of the brain in the limbic system responsible for anxiety, fear, and aggression) grows larger if the person feels these emotions on a regular basis [17]. In the earlier decades, epilepsy and seizures emerged in the temporal lobe (called so because these lobes of the brain that exist around the temples), which is where our auditory cortex also exists. There is medical evidence which depicts that seizures are due to abnormal neuron activity in the brain [18]. This evidence indicates that the healthy development of the brain from birth and even in adulthood stands crucial to the quality of life we have. Maslow's pyramid of needs (as shown in Figure 6.3) describes the needs of a human being [19].

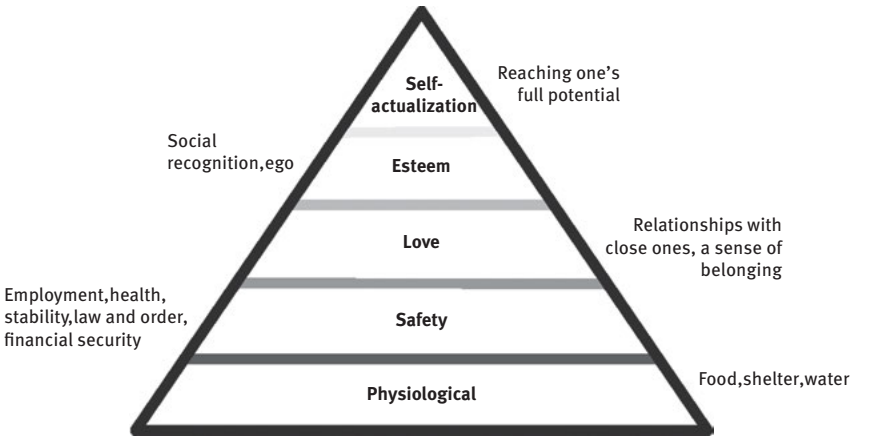

Figure 6.3: Maslow's pyramid. After [19].

The physiological needs of a person with neurological disorders remain challenged and they depend on others due to their inability to perform day-to-day activities, and needs of safety, such as employment, are always questionable due to their decreased ability to perform tasks. They are always offered jobs that are lowly paid. They also face multiple health issues in their development; hence, special care, attention, and funds are required [20]. These people are able to form good relationships with their close ones. Every neurologically disordered person may have limited or no social interaction to have any kind of social recognition, so their self-esteem stands challenged. The self-actualization of the person is challenged by all the below circumstances, often leading them to fall into psychological disorders. All these are prevented if the brain growth is normal in the infant. To ensure the same, there are multiple measures that the mother and the family take for the well-being and development of the infant.

6.3 Impact of music on infants and adults

An aspect that is to be noticed after the baby is born is how it reacts to sound. Therefore, toys are made with sounds, for example, wind chimes that make sounds when they collide together or musical tones triggered by some action. Furthermore, the mother often sings lullabies for the infant to sleep [21]. With their tiny vocabulary of crying for all their day-to-day needs, they pay close attention to and enjoy music even when they are growing. This gives us evidence that music

plays a crucial role in their early formidable years. A complete assessment of music on infants is trickier than that on adults due to their development phase.

An experience of a musician in her own delivery showed that during the preterm delivery of her twins in the 22nd week of her pregnancy, she lost one of her infants, and the development of the other had to take place in the noisy environment of the hospital; hence, she decided to bring her own CDs to the hospital and play it by her infant's station, and to her amazement the vital signs of the baby are as follows: oxygen saturation in the blood improved, heart rate improved, and blood pressure stabilized and so was observed in all the babies who were in the area and could listen to the music. This delivered the healing power of music to the infant born in the 22nd week having a survival rate of only 2% in 2009, of which very few survive and if they survive are prone to disabilities; however, with the continuous presence and exposure to music, the infant was able to grow normally [22].

In adulthood, the positive effect of music in the life of musicians has shown denser interconnections between the left and the right hemispheres of the brain, which is connected by the corpus callosum. This interconnects the two hemispheres of the brain and allows them to communicate via firing neurons [23]. White matter and gray matter distributed normally in humans have been observed to be biased toward more gray matter in musicians and dancers, which allows them to perform better at certain tasks [24]. Furthermore, a reduced amount of gray matter is indicative of the disease that affects the brain causing mental illnesses [25].

We conducted a study on 40 adult subjects to whom we played classical music that was instrumental based on the following experimental setup. For the experiment, we took a composition of each of the composer: J.S. Bach – Invention No. 1 in C Major, BWV 772, L.V. Beethoven – Rondo in C op. 51 No. 1, and W. A. Mozart Bassoon Concerto in B flat, K. 191, each comprising joyful and tense music within them, and conducted a listening test. Based on the suggestions of a general physician, we asked the subjects to sit ideal for a while before the test and we asked them not to have tea or coffee 2 h prior to the test as it affects the dopamine levels, blood pressure, heart rate, and oxygen percentage, and thus, could induce a bias (Figure 6.4). The tests have been performed in accordance with the Declaration of Helsinki [26]. We split the compositions based on these criteria, and perform the experiment in the following manner (Figure 6.5):

- Subjects were asked to fill up a metadata information form giving information about them and their preferences in music, along with their consent for giving their medical information for the purpose of this study.

- Their initial blood pressure (systolic and diastolic), heart rate, and blood oxygen percentage were measured.
- Subjects were asked to listen to joyful music for a minimum duration of 90 s before the measurements were taken again while listening to music. This observation was made for three composers separately.
- They were given a break for 2–3 min before continuing with the test.
- Then they were made to listen to tense music for a minimum duration of 90 s before the measurements were taken again while listening to music. Again separate readings were taken for the composer.
- Last observation was made after taking all the readings for 2–3 min after stopping the music.

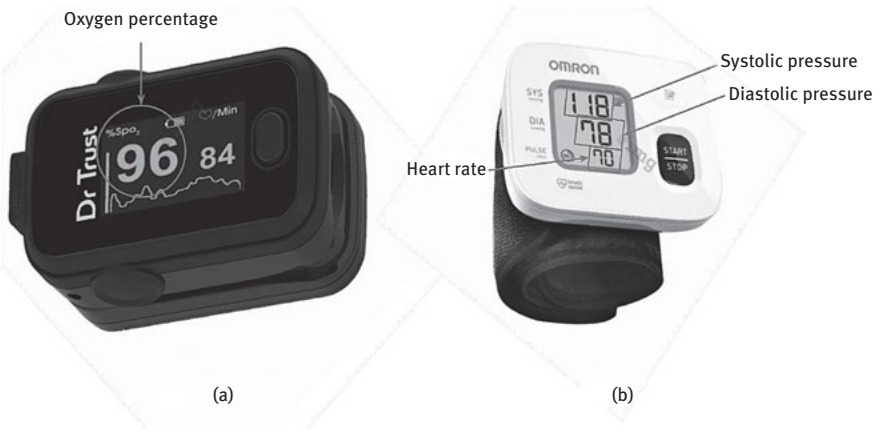

Figure 6.4: (a) Pulse oximeter: Dr. Trust (USA) Signature Series FingerTip with AUDIO VISUAL ALARM water-resistant pulse oximeter, and (b) blood pressure sensor: Omron HEM 6161 Fully Automatic for Measuring Blood Pressure.

Figure 6.5: (a) Measuring the blood pressure, and (b) taking the oxygen saturation readings of the subjects during listening tests. In each picture, left side shows subjects, and right side shows authors conducting data collection.

In Figures 6.6 and 6.7, the *Y*-axis denotes the change in the magnitude of the respective measures: the blood pressure (systolic, diastolic) and heart rate vary from their initial readings by the magnitude indicated in the *Y*-axis, and the width of the violin denotes the concentration of subjects, that is, greater width indicates greater number of subjects showing the corresponding *Y*-axis value of change from their normal readings. The oxygen saturation varies between three classes, that is, 1 for the subjects who showed an increase, 0 for the subjects who showed no change in oxygen saturation levels, and –1 for those subjects who showed a decrease in oxygen saturation levels. We plot a *violin plot* to describe these observations. The violin plot is designed in a manner where the distribution of the data is plotted, and then that distribution is

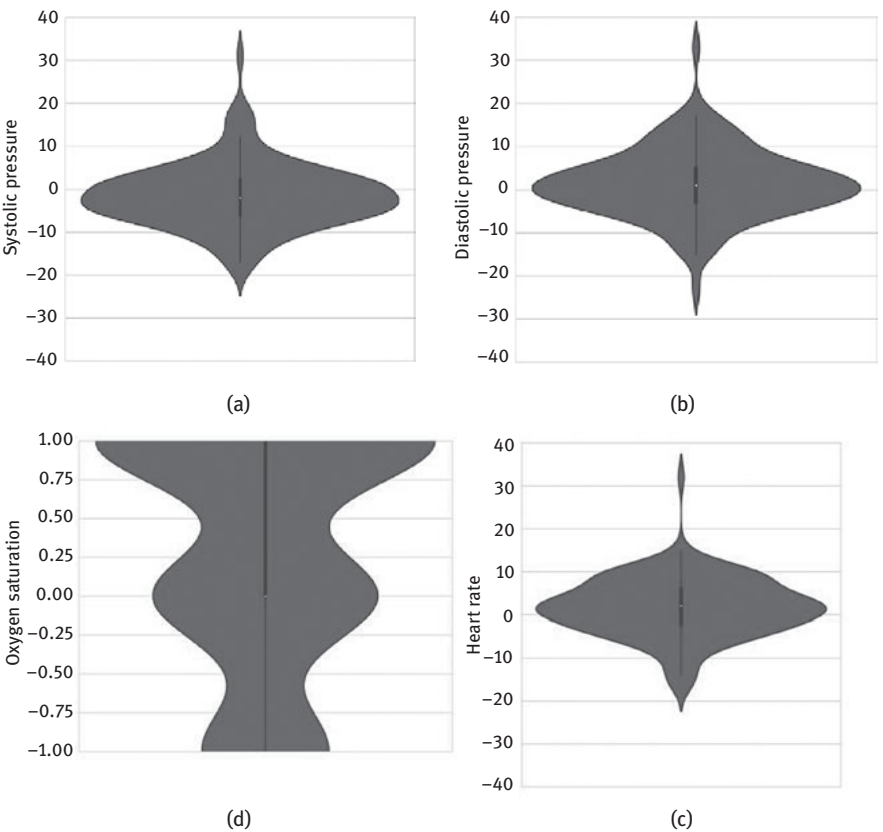

Figure 6.6: Observations for subjects during joyful music: (a) systolic blood pressure, (b) diastolic blood pressure, (c) heart rate, and (d) oxygen saturation. After [27].

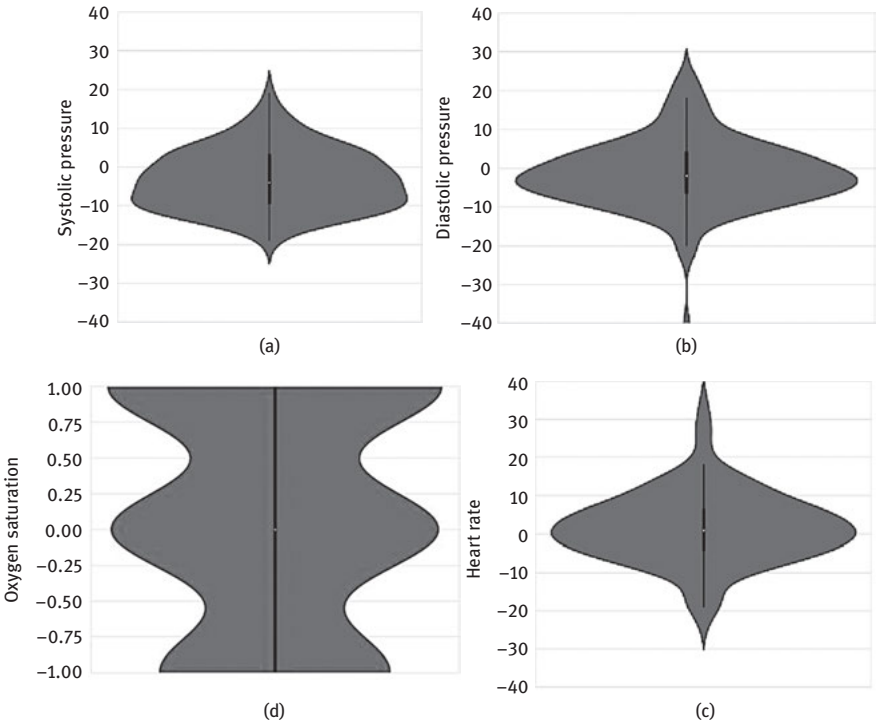

Figure 6.7: Observations for subjects during tense music: (a) systolic blood pressure, (b) diastolic blood pressure, (c) heart rate, and (d) oxygen saturation. After [27].

symmetrically flipped and plotted alongside itself to give an estimation of width. This is done so because our eyes perform a good estimation of lengths in symmetrical objects. The width of the violin plot at corresponding values allows us to observe where the data is concentrated among all the other values to form a conclusion. We obtain the following observations for the happy/joyful parts of the music:

- The systolic blood pressure decreased most for more than 40% of the subjects.
- The diastolic blood pressure showed less variation for most of the subjects; however, for the subjects who showed variation majority showed a drop in the levels of the diastolic blood pressure.
- The heart rate had an overall increase for most of the subjects when subjected to this kind of music.
- The oxygen saturation increased for most of the subjects.

We obtain the following observations for the tense parts of the music:

- The systolic blood pressure tends to decrease for most subjects and as compared to joyful music, it decreases with greater magnitude.
- The diastolic blood pressure shows a significant decrease for most subjects as compared to the joyful music, which stays mostly constant for the subjects.
- The heart rate varies equally on both sides of the subjects.
- The oxygen saturation either increases or remains constant for the majority of the subjects and for a few, it decreases.

These observations indicate a positive result as overall in both the cases, the oxygen saturation level increases and the blood pressure decreases. This may not be very crucial for a healthy person listening to music at their own leisure time but rather more crucial for an unhealthy person whose such vital signs are to be monitored daily and kept in control. Based on the studies in [25–27], it was found that preterm infants kept in the NICU were surrounded by noise from fans, pagers, and other devices. The infants were split into two groups, namely, control infants (who were not subjected to music) and music infants (who were subjected to voice lullabies), and the results shown in Figure 6.8 were obtained. Analyzing previous studies on voice lullabies played to infants on the NICU, we found that a greater average weight gain of 79 g over a 3 day time period for the preterm infants subjected to music and a 62 g weight gain of infants not subjected to music. We also observed that the total stay in the NICU of the preterm infants was reduced to 5 days when subjected to music than the preterm infants not subjected to music. Some studies conducted also reported that while playing lullabies, the oxygen saturation level increased.

In Table 6.1, * represents not significant and REE is resting energy expenditure. We observed the different behaviors of infants, Cohen's d, and the confidence interval of the test. Cohen's d is given as follows [32]:

$$d = \frac{\bar{m} - \bar{c}}{S_c} \tag{6.1}$$

where $\bar{m}$ denotes the mean of the experimental group subjected to music therapy, $\bar{c}$ denotes the mean of the control group not subjected to music, and S_c denotes the standard deviation of the control group of infants. We observe significant deviations from the control group for heart rate, behavior state, oxygen saturation, sucking/feeding ability, and length of stay. It is interesting to observe that this observation is corroborating with our results for the effects of music stimuli on adults as well, indicating the role of music on the human body from infancy to adulthood. Such evidence of music regulating

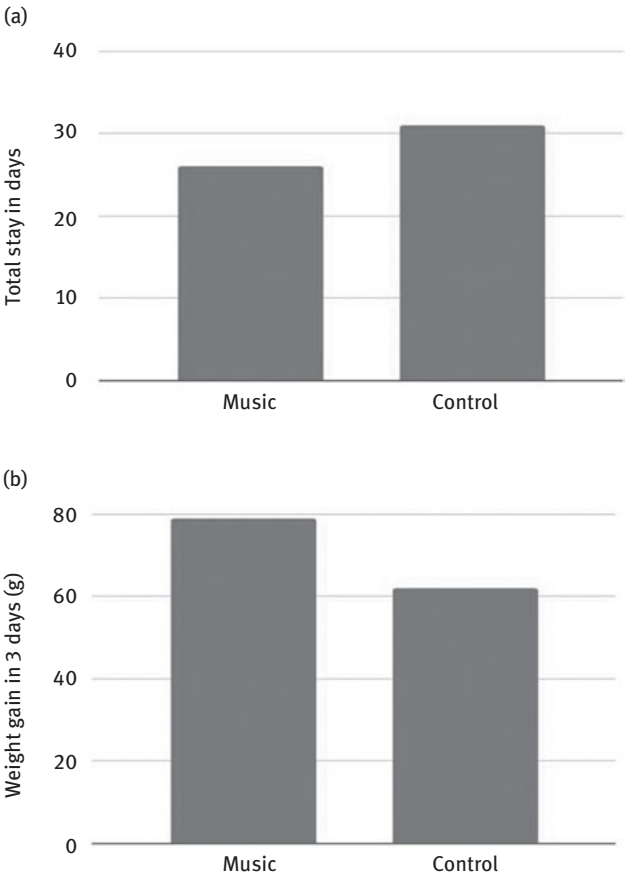

Figure 6.8: Observations on infants [31]: (a) total stay in days versus music and control infants (adapted from [28, 29]) and (b) weight gain in 3 days versus music and control infants (adapted from [29, 30]).

the autonomic nervous system to healthy behavior and response intrigues us to investigate deeply about the structures of music that cause such effects on the human body. Furthermore, subjecting the fetus in the womb via *baby pod* has shown response of the fetuses indicating neural developments in the fetus' brain, which may help track and improve its prenatal development in the mother's womb [33].

Table 6.1: Summing up the studies conducted on infant behavioral effects when subjected to music therapy (adapted from [31]).

Dependent variable	No. of studies	Cohen's *d*	*p*
Heart rate	4	1.19	0.00
Behavior state	11	1.09	0.00
Respiration rate	1	1.07	1.00*
Oxygen saturation	9	0.97	0.04
Sucking/feeding ability	4	0.85	0.00
Length of stay	7	0.71	0.00
Weight/REE	6	0.43	0.38*
Head circumference	2	0.24	0.43*
Blood pressure	1	0.13	1.00*

6.4 The occurrence of learning in the infant

One of the major abilities of the brain is learning. The infant is born with the little vocabulary of crying to call for attention, smiling to show happiness or joy, among other expressions. As the baby grows, it starts to walk by learning to crawl, attempting to stand up sometimes, falling sometimes, and then getting back up to crawl, and on one day after more than a thousand attempts, it learns how to walk. This period for an infant lasts from 4 to 12 months. Another example for it is the language acquisition in which the infant picks up its first few words in the second year, learning only small lettered words. For example, "mama" or "papa," progressing to the third year, it picks up the formation of sentences and finally in the fourth or fifth year, it marks the completion of major *language acquisition*. First, the infant has no recognition of itself but only after the first 15–24 months does it recognize itself as different from others and is conscious of its hands and legs and begins to develop emotions related to the self, for example, shame, guilt, pride, jealousy, and arrogance. This shows us that the learning process for the toddler takes longer time to complete, which consists of learning specific movements, that is, balancing on its own feet, learning to understand new words, fit words in the right organization of the sentence to express themselves, and learn to develop consciousness about themselves.

During these years, the brain develops by connecting neurons from one part to another as the infant grows. The human behavior classifies behavior into two categories, namely, innate and learned, out of which we focus on learned behavior. The iterative learning process from experience is thought of as a learning

behavior that develops after the child is born, and is not present innately. The learned behaviors include habituation, classical conditioning, operant conditioning, and insightful learning. Habituation refers to our ability to stop reacting to a stimulus because it has become common for it to occur. Classical conditioning versus operant conditioning are methods of learning, where the person learns to predict the rewarding stimulus before the actual stimulus in case of the classical conditioning and learns after getting a rewarding or punishing stimulus how to act in the environment. We can distinguish between the two as shown in Figure 6.9.

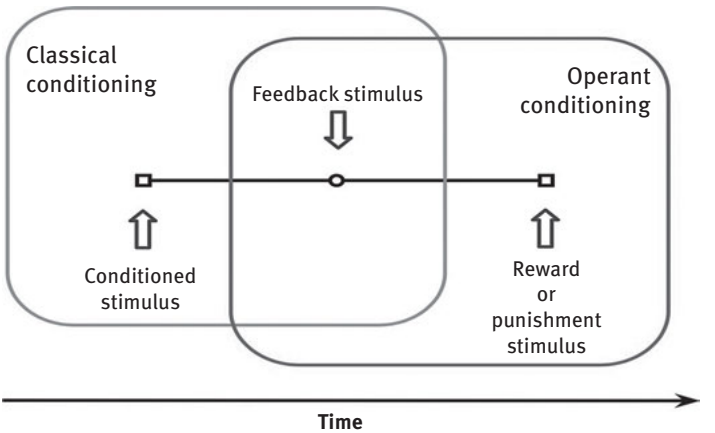

Figure 6.9: Classical conditioning versus operant conditioning.

Figure 6.9 depicts that when the stimulus is received, it allows us to distinguish between classical conditioning and operant conditioning. In classical conditioning, the stimuli before allow us to predict the conditioned response while in the operant conditioning, the stimulus after the event allows us to predict the next action. In the computational world, we have modeled reinforcement learning (RL) algorithms, which is based on the following learn to interact with the environment and reach the required goal. We call the algorithm interacting with the environment as the RL *agent*:

- Design the states space and the action space of the agent: where the state defines all possible states the agent can be in and actions define what are all the possible actions the agent can take in any state.
- Design the goals and subgoals of the RL agent.
- Construct the reward system based on the goals and subgoals.
- Design of the environment of the RL agent. This is the simulation of a world where the agent interacts and learns how to take actions based on

what state it is currently present in, and what reward it has gained previously for each action based on the state.
- Design what kind of reward updates will the agent have:
  - After each episode, where each sequence of moves has a definite end either after taking the set number of actions or after a set number of time steps elapsed or on reaching a terminal state the episode ends.
  - After each action called temporal difference learning.
  - After a set decided the number of epochs of interacting with the environment, useful for approximating the intratransition between states in a Markov decision process.

Our learning behavior occurs due to a dopamine-based learning mechanism in our brain which gives us rewards as per our actions [34]. We describe the dopamine-based rewarding experience of the infant in Figure 6.10. These studies have grown over time to incorporate nature's way of learning into these algorithms, and most of them prove to give better results than the original algorithms proposed [35]. In the studies of infants, the craving for a mother's love and attention is crucial for the behavioral development of the child. This occurs due to the release of oxytocin that triggers the release of dopamine in the child's growth. Dopamine is a neurotransmitter responsible for creating motivation in our minds to pursue actions leading to a reward. The negligence of the mother or her absence causes the child to develop the neurobehavioral disorder, such as attention-deficit hyperactivity disorder (ADHD) [36, 37]. An altered

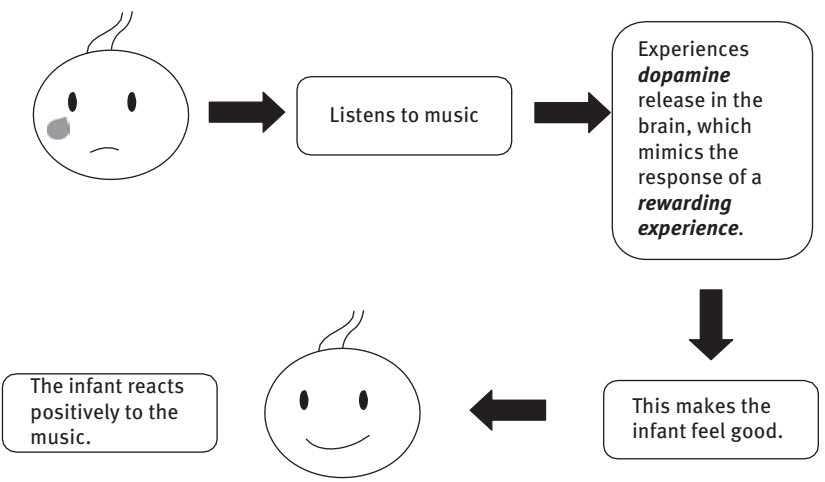

Figure 6.10: The experience of the infant while listening to music.

RL mechanism is present in ADHD [38]. This forms the motivation to administer stimulants that would trigger such behaviors. We investigated that the repetitive structure of the music forms a *melody*, taking into consideration the notes, the time durations of the notes, and the beats that create a dopamine-releasing reaction from humans [39]. Neuroplasticity, the brain's ability to change throughout an individual's life, is enhanced by music [40, 41]. We design an RL algorithm based on the Upper Confidence Bound (UCB) RL algorithm, which is based on the current note that predicts the next note. Based on whether the note is correct or not, the agent gains a reward and learns to take action.

The UCB algorithm is used to allow the RL agent to learn the q-values and, hence, approximate the probability distribution function (PDF) in hearing the next note n given the last note heard was n^l, consequently forming a complete PDF. Here N is the 128 × 128 matrix containing the number of times the notes were played in a sequence, and $R[i]$ is the reward obtained at ith position of the song:

$$q(n/n^j) = q(n/n^j) + \frac{1}{N(n/n^j)}(R[i]\ q(n/n^j)) \tag{6.2}$$

The action and states of the RL agent are 128 notes, namely, C, C#, D, D#, E, F, F#, G, G#, A, A#, B, and their harmonics. According to the current state (note), the RL agent predicts the next action (note) to get the maximum reward. Here, we provide the reward system as a simple binary one, whereupon prediction of the correct note, the RL agent receives a positive reward of value +9 and on the wrong note, the RL agent receives a negative reward of value −2. The q values are assigned a value of 0 initially, which is a nonoptimal policy. We use the UCB algorithm of RL to iterate and predict notes over the composition. We aim at an accuracy of 25–35% of the RL agent; hence, we reiterate until the accuracy is at least 25%, and that is why we limit the iterations to 500 epochs. The plot is made in intervals of 10 epochs, and the pseudocode for the same is given in Table 6.2 (as given in our recent study reported in [27]).

This algorithm allows us to approximate the learning curve of a composition based on the notes used. We extract the notes from the MIDI files and predict accordingly. We compile a list of 15 composers and take a composition of each and run this algorithm to observe the learning curves. The composers and their corresponding curves chosen are given in Table 6.3, and the observations are given in Figure 6.11 (as given in [27]). These learning curves are formed in a manner in which if the agent reaches 25% accuracy, it stops learning and if it is below 25% accuracy it continues learning up to 500 epochs on the composition.

Table 6.2: Proposed algorithm for the reinforcement learning agent (after [27]).

S.no.	Steps
1.	Q[128][128]← 0,N[128][128] ←0
2.	R = [], epoch_ quantum = 10,actions = 128,
3.	note_seq = returnNoteSeq(filename)
4.	runs = length(note_seq)
5.	bandit[0] = −2
6.	bandit[1] = 9
7.	c = 2,count = 0, total = 0
8.	for *i* in range(0, epoch_quantum):
9.	for *m* in range(0, runs −1):
10.	a = 0,max_upperbound = 0
11.	for *k* in range(0,128):
12.	if(N[k][note_seq[m-1]]¿0):
13.	upper_bound = Q[k][note seq[m−1]] + cs $\sqrt{(\log(m)) = N[k][\text{note seq}[m \ldots 1]]}$
14.	else:
15.	upper_bound = 1e400
16.	if(upper_bound ¿ max_upperbound):
17.	max_upperbound = upper_bound, a = k
18.	if(a = = note_seq[m]):
19.	R.append(bandit[1]), count + = 1
20.	else:
21.	R.append(bandit[0])
22.	N[a][note_seq[m−1]] + = 1
23.	Q[a][note seq[m−1]] = Q[a][note_seq[m−1]] + (1/N[a] [note_seq[m−1]] * (R[length(R) − 1]-Q[a][note_seq[m−1]])
24.	total + = 1
25.	if(count/total*100 ¡ 25):
26.	repeat 2–24

Grouping the observations we see a shorter learning curve for Byrd (a), Haydn (d), Scirabian (h), Bartok (i), Tchaikovsky (l), Beethoven (n), and a larger learning curve for Handle (a), Bach (c), Mozart (e), Rachmanioff (f), Hummel (g), Mandelssohn (j), Schumann (n), and Liszt (o). This indicates that as per the repetitive structure in music, and the PDF of one note to the next in a given composition, the reward accumulated changes in coherence with the accuracy. A composition with a much simpler concentrated PDF reduces the randomness in the prediction of the next note and, hence, attains a higher accuracy. Another composition with a distributed PDF results in a poorer accuracy for the agent as it has many choices of notes to pick from having close probabilities.

Table 6.3: Composition chosen for each composer (after [27]).

S.no.	Name of composer	Name of composition
(a)	G. F. Handel	Concerto for Clarinet and Strings first overture
(b)	William Byrd	The Leaves Be Green
(c)	J. S. Bach	Inventio 1
(d)	Joseph Haydn	Gypsy Rondo
(e)	W. A. Mozart	Piano Concerto No.5 in D Allegro
(f)	Sergei Rachmaninoff	Russian Rhapsody for two Pianos
(g)	J. N. Hummel	8 piano Pieces op. 37
(h)	Alexander Scriabin	Etude Op. 8 No. 12
(i)	Bela Bartok	Bagatelle
(j)	Felix Mandelssohn	Rondo & Capriccioso in E, Op. 14
(k)	Johannes Bhrams	Waltz 1
(l)	P. I. Tchaikovsky	4*th* Movement
(m)	L. V. Beethoven	Rage over lost penny
(n)	Robert Schumann	In the Evening
(o)	Franz Liszt	Hungarian Rhapsody No. 2.

This algorithm mimics the reward-based learning mechanism in our brain activated through music, and can be useful for being a perceptual-computational model for music compositions, and the positive neural development in the infant.

6.5 Summary and conclusions

The presence of sound in nature has evolved to ensure communication and safety in organisms. The development of musical sounds in an organized structure took place in order to mimic the human-singing voice, and orchestrate the production of a variety of sounds that could be combined to form a *melody*. These melodies make their presence in the early beginning of the infant life creating a soothing and safe feeling for the infant. The autonomic nervous system controlled activities get stabilized in the presence of music for the adults as well as infants. Furthermore, in the NICUs when the preterm infants were made to listen to lullabies, they showed a greater weight gain in 3 days and an earlier discharge from the NICU. The development of behavior through our motivation creating dopamine system occurs as the child learns how to walk and communicate. This mechanism is also the same which is present when we listen to music in eliciting responses. This detail allows us to model a dopamine-based RL system present

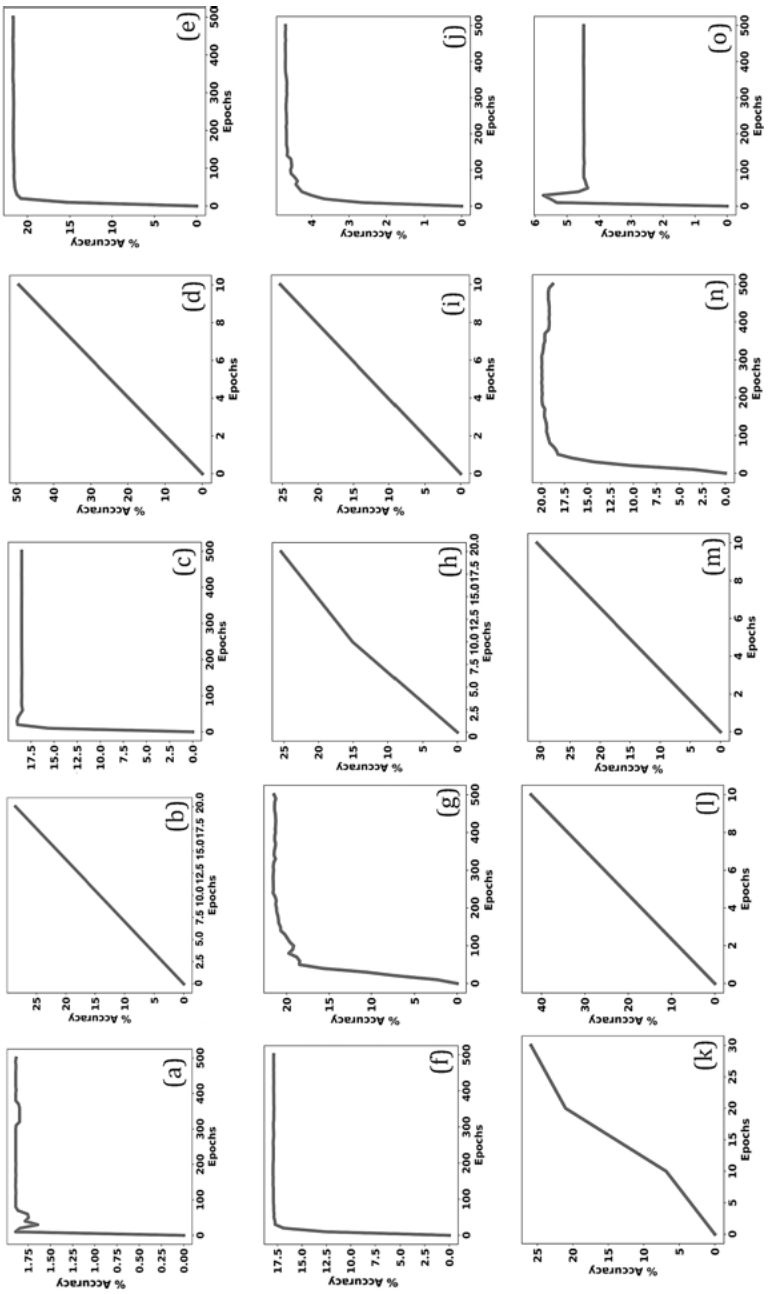

Figure 6.11: Learning accuracy versus epoch curve for compositions given in Table 6.3. After [27].

and model the question: how are we able to predict notes when a note is played? Music influencing the dopamine-based learning mechanism could help in treating neurobehavioral disorders, for example, ADHD and also improve the neuroplasticity of our brain. This study also suggests that a simpler note pattern, that is, more frequent repetitive structure of the song, can be among the best kinds of musical melodies played to the child. Music, when examining structurally bit by bit, allows us to draw our scientific facts and reasons to elicit a theory that can then be used to form theorems of science incorporating music as a part of it.

Our current research conducts a study on the music mechanisms forming a model around the structural pattern of a composition. It does not consider the emotional content in a song, which is a crucial aspect of a composition. The infant's cry is another melodic composition just as the musical composition where each cry is based upon eliciting a particular type of response from the listener and the child innately regulates between the frequencies to indicate what it is feeling [42, 11]. Just as the networks formed for music to indicate their simplicity or complexity, networks can be formed for the infant cries so as to classify which type of structures belong to which type of cry. In addition, pathologies can be detected in infant cry mechanism for earlier detection and treatment [43, 44].

Acknowledgments: The authors thank the authorities of Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, and subjects for their kind support and cooperation during this research work. The authors also thank Dr. Jayesh Shah (a general physician at Gandhinagar, India) for his valuable inputs on dopamine levels and data collection protocol during this work.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, Advances in Neural Information Processing Systems (NIPS), 2017, 5998–6008.

[2] A. Wils, "A Chick Named Albert: Meet Pete," https://www.youtube.com/watch?v=n55-nVy_PCs, 2016, {Last Accessed 12-April-2020}.

[3] Y. Nishimura, and T. Kumoi. The embryologic development of the human external auditory meatus, Acta Oto-laryngologica, 112(3), 496–503, 1992.

[4] O.D. Chorna, J.C. Slaughter, L. Wang, A.R. Stark, and N.L. Maitre. A pacifier-activated music player with mother's voice improves oral feeding in preterm infants, Pediatrics, 133(3), 462–468, 2014.

[5] T. Humphrey. The development of the human amygdala during early embryonic life, Journal of Comparative Neurology, 132, 1, 135–165, 1968.

[6] L. Thompson, P. Barraud, E. Andersson, D. Kirik, and A. Björklund. Identification of dopaminergic neurons of nigral and ventral tegmental area subtypes in grafts of fetal ventral mesencephalon based on cell morphology, protein expression, and efferent projections, Journal of Neuroscience, 25(27), 6467–6477, 2005.
[7] Y.-w. Tseng, J. Diedrichsen, J.W. Krakauer, R. Shadmehr, and A.J. Bastian. Sensory prediction errors drive cerebellum-dependent adaptation of reaching, Journal of Neurophysiology, 98(1), 54–62, 2007.
[8] P. Gloor, and A.H. Guberman. The temporal lobe and limbic system, Canadian Medical Association. Journal, 157(11), 1597, 1997.
[9] G.M. McAlonan, J. Suckling, N. Wong, V. Cheung, N. Lienenkaemper, C. Cheung, and S.E. Chua. Distinct patterns of grey matter abnormality in high-functioning autism and Asperger's syndrome, Journal of Child Psychology and Psychiatry, 49(12), 1287–1295, 2008.
[10] H.C. Kinney, P.C. Burger, F.E. Harrell, and R.P. Hudson. 'Reactive gliosis' in the medulla oblongata of victims of the sudden infant death syndrome, Pediatrics, 72(2), 181–187, 1983.
[11] H.A. Patil. Cry baby: using spectrographic analysis to assess neonatal health status from an infant's cry, Advances in Speech Recognition Mobile Environments, Call Centers and Clinics, A. Neustein (Ed.), Springer, 2010, 323–348.
[12] S.E. Nasrabady, B. Rizvi, J.E. Goldman, and A.M. Brickman. White matter changes in Alzheimer's disease: a focus on myelin and oligodendrocytes, Acta Neuropathologica Communications, 6(1), 22, 2018.
[13] J.A. Duarte, R. Massuda, P.D. Goi, M. Vianna-Sulzbach, R. Colombo, F. Kapczinski, and C.S. Gama. White matter volume is decreased in bipolar disorder at early and late stages, Trends in Psychiatry and Psychotherapy, 40(4), 277–284, 2018.
[14] S.A. Mitelman, M.S. Buchsbaum, D.S. Young, M.M. Haznedar, E. Hollander, L. Shihabuddin, E.A. Hazlett, and M.-C. Bralet. Increased white matter metabolic rates in autism spectrum disorder and schizophrenia, Brain Imaging and Behavior, 12(5), 1290–1305, 2018.
[15] A. Keymer-Gausset, A. Alonso-Solís, I. Corripio, R.B. Sauras-Quetcuti, E. Pomarol-Clotet, E.J. Canales- Rodriguez, E. Grasa-Bello, E. Álvarez, and M.J. Portella. Gray and white matter changes and their relation to illness trajectory in first episode psychosis, European Neuropsychopharmacology, 28(3), 392–400, 2018.
[16] C. Power, P.-A. Kong, T.O. Crawford, S. Wesselingh, J.D. Glass, J.C. McArthur, and B.D. Trapp. Cerebral white matter changes in acquired immunodeficiency syndrome dementia: alterations of the blood-brain barrier, Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society, 34(3), 339–350, 1993.
[17] J.F. Mustard. Experience-based brain development: scientific underpinnings of the importance of early child development in a global world, Early Child Development: From Measurement to Action. Washington DC, The World Bank, 43–86, 2007.
[18] E.H. Bertram. Temporal lobe epilepsy: where do the seizures really begin? Epilepsy & Behavior, 14(1), 32–37, 2009.
[19] A.H. Maslow. A theory of human motivation, Psychological Review, 50(4), 370, 1943.
[20] S.M. Robertson. Neurodiversity, quality of life, and autistic adults: shifting research and professional focuses onto real-life challenges, Disability Studies Quarterly, 30(1), 2009.
[21] F. Baker and E. Mackinlay. Sing, soothe and sleep: a lullaby education programme for first-time mothers, British Journal of Music Education, 23(2), 147–160, 2006.
[22] R. Spielberg, "The Healing Power of Music: Robin Spielberg at TEDxLancaster," https://www.youtube.com/watch?v=8LTusPwrH9Et=410s, 2014, {Last Accessed 12-April-2020}.

[23] I. Burunat, E. Brattico, T. Puoliväli, T. Ristaniemi, M. Sams, and P. Toiviainen. Action in perception: prominent visuo-motor functional symmetry in musicians during music listening, PloS One, 10(9), 2015.
[24] F.J. Karpati, C. Giacosa, N.E. Foster, V.B. Penhune, and K.L. Hyde. Dance and music share gray matter structural correlates, Brain Research, 1657, 62–73, 2017.
[25] P.J. Shah, K.P. Ebmeier, M.F. Glabus, and G.M. Goodwin. Cortical grey matter reductions associated with treatment-resistant chronic unipolar depression: controlled magnetic resonance imaging study, The British Journal of Psychiatry, 172(6), 527–532, 1998.
[26] W.M. Association et al. World Medical Association Declaration of Helsinki. ethical principles for medical research involving human subjects, Bulletin of the World Health Organization, 79(4), 373, 2001.
[27] K.S. Phatnani, and H.A. Patil. Modeling music structure: relevance to music cognition, Submitted to Transactions of the International Society for Music Information Retrieval (TISMIR), 2020.
[28] J.W. Cassidy, and J.M. Standley. The effect of music listening on physiological responses of premature infants in the NICU, Journal of Music Therapy, 32(4), 208–227, 1995.
[29] F.J. Schwartz, and R. Ritchie. Music listening in neonatal intensive care units, Music Therapy and Medicine: Theoretical and Clinical Applications, 13–22, 1999.
[30] J.M. Coleman, R.R. Pratt, R.A. Stoddard, D.R. Gerstmann, and H.-H. Abel. The effects of the male and female singing and speaking voices on selected physiological and behavioral measures of premature infants in the intensive care unit, International Journal of Arts Medicine, 5(2), 4–11, 1997.
[31] J. Standley. Music therapy research in the NICU: an updated meta-analysis, Neonatal Network, 31(5), 311–316, 2012.
[32] K. Kelley, and K.J. Preacher. On effect size, Psychological Methods, 17(2), 137, 2012.
[33] M.K. Philbin, "The Sound Environments and Auditory Perceptions of the Fetus and Preterm Newborn," A. Neustein, Ed.
[34] R.A. Wise. Dopamine, learning and motivation, Nature Reviews Neuroscience, 5(6), 483–494, 2004.
[35] W. Dabney, Z. Kurth-Nelson, N. Uchida, C.K. Starkweather, D. Hassabis, R. Munos, and M. Botvinick. A distributional code for value in dopamine-based reinforcement learning, Nature, 577,671–675, 2020, (2020), pp. 1–5.
[36] L. Strathearn. Maternal neglect: oxytocin, dopamine and the neurobiology of attachment, Journal of Neuroendocrinology, 23(11), 1054–1065, 2011.
[37] J. Williams, and P. Dayan. Dopamine, learning, and impulsivity: abiological account of attention- deficit/hyperactivity disorder, Journal of Child & Adolescent Psychopharmacology, 15(2), 160–179, 2005.
[38] G. Tripp, and J.R. Wickens. Research review: dopamine transfer deficit: a neurobiological theory of altered reinforcement mechanisms in ADHD, Journal of Child Psychology and Psychiatry, 49(7), 691–704, 2008.
[39] V.N. Salimpoor, M. Benovoy, K. Larcher, A. Dagher, and R.J. Zatorre. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music, Nature Neuroscience, 14(2), 257, 2011.
[40] E.L. Stegemöller. Exploring a neuroplasticity model of music therapy, Journal of Music Therapy, 51(3), 211–227, 2014.

[41] Editor, "Times of India: Music Helps Neural Plasticity," https://timesofindia.indiatimes.com/city/ahmedabad/music-helps-neural-plasticity/articleshow/74016649.cms, 2020, Last Accessed 13-April-2020.
[42] A. Chittrora, "Crying For a reason: a signal processing based approach for infant cry analysis and classification," Ph.D. Thesis, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India, June 2016.
[43] C.C. Onu, J. Lebensold, W.L. Hamilton, and D. Precup. Neural transfer learning for cry-based diagnosis of perinatal asphyxia, arXiv preprint arXiv:1906.10199, 2019, Last Accessed 13-April-2020.
[44] C.V. Bellieni, R. Sisto, D.M. Cordelli, and G. Buonocore. Cry features reflect pain intensity in term newborns: an alarm threshold, Pediatric Research, 55(1), 142–146, 2004.

Speech Technology and Text Mining in Medicine and Health Care

Patil, Neustein (Eds.), *Voice Technologies for Speech Reconstruction and Enhancement*, 2020
ISBN 978-1-5015-1041-0, e-ISBN 978-1-5015-0126-5,
e-ISBN (EPUB) 978-1-5015-0130-2

Patil, Neustein, Kulshreshtha (Eds.), *Signal and Acoustic Modeling for Speech and Communication Disorders*, 2018
ISBN 978-1-61451-759-7, e-ISBN 978-1-5015-0241-5,
e-ISBN (EPUB) 978-1-5015-0243-9

Ganchev, *Computational Bioacoustics*, 2017
ISBN 978-1-61451-729-0, e-ISBN 978-1-61451-631-6,
e-ISBN (EPUB) 978-1-61451-966-9

Beals et al., *Speech and Language Technology for Language Disorders*, 2016
ISBN 978-1-61451-758-0, e-ISBN 978-1-61451-645-3,
e-ISBN (EPUB) 978-1-61451-925-6

Neustein (Ed.), *Speech and Automata in Healthcare*, 2014
ISBN 978-1-61451-709-2, e-ISBN 978-1-61451-515-9,
e-ISBN (EPUB) 978-1-61451-9607, Set-ISBN 978-1-61451-516-6

Neustein (Ed.), *Text Mining of Web-Based Medical Content*, 2014
ISBN 978-1-61451-541-8, e-ISBN 978-1-61451-390-2,
e-ISBN (EPUB) 978-1-61451-976-8, Set-ISBN 978-1-61451-391-9