UNIVERSITY OF CALIFORNIA

Los Angeles

Speech Normalization and Data Augmentation Techniques
Based on Acoustical and Physiological Constraints and
Their Applications to Child Speech Recognition

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Gary Joseph Yeung

2021

ABSTRACT OF THE DISSERTATION

Speech Normalization and Data Augmentation Techniques
Based on Acoustical and Physiological Constraints and
Their Applications to Child Speech Recognition

by

Gary Joseph Yeung
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2021
Professor Abeer Alwan, Chair

Recently, adult automatic speech recognition (ASR) system performance has improved dramatically. In contrast, the performance of child ASR systems remains inadequate in an era where demand for child speech technology is on the rise. While adult speech data is abundant, publicly available child speech data is sparse due, in part, to privacy concerns. Hence, many child ASR systems are trained using adult speech data. However, child ASR systems perform poorly when trained on adult speech due to the acoustic mismatch that results from body size differences, especially the vocal folds and the vocal tract, as well as the high variability of child speech.

This research analyzes the acoustical properties of child speech across various ages and compares them to the acoustic properties of adult speech. Specifically, the subglottal resonances (SGRs), fundamental frequency (f_o), and formant frequencies of vowel productions are investigated. These acoustic features are shown to be capable of predicting acoustic structures across speakers. As such, we propose feature extraction methods utilizing these properties to normalize the acoustic structure across speakers and reduce the acoustic mismatch between adult and child speech. This allows child ASR systems to leverage adult data for training and suggests a framework for a universal ASR system that need not be adult or child dependent. Furthermore, we demonstrate that when child speech data is limited, these feature normalization methods are capable of producing significant improvements in child ASR for both Gaussian mixture model (GMM) and deep neural network (DNN)-based systems.

The dissertation of Gary Joseph Yeung is approved.

Xiaodong Cui

Achuta Kadambi

Gregory J. Pottie

Lieven Vandenberghe

Abeer Alwan, Committee Chair

University of California, Los Angeles

2021

To my father . . .

who somehow knows

the answer to everything.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation would not have been possible with the support of many people. First and foremost, I wish to humbly thank Prof. Abeer Alwan for being a wonderful advisor and mentor, both guiding me towards my goal of becoming a successful researcher and providing me with her generous support throughout my studies. With her guidance, I was able to develop my own mastery of the field of speech processing. Under her supervision, I was able to achieve success that was far beyond my reach.

Many thanks to the members of my Ph.D. committee. From the University of California, Los Angeles (UCLA), thanks to Prof. Gregory Pottie, Prof. Lieven Vandenberghe, and Prof. Achuta Kadambi, professors that I have gotten to know and learn from throughout my many years at UCLA. From International Business Machines (IBM), thanks to Dr. Xiaodong Cui for his mentorship throughout my studies, through many talks provided me with invaluable knowledge and insight towards my research.

I would like to thank Prof. Steven M. Lulich for his guidance and advice throughout my research on subglottal resonances. His insights provided me with the inspiration I needed to make my own discoveries. Additionally, I wish to thank Dr. Alan V. McCree, Dr. Gregory Sell, and Dr. Daniel Garcia-Romero for their advice and tutelage throughout my relationship with Johns Hopkins University researching speaker recognition and verification. Finally, many thanks to Prof. Cynthia Breazeal, Dr. Hae Won Park, Dr. Samuel Spaulding, and the Massachusetts Institute of Technology's Personal Robotics Laboratory for their assistance and dedication to the JIBO data collection and associated child speech research.

Many thanks to my former and current labmates who have put up with my shenanigans throughout my time at UCLA. My senior labmates, Dr. Kantapon Kaewtip, Dr. Soo Jin Park, and Dr. Jinxi Guo, were invaluable mentors and I am grateful to have worked with them. My junior labmates, Vijay Ravi, Alexander Johnson, Morgan Tinkler, Ruchao Fan, Jinhan Wang, and Yunzheng Zhu, for providing me with support and inspiration that was very much needed to persevere towards my graudation. Finally, a special thanks to Amber Afshan, who spent her entire graduate career with me and surprisingly has not gone crazy from dealing with my insanity. I am truly thankful for the many chats and discussions that she and I had together.

A big thank you to all the administrative staff that helped me throughout my time at UCLA, including Deeona Columbia, Ryo Arreola, Julio Vega Romero, Ylena Requena, Jose Cano, Adriana “Vanessa” Ramirez, and Lexi Columbia, for being a huge help throughout my studies. I could not imagine surviving through the administrative details of my time at UCLA without their gracious help and support.

Finally, I wish to thank my father, Dr. Wing K. Yeung, and my sister, Dr. Samantha L. Yeung, for all their encouragement and support through many difficult and frustrating times during my life at UCLA. Their willingness to help me throughout my struggles and cheer for my successes was truly the motivation I needed to persevere throughout this period of my life.

VITA

2015	B.S. (Electrical Engineering), University of California, Los Angeles (UCLA)
2015	Harry M. Showman Prize, Henry Samueli School of Engineering, UCLA
2015	Outstanding Undergraduate Award, Dept. of Electrical Engineering, UCLA
2015	First Year Graduate Fellowship, Dept. of Electrical Engineering, UCLA
2017	M.S. (Electrical Engineering), UCLA
2017	Ph.D. Preliminary Examination Fellowship, Dept. of Electrical and Computer Engineering, UCLA
2018	Teaching Assistant Excellence in Teaching Award, Dept. of Electrical and Computer Engineering, UCLA
2018–2021	Teaching Assistant, Dept. of Electrical and Computer Engineering, UCLA
2018–2020	IBM Ph.D. Fellowship, IBM, Armonk, NY, USA

PUBLICATIONS

G. Yeung, A. Afshan, K. E. Ozgun, K. Kaewtip, S. M. Lulich, and A. Alwan. “Predicting Clinical Evaluations of Children’s Speech with Limited Data Using Exemplar Word Template References.” In *Proc. of SLaTE*, pp. 161–166, 2017.

G. Yeung and A. Alwan. "On the Difficulties of Automatic Speech Recognition for Kindergarten-aged Children." In *Proc. of INTERSPEECH*, pp. 1661–1665, 2018.

G. Yeung, S. M. Lulich, J. Guo, M. S. Sommers, and A. Alwan. "Subglottal Resonances of American English Speaking Children." *The Journal of the Acoustical Society of America*, **144**(6), pp. 3437–3449, 2018.

G. Yeung, A. L. Bailey, A. Afshan, M. Q. Pérez, A. Martin, S. Spaulding, H. W. Park, A. Alwan, and C. Breazeal. "Towards the Development of Personalized Learning Companion Robots for Early Speech and Language Assessment, in *Proc. of AERA*, 2019.

G. Yeung and A. Alwan. "A Frequency Normalization Technique for Kindergarten Speech Recognition Inspiried by the Role of F0 in Vowel Perception." In *Proc. of INTERSPEECH*, pp. 6–10, 2019.

G. Yeung, A. L. Bailey, A. Afshan, M. Tinkler, M. Q. Pérez, A. Martin, A. A. Pogossian, S. Spaulding, H. W. Park, M. Muco, A. Alwan, and C. Breazeal. "A Robotic Interface for the Administration of Language, Literacy, and Speech Pathology Assessments for Children." In *Proc. of SLaTE*, pp. 41–42, 2019.

G. Yeung, R. Fan, and A. Alwan, "Funddamental Frequency Feature NOrmalization and Data Augmentation for Child Speech Recognition." In *Proc. of ICASSP*, pp. 6993–6997, 2021.

G. Yeung, R. Fan, and A. Alwan, "Fundamental Frequency Feature Warping for Frequency Normalization and Data Augmentation in Child Automatic Speech Recognition." *Speech Communication*, **135**, pp. 1–10, 2021.

CHAPTER 1

Introduction

1.1 Motivation

The need for child automatic speech recognition (ASR) has grown dramatically in recent years. A major reason for this is the increased usage of electronic devices such as home and living-room personal assistants. Often, speech is one of the only mechanisms young children have to interact with these devices due to their limited reading, writing, and typing abilities. Furthermore, improved child ASR performance can greatly benefit the development of teaching, langauge assessment, and clinical diagnostic tools [KWE91, BYP00, TSK06, YAO17, SZ15, RBS96, LNB14, SLR09] through interactive media [KLM17, SCA18, YBA19a]. Yet, while adult ASR has experienced significant improvement in recent years, child ASR continues to perform quite poorly in comparison [KLM17, GGN09, YA18].

Previous analyses of child ASR systems have revealed that current child ASR performance is inadequate for practical usage. For instance, [KLM17] examined the ASR performance for 5-year-old child speech using the Alderbaran NAO, a social robot commonly used for human robot interaction research. In that study, the child ASR system performed inadequately on even the most basic tasks. This included digit recognition, which had a word error rate of over 15%, and scripted speech recognition, which had a sentence error rate of over 88% on four commercial ASR application programming interfaces (APIs) (Google, Bing, Sphinx, Nuance). In contrast, the word error rate for adult scripted speech recognition is generally less than 5% [PCP15].

Similarly, other studies have investigated the performance of child ASR systems on various age ranges. [SPL14] examined child ASR systems using Gaussian mixture model (GMM) hidden Markov model (HMM)-based acoustic models and found that small differences in a child's age can result in dramatic performance changes. Similarly, [YA18] examined child ASR systems using both GMM-HMM-based and deep neural network (DNN)-HMM-based acoustic models and discovered that as the age range between training and testing speakers increases, child ASR performance degrades rapidly.

A significant impediment to the development of child ASR is the lack of publicly available child speech databases, especially for young child speech. This is further complicated when considering that deep learning, which require large amounts of speech data to train, is becoming the most prominent method of developing ASR systems. To compensate for this lack of data, young child ASR systems often employ speech data from other speech domains, such as older child speech or even adult speech, to supplement the training data. However, there are many differences between child and adult speech acoustics, further complicated by the fact that children's speech acoustics change as they grow [LPN97, LPN99, VK07, Smi92, KLP08, KL08], in part due to the growth of the physical components used to produce speech (e.g., mouth, neck, larynx). Pertaining to speech acoustics, these changes often include the rapid lowering of the fundamental frequency (f_o) [LPN97, LPN99, VK07], formant frequencies [LPN97, LPN99, VK07], and subglottal resonances (SGRs) [Lul10, LAM11, YLG18, GPY15], three defining acoustic features of the speech space.

1.2 Acoustic Theory of Speech Production

The classical linear time-invariant model of speech production treats speech as the output of a filter representing the vocal tract, including the tongue, teeth, mouth, and throat, and a quasi-periodic harmonic-rich input signal representing the voice source created by the vibration of the vocal folds in the larynx [RS11]. A representation of the source-filter speech model is shown in Figure 1.1 [ESW97]. Two sets of features from this model have defining properties for speech production: f_o and formants.

1.2.1 Fundamental Frequency (f_o)

Fundamental frequency (f_o) is the frequency of the voice source, defined by the periodicity of the quasi-periodic voice source signal. It is also defined as the distance between harmonics of the voice source in the frequency domain. Several past analyses have noted that the f_o values of children are markedly higher than both male and female adults, and this f_o generally decreases as children age [LPN97, LPN99, VK07]. This is shown graphically in Figure 1.2 [BN13], which shows the f_o values of several adult males, adult females, and children uttering the vowel /i/. Additionally, in terms of the source-filter speech production model, f_o is inversely related to the sampling of the vocal tract filter in the frequency domain [RS11]. This often results in a less sampled vocal tract filter for children as they have higher values of f_o than adult males or females.

1.2.2 Formant Frequencies

Formant frequencies ($F1$, $F2$, $F3$, $\ldots$) are the resonances of the vocal tract filter. In the frequency domain, they are characterized by the peaks of the vocal tract filter's transfer function. Formants are known to have defining properties for vowel identity [PB52, RS11]. For instance, the vowel /i/ is characterized by a low $F1$ and high $F2$ while the vowel /u/ is characterized by low $F1$ and $F2$. Thus, as speech changes over time, formants change dramatically across the utterance of a word or sentence. A plot of $F1$ and $F2$ values and the corresponding vowels for many speakers is shown in Figure 1.3 [PB52].

Similar to f_o, formants are known to be higher for children than for adults and decrease as children age [LPN97, LPN99, VK07, Smi92, KLP08, KL08]. The means of the first three

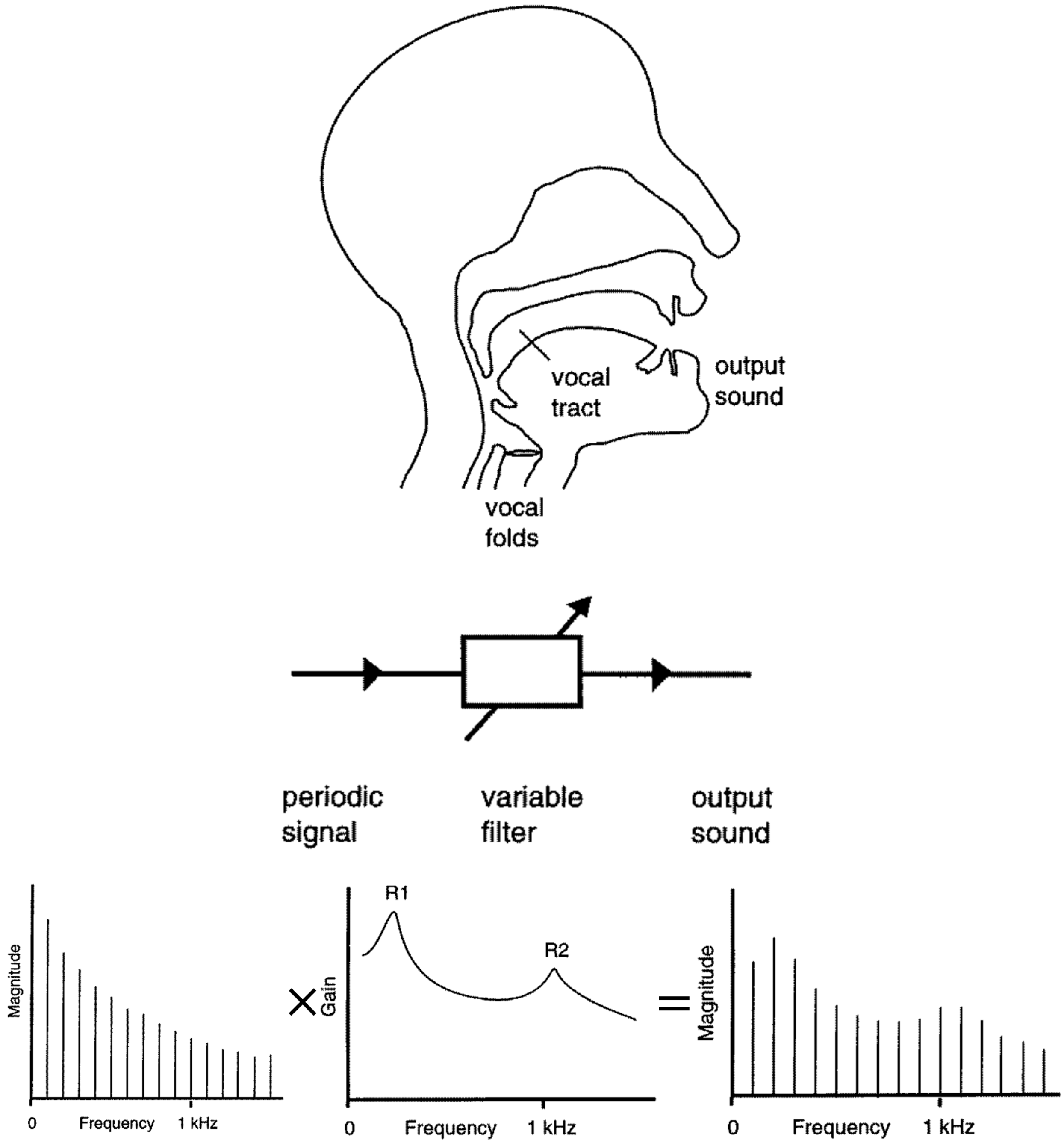


Figure 1.1: A representation of the source-filter model of speech production for voiced speech. The speech production system is shown (top), modeled as an input and filter combination (middle), and decomposed into a source signal and a transfer function (bottom). An example of the frequency content of the voice source (left), vocal tract transfer function (middle), and resulting speech (right), are shown in the decomposition. (Adapted from [ESW97].)

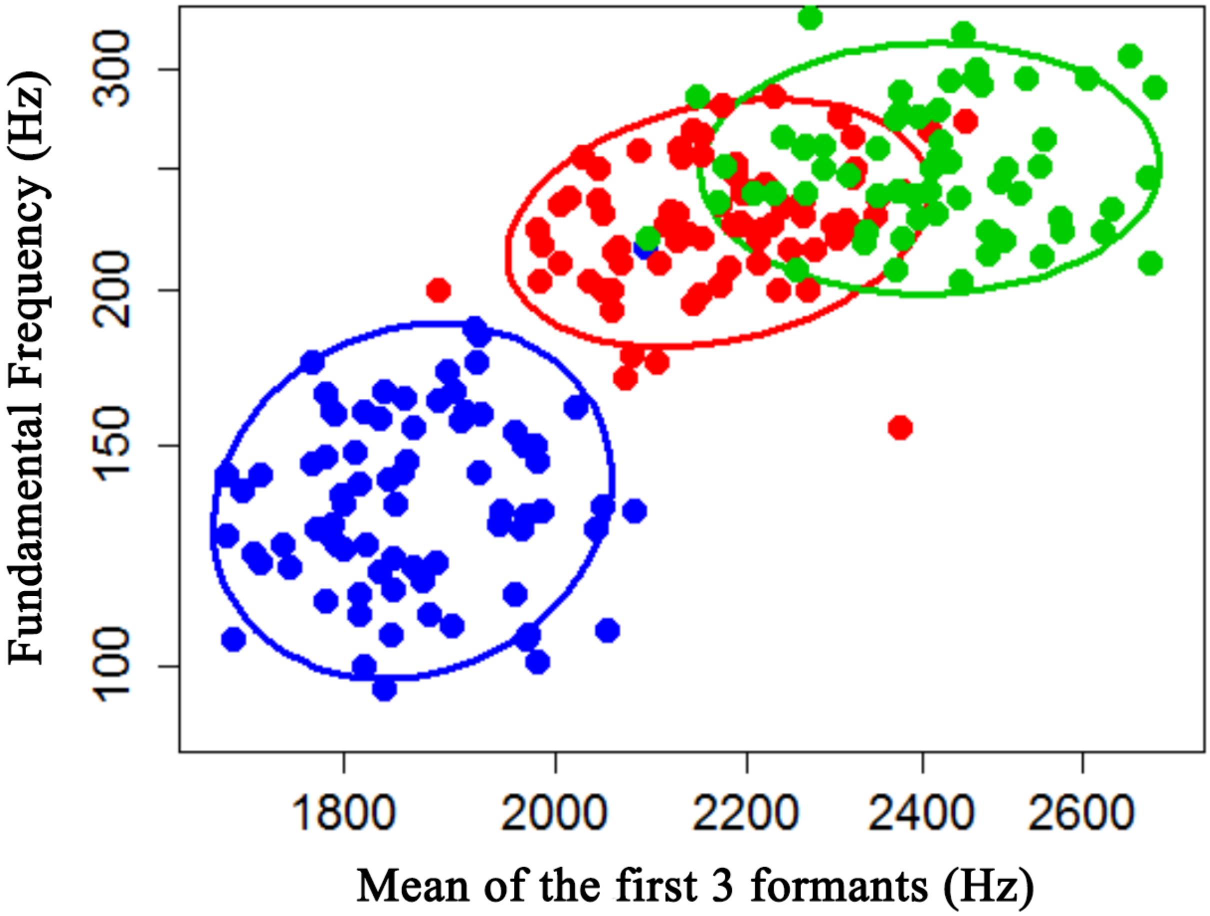


Figure 1.2: f_o vs. mean of the first three formants of utterances of the vowel /i/ using the adult male (blue), adult female (red), and child (green) data in [PB52]. Adult males have the lowest f_o and formant values while children have the highest f_o and formant values. (Adapted from [BN13].)

formants of several adults and children uttering the vowel /i/ are shown in Figure 1.2. Once again, this displays that child formants are markedly higher than adult formants. Notably, Figure 1.2 also reveals a correlation between formants and f_o.

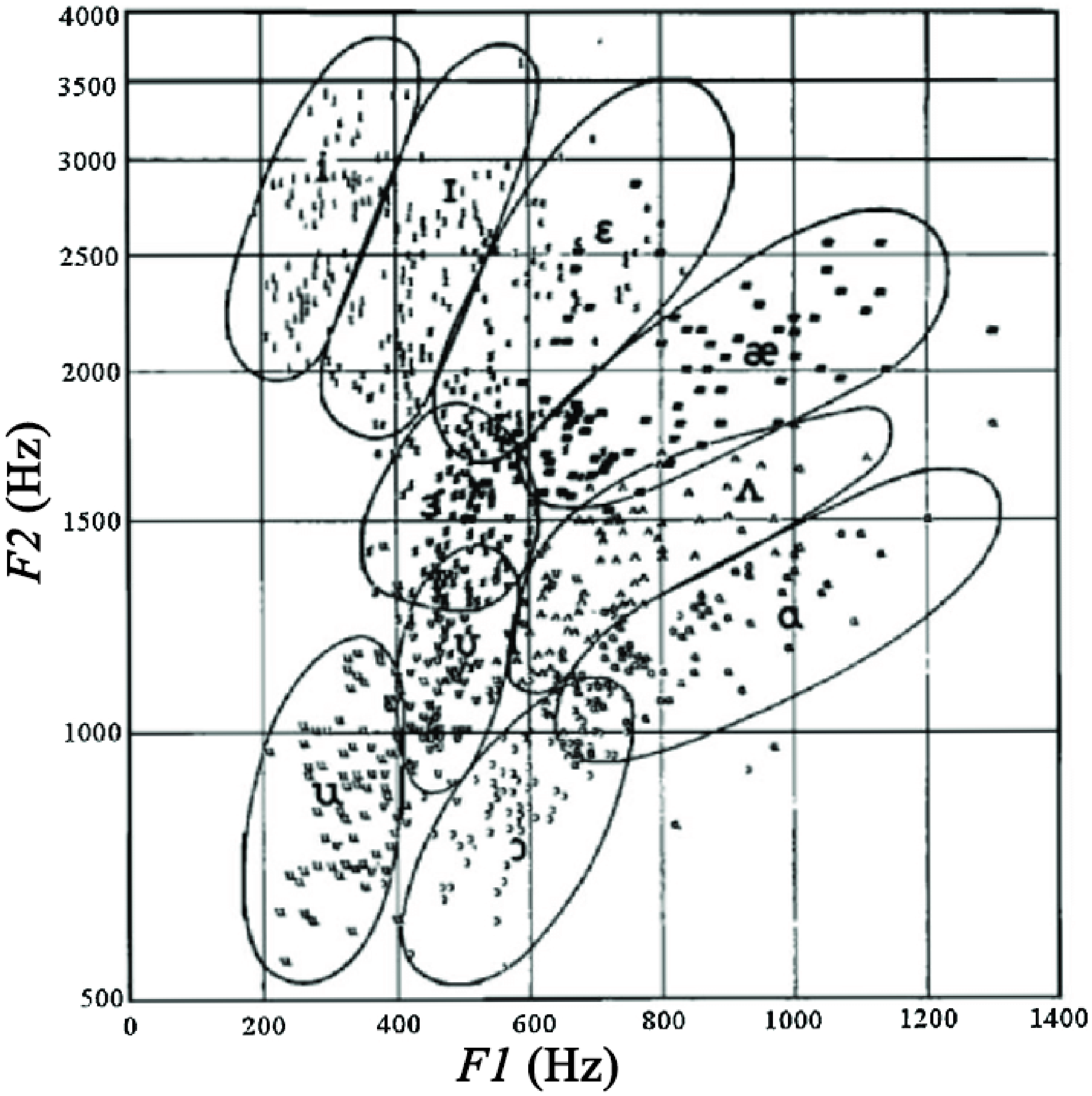


Figure 1.3: $F2$ vs. $F1$ from utterances of various vowels. Note that productions of the same vowel have similar formant volues. The general $F2$ vs. $F1$ region for several vowels are also marked. (Adapted from [PB52].)

1.3 Child Speech Databases

While there is a lack of large-scale child speech databases, some smaller-scale child speech databases are available to the public. The most common style of speech in these databases is read speech with single-word, phrase, or sentence transcripts. While databases of spontaneous styles of speech also exist, mispronunciations by children often make transcription unreliable or inconsistent, both within and across databases. Databases for English speakers include the OGI Kids' Speech Corpus (5-16 years old) [SHC00], CMU Kids Corpus (6-11 years old) [EMG97], TIDIGITS (6-15 years old) [LD93], and PF-STAR Children's Speech Corpus (4-13 years old) [BBD05]. However, unlike adult speech databases that have hundreds or thousands of hours of speech data, child speech databases generally have at most tens of hours of speech data. Additionally, due to the lack of child speech data, when training ASR systems using child speech databases, it is usually necessary to consider specific tasks or applications, such as using speech data from children in educational settings to train ASR for classroom appropriate speech technology.

1.4 Subglottal System and Resonances

The subglottal system, consisting of the trachea, bronchi, lungs, and surrounding tissues, serves as the main source of airflow that powers the larynx, and thus vocal tract, during speech production. The subglottal resonances (SGRs) are the natural frequencies of the subglottal system. These SGRs, also referred to as subglottal formants, are analogous to the formant frequencies of the vocal tract [FIL72, CB87]. However, unlike formants, SGRs generally remain stable during speech production due to the limited physical changing of the tracheobronchial tree. Past studies have analyzed the subglottal resonances and their impact on the speech waveform [ALL13, CS07, CB87, IMK76, KK90, Lul10, LAM11, LAA11, LMA12, ZNB06].

During speech production, the subglottal system is acoustically coupled with the vocal tract through the larynx [Fan60, Lul13, LZM09, Ste98, Tit08, Tit06, ZMH11]. Due to this coupling, the SGRs often manifest themselves as zeros in the vocal tract filter's transfer function. Previous studies have demonstrated that the formants manifest themselves with respect to these acoustic zeros as boundaries for the production of different vowel phonemes in adults [CS07, CBG09, DLM11, GLC11, Jun09, Lul10, LBM07, MLW08, Ste98]. Furthermore, SGRs are correlated with a speaker's height and formants [GPY15, GYM16, WLA09b, WLA09a, ALL13, LAM11, LAA11, LMA12]. As such, similar to both f_o and formants, SGRs also generally decrease as children grow [YLG18, LAM11].

As the measurement of SGRs requires considerable effort through the recording of accelerometer signals or other more invasive techniques, several studies have attempted to model SGRs using speech parameters that are easily measured. [WLA09a] used the fact that the second SGR contributed a zero in the microphone signal within the range of the second formant. This was used to estimate the location of the second SGR for adults by identifying where a discontinuity occurred in the second formant's trajectory. [ALL13] used the differences between the first three formants (in Bark scale) to estimate the first three SGRs for adults. Similarly, both [LAM11] and [GPY15] used the differences between the first three formants to estimate the first three SGRs for children. However, the modeling, analysis, and applications of SGRs have still not yet been thoroughly explored.

1.5 Automatic Speech Recognition

Automatic speech recognition (ASR) is generally composed of four modules: a front-end feature extraction module that converts a raw speech signal into a set of low-dimensional speech representations, an acoustic model that attempts to classify (or assign probabilities to) the potential phonemes (or sounds) produced by using the features, a language model that attempts to convert the series of phoneme probabilities into meaningful words, phrases, or sentences, and a decoder that interprets the scores by the previous modules and outputs the resulting transcript. This dissertation will mostly focus on the front-end feature extraction techniques and acoustic models.

1.5.1 Mel Frequency Cepstral Coefficients (MFCCs)

The Mel frequency cepstral coefficients (MFCCs) are one of the most common features used in ASR. The main characteristic of MFCCs is the use of the Mel scale, a perceptual frequency scale, to filter the frequency content of a speech frame, which in turn smooths the frequency domain envelope. The relationship between Mel and Hz is shown in Figure 1.4 and given by Eq. 1.1 as follows:

$$F_{Mel} = 1127 \ln(1 + \frac{F_{Hz}}{700}) \tag{1.1}$$

where F_{Hz} is the frequency in Hz and F_{Mel} is the pitch in Mels [OS87]. Notably, the Mel and Hz scales are approximately linearly related at low frequencies but develop into an exponential relationship as the frequency increases. An example of a 22 filter Mel filter bank using triangular filters is shown in Figure 1.5 [RS11]. Note that the filters at the low frequencies are narrower than the filters at the high frequencies due to the Mel scale.

MFCCs are frequency-based features and require the discrete Fourier transform (DFT) to be computed. As such, variations of the DFT can be used to compute different sets of MFCCs. These DFT variations are relevant as frequencies are often scaled to compensate for acoustic mismatches. This dissertation will use MFCCs and their variants as the ASR feature of choice. See [Mer76] for more details about MFCCs.

1.5.2 Acoustic Modeling

The two most common ASR acoustic models are the Gaussian mixture model (GMM) and the deep neural network (DNN) and its variations, including feedforward networks, recurrent neural networks (RNN), long-short term memory (LSTM) networks, and bidirectional long-short term memory (BLSTM) networks. When using GMMs, the acoustic model uses a linear combination of Gaussian distributions to model phoneme probabilities over some set of features. As this combination of Guassians is modeled in the input feature space, the GMM can be interpreted to model the distribution of phonemes over the feature space. This

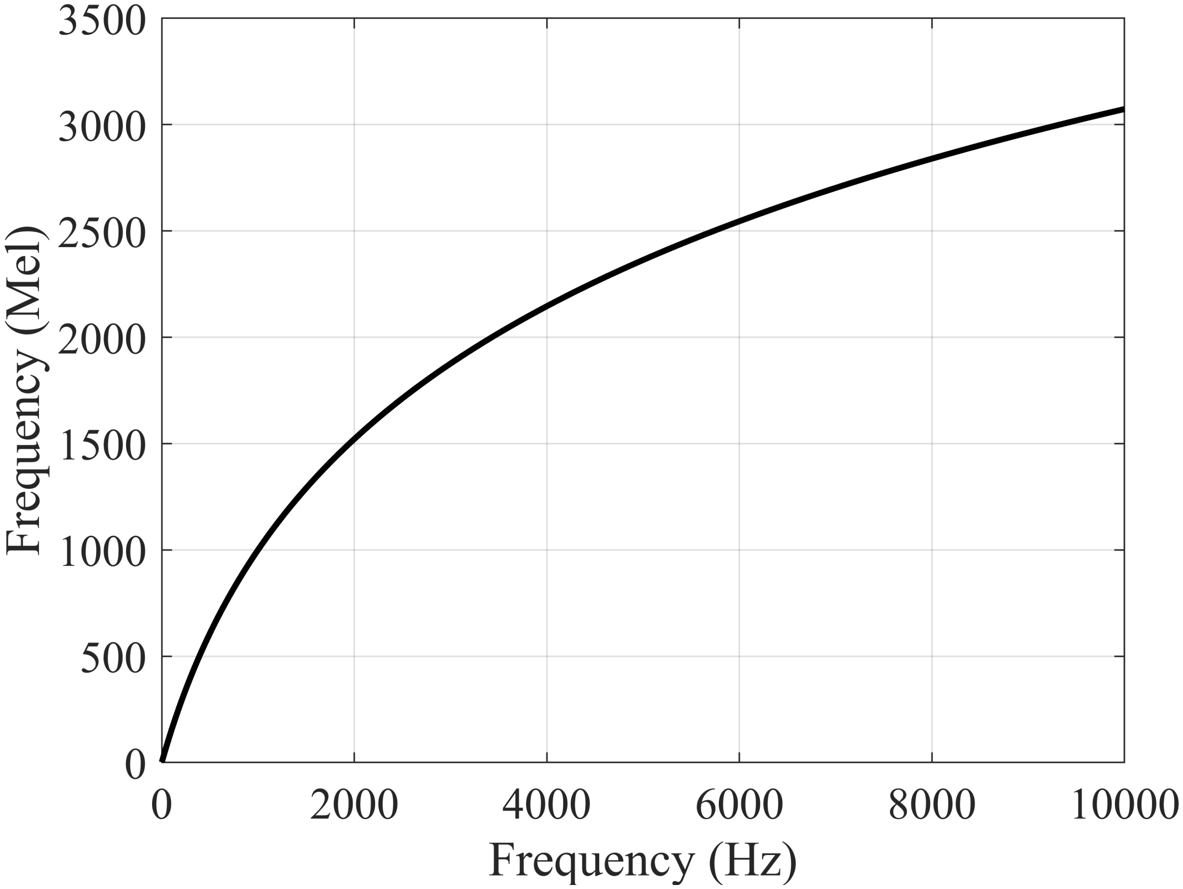


Figure 1.4: Mels vs. Hz. Note that Mels and Hz are approximately linearly related at lower frequencies but develop an exponential relationship as frequencies increase.

dissertation will focus on feedforward DNNs and BLSTMs as acoustic models but will also use GMMs for comparison purposes. See [YD15] for further information about GMM and DNN acoustic modeling in ASR.

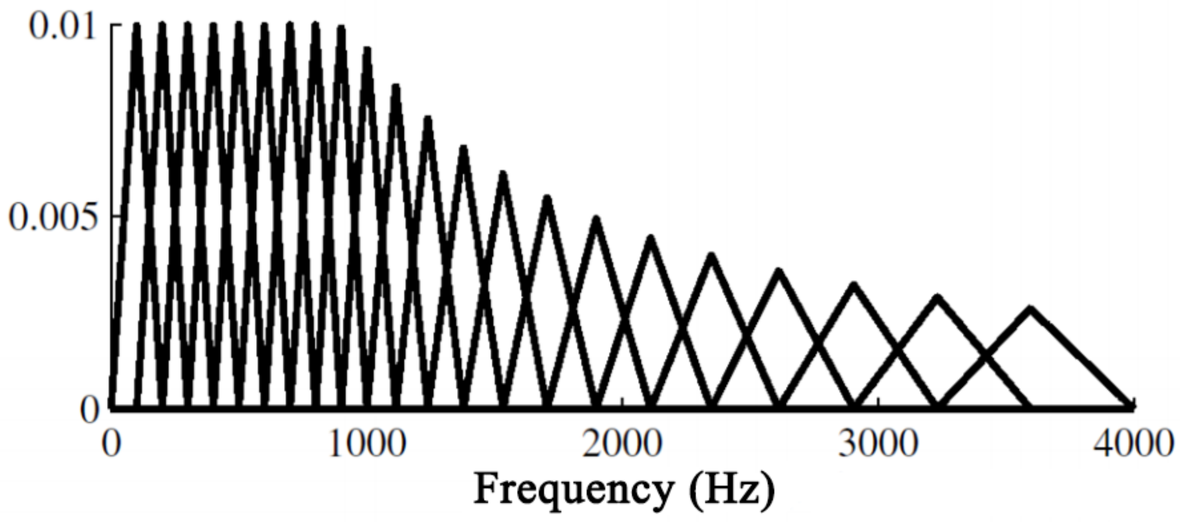


Figure 1.5: A Mel-scaled triangular filter bank of 22 filters over a bandwidth of 4 kHz. Note that the filters are narrow in the low frequencies but wider in the high frequencies. (Adapted from [RS11].)

1.6 Frequency Warping Techniques for ASR

1.6.1 Frequency Normalization

Frequency normalization techniques attempt to warp the speech spectra to a normalized speech space, reducing inter-speaker variability in the spectral domain. For instance, vocal tract length normalization (VTLN) [LR98] uses a maximum likelihood approach to warping the speech spectra, and various implementations have been successful in child ASR [SPL14, SHS03, CA05, SG14, PA06, GWL14]. Alternatively, acoustically relevant speech parameters, such as the subglottal resonances [GPY15] or third formant frequency (F3) [CA06], can be used as a normalization factor by warping the spectra to match a default speaker.

Fundamental frequency (f_o) has also been used successfully as a feature to improve adult ASR performance, even in atonal languages [FNS01, Ljo02, MSB03]. In [FG05], the authors found that f_o could be used to predict the VTLN warping factor of an utterance with a maximum likelihood approach, while in [SDS16], f_o was used to determine lifter sizes when extracting cepstral features. Although many studies examining the use of f_o in ASR were performed on adult speech, f_o may also be relevant to child ASR.

Research on human speech perception may provide further insight into the use of f_o in ASR. The tonotopic distances between formants, the distance between adjacent formants in some perceptual scale, along with the tonotopic distance between the first formant and f_o, are a set of features that have been successfully used to model human vowel perception [CL79, Chi85, SG86, Tra81]. This set of features can be interpreted as a normalization of formant-based vowel models. The inclusion of f_o in the tonotopic distance model suggests that f_o contains information that can be exploited to normalize speech spectra. This is supported by studies that suggest that the perception of vowel quality, vowel production, and voice naturalness are dependent on f_o when formants are fixed [BN12, BN13, AN07]. Furthermore, f_o and the tonotopic distances may also be useful for data augmentation.

1.6.2 Data Augmentation

Recently, ASR systems based on deep learning have used data augmentation techniques to increase the available training data to train large neural networks such as BLSTM networks. There are several ways to implement these deep learning techniques such as feature warping [JH13, CGK14], adding noise [KPP17, HCC14], and masking in time or frequency [PCZ19]. An analogue to VTLN, vocal tract length perturbation (VTLP) uses VTLN warping factors to extract features from the same utterance several times, creating additional variability in the available training data [JH13]. Data augmentation has not yet been fully explored for child speech, although some researchers have evaluated techniques that include adding noise and reverberation [WGP19] and applying out-of-domain adult data to the training data [FBL16]. Notably, while data augmentation techniques increase the amount of available training data, many techniques simply create variability without considering whether these additional features adhere to the acoustic properties of speech.

1.7 Dissertation Overview

This dissertation makes several contributions to the fields of child speech and child ASR. Two databases that were published as part of this dissertation are described. These databases provide child speech data, as well as child subglottal signal data, for eventual use in education, child speech science, child ASR, and biometric research. Additionally, a model of SGRs is proposed, which provides a more effective method for estimating child SGRs. Finally, an f_o-based frequency warping method is proposed for both frequency normalization and data augmentation in child ASR. These methods are evaluated with several child ASR experiments; the proposed techniques perform better than state-of-the-art techniques.

The remainder of the dissertation is organized as follows. Chapter 2 describes the databases and speech software published or used throughout the dissertation. Chapter 3 discusses subglottal resonance modeling and introduces a more effective way of modeling the third SGR. Chapter 4 examines the properties of f_o and proposes a frequency warping technique using these properties. Chapter 5 describes several child ASR experiments evaluating the performance of the SGR normalization and f_o warping techniques for both frequency normalization and data augmentation against other commonly used techniques. Finally, Chapter 6 concludes the dissertation with a brief summary, a discussion of potential applications of this work, and directions for future work.

CHAPTER 2

Databases and Speech Software

Several databases and speech software tools were used throughout this dissertation and are described in this chapter.

2.1 The Child Subglottal Resonances Database

The Child Subglottal Resonances Database [YLG18] is a corpus intended for use in speech science and automatic speech recognition technology. This database contains the speech utterances of 43 native speakers of American English (31 male, 12 female), aged 6-18 years old. These children were recruited through the Washington University in St. Louis psychology department subject pool, as well as through advertisements posted in public spaces around the greater St. Louis, MO area. The parents of the recruited children were asked if their children had any history of speech or hearing disorders, and none were reported. Each speaker's standing height, age, and gender were also documented.

The corpus consists of recordings that simultaneously capture the speech and subglottal acoustics of the participants. To capture the speech acoustics, a free-standing SHURE PG27 microphone (Shure, Niles, IL, USA) was used. The microphone was placed approximately 20 cm in front of the speaker and slightly to the side to avoid distortion due to high airflow sounds (e.g., the plosive /p/). To capture the subglottal acoustics, a K&K Sound HotSpot accelerometer (K&K Sound Systems, Coos Bay, OR, USA) was used. The participants were instructed to press and hold the accelerometer firmly against the skin at the cricoid cartilage

Table 2.1: The complete list of CVCs recorded for The Child Subglottal Resonances Database. Various vowels (including the approximant /ɹ/ were recorded in up to four different consonant contexts. Phonological feature specifications are also given for the features [low] and [back] [Ste98].

hVd	i	ɪ	e	ɛ	æ	ɑ	ʌ	o	ʊ	u	ɑɪ	ɑʊ	ɔɪ	ɹ
bVb	i	-	-	ɛ	-	ɑ	-	-	-	u	ɑɪ	ɑʊ	ɔɪ	-
dVb	i	-	-	ɛ	-	ɑ	-	-	-	u	ɑɪ	ɑʊ	ɔɪ	-
gVb	i	-	-	ɛ	-	ɑ	-	-	-	u	ɑɪ	ɑʊ	ɔɪ	-
[low]	−	−	−	+	+	+	+	−	−	−				−
[back]	−	−	−	−	−	+	+	+	+	+				+

just below the glottis. This placement helped prevent the formant frequencies from interfering with the accelerometer signal's frequency response, which is common when formant frequencies and subglottal resonances (SGRs) are near each other [CS07]. The participants sat inside a double-walled sound attenuating booth while being recorded. Simultaneous recording was performed using a two-channel M-Audio MobilePre USB pre-amplifier (M-Audio, Cumberland, RI, USA) connected to a computer running Windows Vista (Microsoft, Redmond, WA, USA). All signals were recorded using MATLAB (MathWorks, Natick, MA, USA). Both microphone and accelerometer recordings were sampled at a sampling rate of 48 kHz and quantized at 16 bits per sample.

A number of consonant-vowel-consonant (CVC) utterances were embedded in the carrier phrase "I said a CVC again." These phrases were displayed using a computer monitor placed directly in front of the speaker to be read aloud. However, in the case of young children who have not yet learned to read, a researcher sat inside the sound booth with the child and read the sentences aloud to be repeated. The complete list of CVC utterances recorded (bVb, dVb, gVb, hVd) is listed in Table 2.1. Each CVC utterance, along with the associated carrier phrase, was repeated at least six times, and both microphone and accelerometer signals were recorded simultaneously.

For illustration, spectrograms from both the speech and subglottal signal from a 13-year-

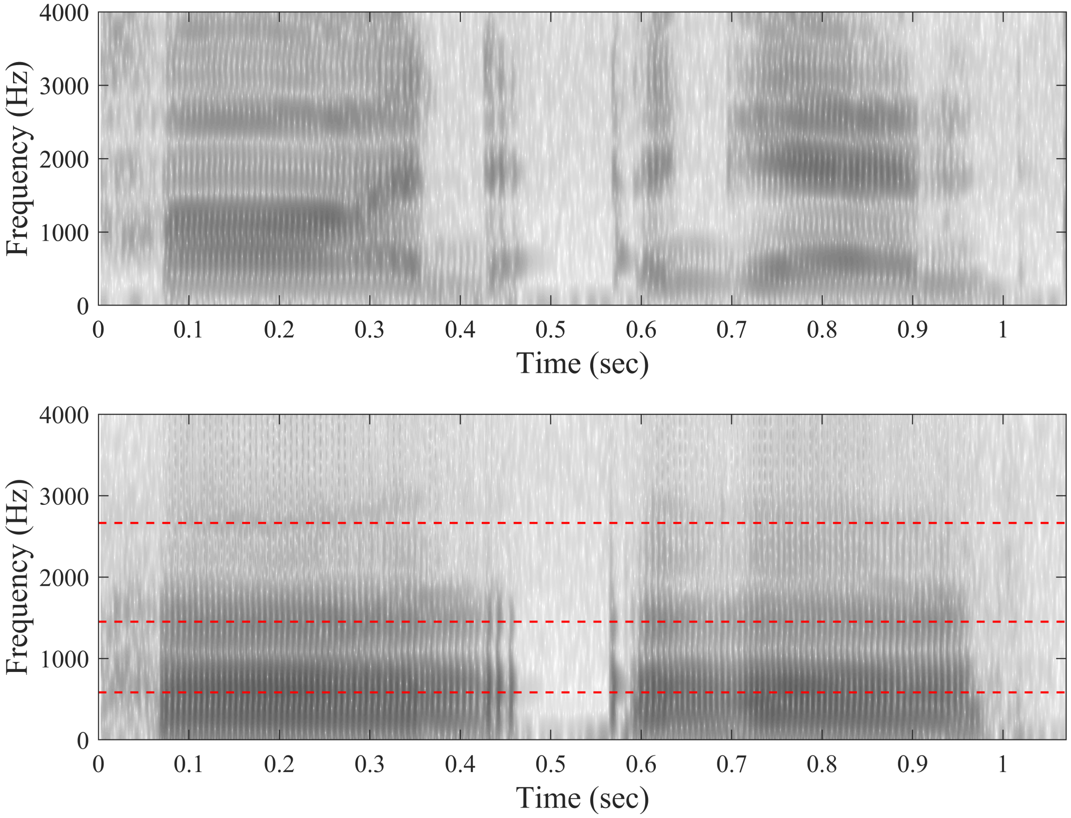


Figure 2.1: A wideband spectrogram of the speech (top) and subglottal accelerometer (bottom) signal of a 13-year-old male saying "hod again." A 2048 length discrete Fourier transform (DFT), window length of 6 ms, and frame shift of 1 ms were used. The dashed lines in the subglottal accelerometer spectrogram show the means of the speaker's first three SGRs. Notably, the subglottal acoustics remain relatively stationary while the speech acoustics change rapidly.

old male saying the phrase "hod again" are shown in Figure 2.1. Unlike the speech signal, the frequency content of the subglottal signal remains relatively stationary across the phonation, which is a characteristic of the subglottal system acoustics.

As the accelerometer signal quality for the various CVC words was variable, additional subglottal accelerometer recordings of each participant sustaining the vowel /ɑ/ were recorded with an emphasis on high-quality recordings. The signal quality was optimized by allowing the subject and researcher to interact while visually inspecting the spectrogram produced by the accelerometer signal. The position of the accelerometer, loudness of the speaker, and fundamental frequency of the speaker were all adjusted until the highest quality signal was achieved. This procedure of recording the vowel /ɑ/ was repeated twice for each subject.

Of the 43 child speakers from the initial data collection, the speech data from 28 of the speakers have been released to the public as The Child Subglottal Resonances Database. The remaining 15 children were not released publicly due to privacy issues, parental consent issues, or signal quality issues. The data will be available through the Linguistic Data Consortium (LDC) website (https://www.ldc.upenn.edu). This dissertation will use both the microphone and accelerometer signals of the hVd utterances to measure and analyze formants and SGRs in various vowel productions. A further analysis of the database was conducted in [YLG18].

2.2 The GFTA-JIBO Kids Corpus

The GFTA-JIBO Kids Corpus [YBA19a, YBA19b] is a database intended for use in automatic speech technology, human-robot interaction (HRI), education, and clinical research. At the time of writing, the database contains a total of 80 speakers, from pre-kindergarten to 1st grade, including both native and non-native speakers of American English. These children were recruited through the UCLA Laboratory School research partnership, part of UCLA's School of Education and Information Studies. The parents of the participants were asked to fill out a questionnaire containing the child's first language and reading habits. Additionally, each speaker's standing height, age, and gender were documented.

The corpus consists of the speech of children interacting with the social robot JIBO (Jibo Inc., https://jibo.com). A JIBO robot is shown in Figure 2.2 for reference. All images that were shown during the interaction were shown on JIBO's screen. Instructions, prompts, and friendly interactions administered by JIBO were recorded by a female researcher and were pitch-shifted to sound like a young child's voice.

During each recording session, a student interacted with JIBO alongside two researchers.

One researcher sat next to the child as an instructor assisting the child, while the other researcher operated JIBO. JIBO was placed directly in front of the child on a desk or table approximately 1.5 feet away from where the child was sitting. A Logitech C390e webcam (Logitech, Lausanne, Switzerland) was used as a microphone and placed at a 45-degree angle to the direction the child was facing approximately 1-2 feet away. This placement prevented distortion due to high airflow sounds and ensured that the webcam did not interfere with JIBO's movements. Audio files were sampled at 48 kHz.

The educational tasks that JIBO was programmed to administer included the 3rd Goldman Fristoe Test of Articulation (GFTA-3) Sounds in Sentences (SIS) and Sounds in Words (SIW), letter and number naming, and explanatory discourse. The GFTA-3 SIS task was suitable for children from ages 2-7 years old. JIBO narrated a story about children walking home. Five pictures corresponding to the story were shown in chronological order, one at a time, with each picture having 3-5 associated sentences. After the first telling of the story, JIBO told the story again and instructed the student to repeat each sentence.

For the GFTA-3 SIW task, children were prompted to say 58 different words by showing a picture on JIBO's screen and asking the child various questions about the picture (e.g.,

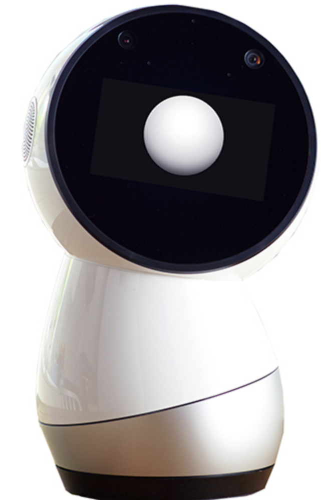


Figure 2.2: A JIBO robot.

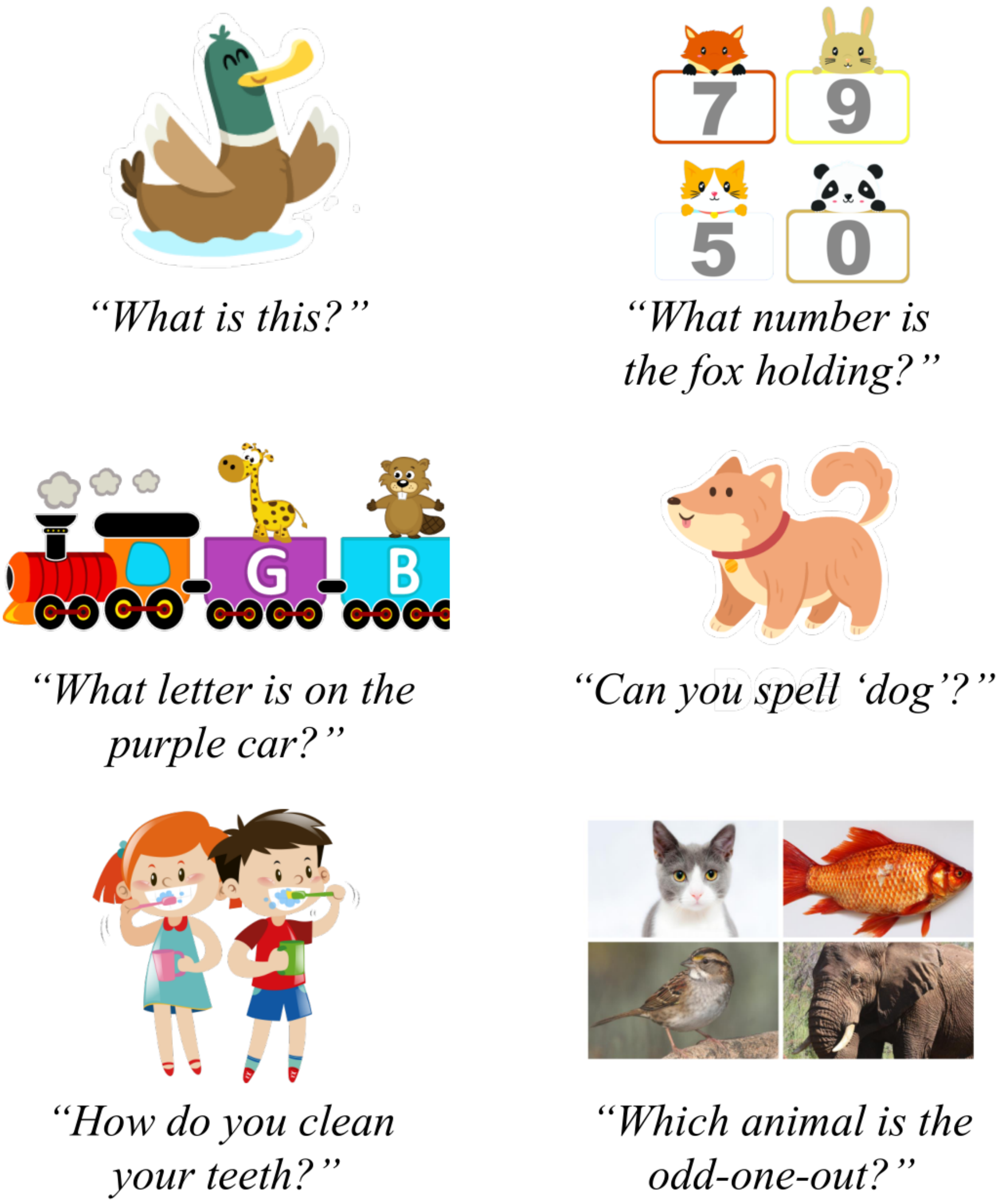
“What is this?”, “What is he wearing?”). Some pictures corresponded to multiple words, requiring JIBO to ask several questions about the picture. If the child responded incorrectly, JIBO would tell the child the solution as well as a secondary prompt for the word.

For the letter and number naming task, a random sequence of letters or numbers was randomly generated. JIBO would show a picture of a letter or number to the child and prompt the child to identify the correct word, often by asking “What is this letter/number?” Due to the simplicity of this task, there was no secondary prompt. Instead, the instructor would help the child identify the letters or numbers.

For children in 1st grade and older, JIBO administered more complicated letter and number naming games. For letters, the child’s ability to spell simple words was assessed by showing a picture of an item and asking the child to spell that word (e.g., a picture of a hat along with the question “How do you spell hat?”). For numbers, the child’s ability to apply numbers and math was assessed. This included basic arithmetic (e.g., “What is two plus five?”) and the use of numbers in real scenarios (e.g., a picture of a birthday cake with five candles along with the question “How old is the boy?”).

For the explanatory discourse task, JIBO conversed with the child about open-ended reasoning tasks or their daily routines. These conversations were accompanied by pictures that JIBO showed on its screen. For instance, JIBO could say, “Here is a picture of a boy brushing his teeth. Tell me how you clean your teeth.” JIBO could then follow up with, “Why do you clean your teeth?” or “Could you explain to me how you do that?” In another discourse conversation, JIBO would show four animals (bird, cat, elephant, fish) and ask the child, “Which animal is the odd-one-out” and “Why do you think that?” Several of the various prompts administered by JIBO are shown in Figure 2.3.

Additional procedures for the child-robot interaction were also considered and recorded throughout the database collection. When JIBO greeted the child at the beginning of the recording session, the child generally became more open to the robot interaction and less tense. The greeting that was used for this database included questions in which JIBO showed



"What is this?"

"What number is the fox holding?"

"What letter is on the purple car?"

"Can you spell 'dog'?"

"How do you clean your teeth?"

"Which animal is the odd-one-out?"

Figure 2.3: Image prompts shown on JIBO's face screen. JIBO displays the image to the child while asking the corresponding question in a friendly child-like voice.

interest in getting to know the child (e.g., “What is your favorite color?”). Additionally, occasional praise or encouragement (e.g., “Nice job”, “Good try”) seemed to provide the child with some motivation to continue playing with JIBO.

The participants of the data collection attended a university demonstration elementary school in California. Social robots were introduced to teachers and students as part of an early science and technology inquiry-based curriculum. During the first year of data collection, approximately 40% of the students were enrolled in Spanish-English dual language immersion classrooms. Additionally, approximately one-third of the participants were biracial. Parents of the children were asked to complete a survey about their child’s languages spoken, reading habits, and familiarity with technology. Approximately 90% of the parents responded that their household primarily speaks English. Over 70% responded that their child has had some exposure to computers, smartphones, or tablet devices. Further information about the data collection procedure can be found in [YBA19a, YBA19b].

2.3 Publicly Available Databases Used

2.3.1 The Subglottal Resonances Database

The Subglottal Resonances Database [LMA12] is the partner corpus to The Child Subglottal Resonances Database consisting of English-speaking adults reading various CVC utterances. The database was developed by the Speech Processing and Auditory Perception Group at UCLA and researchers at Washington University in St. Louis and Indiana University. A total of 50 adults (25 male, 25 female) were recorded saying various CVC utterances in the carrier phrase “I said a CVC again.” The list of utterances is the same as in The Child Subglottal Resonances Database shown in Table 2.1. Each CVC utterance, along with the associated carrier phrase, was repeated at least 10 times, and both microphone and subglottal accelerometer signals were recorded simultaneously in the same fashion as The Child Subglottal Resonances Database. Additionally, subglottal accelerometer signals of speakers sustaining the vowel /ɑ/ were recorded with an emphasis on high quality. Both microphone and accelerometer recordings were sampled at 48 kHz. This dissertation will use the recordings of h-vowel-d (hVd) words with 8 monophthongs (/ɑ/, /i/, /u/, /æ/, /ʌ/, /ɪ/, /ʊ/, /ɛ/). This data is available through the LDC. Further information about this corpus can be found in [LMA12].

2.3.2 The LibriSpeech ASR Corpus

The LibriSpeech ASR Corpus [PCP15] is a speech corpus consisting of English-speaking adults reading various audio books. The main training set consists of approximately 460 hours of clean speech and 500 hours of noisy speech for a total of 960 hours of speech from 2338 speakers (1210 male, 1128 female). The development set consists of approximately 5 hours of clean speech and 5 hours of noisy speech from 66 speakers (33 male, 33 female). Similarly, the testing set consists of approximately 5 hours of clean speech and 5 hours of noisy speech from 80 speakers (40 male, 40 female). The audio was sampled at 16 kHz. This dissertation will use the training set and testing set of the LibriSpeech ASR Corpus. Further information about this corpus can be found in [PCP15].

2.3.3 The OGI Kids' Speech Corpus

The OGI Kids' Speech Corpus, also known as the CSLU Kids' Speech Corpus, [SHC00] is a speech corpus consisting of English-speaking children from kindergarten to 10th grade. Each grade consisted of approximately 100 speakers, and each speaker participated in two speech tasks. The first task was a scripted speech task where the speaker would read either single words or phrases presented to them. Furthermore, for this task, the corpus also includes a corresponding text file documenting the accuracy of the child's reading for each audio file. The second task was a spontaneous speech task where the speakers were prompted to talk about certain topics such as their favorite movie. The audio was sampled at 16 kHz. This dissertation will use the scripted speech task of the OGI Kids' Speech Corpus. Only the files labeled as "1" in the verification files, which indicates that the audio file contains the speech of a child saying the correct word with limited noise, were used. This subset was further split into two subsets based on whether the child said a single word or a multiword phrase. Additionally, the number of utterances of the words "push" and "spoons" was substantially larger than that of the other words. As such, we only used a random subset of these words to remove any potential bias. Further information about this corpus can be found in [SHC00].

2.3.4 The CMU Kids Corpus

The CMU Kids Corpus [EMG97] is a speech corpus consisting of 76 English-speaking children from 1st to 3rd grade, as well as one child from 6th grade and one child from kindergarten for a total of 78 speakers. Each child was asked to read a series of several phrases or sentences for a total of 5180 read sentence utterances across all speakers. The audio was sampled at 16 kHz. In this dissertation, exactly 70% of these utterances (3626 utterances) were randomly chosen from this corpus to be used as a training set. The rest were used as a testing set. Further information about this corpus can be found in [EMG97].

2.4 Speech Tools and Software

2.4.1 Wavesurfer

Wavesurfer [SB00] is a publicly available speech analysis software developed by Kåre Sjölander and Jonas Beskow of The KTH Royal University of Technology using the Snack Sound Toolkit [Sjo97]. In this dissertation, Wavesurfer's spectrogram, formant, and pitch visualization tools were used. Further information about this software can be found in [SB00].

2.4.2 Praat

Praat [BW17] is a publicly available speech analysis software and feature computation toolbox developed by Paul Boersma and David Weenink of the University of Amsterdam. In this dissertation, Praat's automatic formant estimation tools were used, specifically the "To Formant (burg)" function. Further information about this software can be found in [BW17].

2.4.3 VOICEBOX

VOICEBOX [Bro06] is a speech processing toolbox consisting of MATLAB functions developed by Mike Brookes of Imperial College London. This toolbox was used for feature extraction in the ASR experiments. Further information about this toolbox can be found in [Bro06].

2.4.4 Multi-Band Summary Correlogram-based Pitch Detection

The Multi-Band Summary Correlogram (MBSC)-based Pitch Detection [TA13] is a pitch detection algorithm developed by Lee Ngee Tan and Abeer Alwan of UCLA, and the MATLAB code for the algorithm is available online (http://www.seas.ucla.edu/spapl/shareware.html). This algorithm extracts pitch by decomposing the frequency content of the signal with multiple wideband filters. This dissertation will use MBSC-based Pitch Detection as the pitch detection algorithm of choice; the f_o extraction methods from Praat and Wavesurfer were used for comparison. Further information about this algorithm and software can be found in [TA13].

2.4.5 Kaldi

Kaldi [PGB11] is an open-source speech recognition toolkit developed by researchers at Microsoft Research, Saarland University, Centre de Recherche Informatique de Montreál, SRI International, and Technical University of Liberec. The tookit includes C++ code and shell script recipes to train complete ASR systems including GMM and DNN-based acoustic models, language models, phonetic decision trees, and decoding graphs. This dissertation will use Kaldi to train and test ASR systems in Section 5. Further information about this tookit can be found in [PGB11].

2.5 Chapter Summary

In this chapter, we introduced several databases and speech analysis software tools that will be used throughout this dissertation. Both the Child Subglottal Resonances Database [YLG18] and the GFTA-JIBO Kids Corpus [YBA19a, YBA19b] were developed and published in part by the author. The publicly available databases introduced included the Subglottal Resonances Database [LMA12], LibriSpeech ASR Corpus [PCP15], OGI Kids' Speech Corpus [SHC00], and CMU Kids Corpus [EMG97]. The speech software tools introduced included Wavesurfer [SB00], Praat [BW17], VOICEBOX [Bro06], the MBSC Pitch Detection software [TA13], and Kaldi [PGB11]. The next chapter will discuss the modeling and usage of the subglottal signal in speech recognition.

CHAPTER 3

A Model of the Subglottal System for Children

The subglottal system and subglottal resonances (SGRs) play important roles in speech production and perception for both adults and children. The ability to reliably model SGRs is important for speech science and for utilizing the information contained in SGRs for technologies such as automatic speech recognition (ASR). In this chapter, we analyze the quarter-wavelength resonator model of the subglottal system using both adult and child SGR data. Adjustments are made to improve the model for child SGRs, specifically the third (SGR_3). Additionally, we examine the speech normalization properties of SGRs in child speech.

3.1 The Subglottal System Model

Large-scale investigations of adult SGRs have demonstrated that the adult subglottal system and corresponding SGRs are well-modeled by a tube model [LMA12, LAA11]. This is synonymous with a classical technique for modeling speech acoustics where the vocal tract is modeled as a series of tubes and formants are approximated as the resonances of the corresponding quarter-wavelength resonators for tubes with one end open, half-wavelength resonators for tubes with either both ends open or both ends closed, and Helmholtz resonators for large volumes with narrow openings [RS11]. However, due to the physical structure of the larynx, bronchi, and lungs, only a single quarter-wavelength resonator is necessary to model the SGRs of the subglottal system. An example of a vocal tract tube model for the production of the vowel /i/ and a subglottal tube model are shown in Figure 3.1. The vocal tract

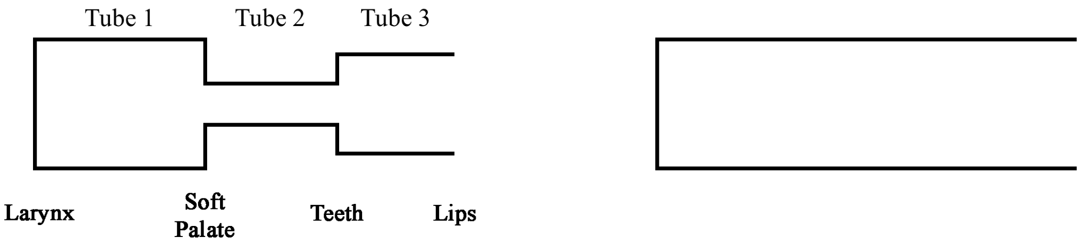


Figure 3.1: Approximate tube models of the vocal tract during the production of the vowel /i/ (left) and the subglottal system (right).

tube model consists of a quarter-wavelength resonator (Tube 3), half-wavelength resonators (Tube 1, Tube 2), and a Helmholtz resonator (the combination of Tube 1 and Tube 2). However, the subglottal system can be approximated with only a single tube quarter-wavelength resonator.

The resonances of a quarter-wavelength resonance model can be computed as follows:

$$R_N = \frac{(2N-1)c_{R_N}}{4l} \tag{3.1}$$

where R_N is the frequency of the N-th resonance in Hz, c_{R_N} is the propagation velocity of the wave for the N-th resonance, and l is the length of the tube. For the subglottal system, it has been shown that for adults, l can be approximated by scaling the speaker's height as $l = h/k_a$, where h is the speaker's height and k_a is the acoustic scaling factor that can be determined using SGR and height data [LMA12, LAA11]. As such the model for computing SGRs is given by the following:

$$SGR_N = \frac{(2N-1)c_{SGR_N}}{4h/k_a} \tag{3.2}$$

where SGR_N is the N-th SGR, and c_{SGR_N} is the speed of sound for that resonance. When modeling SGR_N where $N >= 2$, c_{SGR_N} is approximated as the speed of sound at body temperature, $c_0 \approx 35,900$ cm/s. However, it has been previously noted that c_{SGR_1} is larger than c_0 due to the inertive properties of the subglottal system tissue walls in the lower frequency range of SGR_1 [LAA11]. This value has been previously denoted as $c_{SGR_1} = c_w$ for the "walls" of the subglottal system, and this notation will be retained for the rest of this chapter. Thus, when using Eq. 3.2 to model the subglottal resonances, the values of k_a and c_w must first be determined using SGR and height data by minimizing the root mean-squared (RMS) error of the model estimates as follows:

1. Determine k_a using Eq. 3.2 and SGR_2 vs. height data.

2. Determine c_w using Eq. 3.2, k_a, and SGR_1 vs. height data.

To compute k_a and c_w, the first three SGRs of 50 adult speakers (25 male, 25 female) from The Subglottal Resonances Database [LMA12] were measured. The SGRs were estimated using Praat's "To Formant (burg)" function [BW17]. All estimates were then verified using a spectrogram and LPC visualization of the vowel spectrum, and manual corrections were performed as needed. These SGR data were used to estimate the modeling curves shown in Figure 3.2 using Eq. 3.2 and the minimum RMS error criterion. For the resulting model, $k_a = 8.795$ and $c_w = 46,532$ cm/s, which is less than a 0.2% difference from the values derived in [LMA12]. The minimum RMS errors for the SGR_1, SGR_2, and SGR_3 models were 45.2 Hz, 71.3 Hz, and 108.6 Hz, respectively. Measured SGR vs. height data are shown in Figure 3.2, along with the modeling curves. The SGR and height data clearly follow the general trend of the curves, demonstrating the validity of Eq. 3.2.

3.2 The Subglottal Resonance Model for Children

While Eq. 3.2 has been verified using adult speech, this model has yet to be evaluated using child speech. Of particular importance to the child subglottal system is the fact that children are shorter than adults, resulting in a shorter acoustic length and consequently higher SGRs.

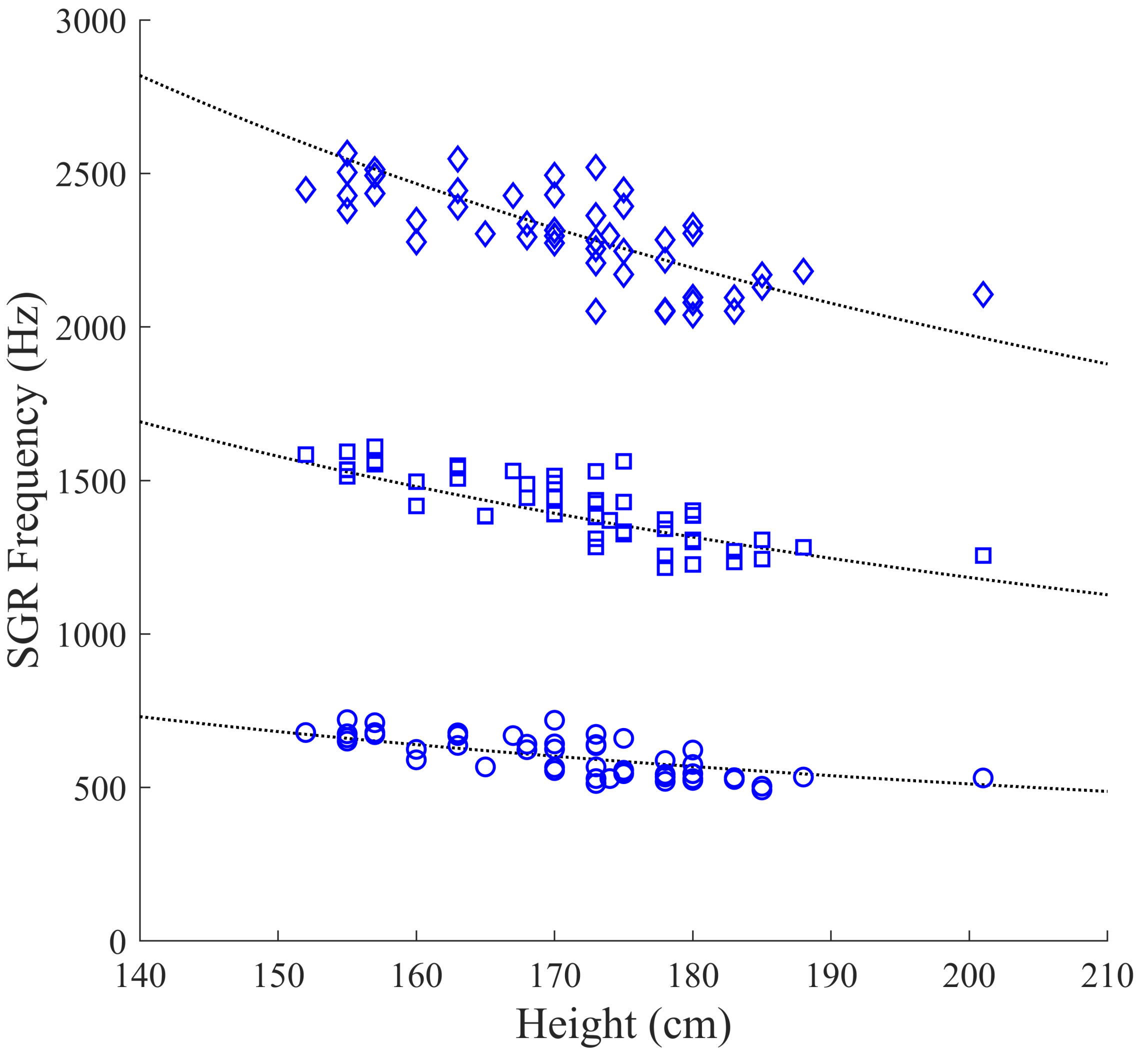


Figure 3.2: Measured SGR_1 (○), SGR_2 (□), and SGR_3 (◊) vs. height for 50 adult speakers. The modeling curves are estimated using Eq. 3.2 along with the minimum RMS error criterion.

To evaluate the validity of Eq. 3.2 for children, we measured the SGRs from 43 children (31 male, 12 female), aged 6-18 years old, from The Child Subglottal Resonances Database [YA18]. The SGRs were estimated using Praat's "To Formant (burg)" function [BW17]. All estimates were then verified using a spectrogram and LPC visualization of the vowel spectrum, and manual corrections were performed as needed. Additionally, the data were

Table 3.1: Minimum RMS estimates of k_a and c_w using Eq. 3.2 for SGR estimation, along with the RMS errors (ϵ_{RMS}) computed using either all adult and child data or only child data.

			Child ϵ_{RMS} (Hz)			Adult ϵ_{RMS} (Hz)		
Data	c_w (cm/s)	k_a	SGR_1	SGR_2	SGR_3	SGR_1	SGR_2	SGR_3
Child and Adult	45,400	8.760	88	136	190	49	75	108
Child	44,648	8.735	87	138	187	55	77	107

further supplemented by SGR and height data from 12 children (5 male, 7 female), aged 6-18 years old, reported in [LAM11], for a total of 55 children. Finally, the child SGR and height data were combined with SGR and height data of the 50 adults from Figure 3.2. These data were used to estimate k_a and c_w for the model in Eq. 3.2 using the minimum RMS error criterion and two sets of data. The first set included all 55 children and 50 adults. The second set only included the 55 children. The resulting model parameters and RMS errors for each of the SGRs is shown in Table 3.1. Measured SGR vs. height data are shown in Figure 3.3, along with the modeling curves. Notably, the curve estimated using all data and the curve estimated using only child data are almost identical.

Unlike the case for only adults in Figure 3.2, the minimum least squares regression curves for the combination of adults and children in Figure 3.3 systematically overestimate the values of SGR_3 and underestimate the values of SGR_2, especially at the higher frequencies and shorter heights. To understand this effect further, SGR_2 vs. SGR_1, SGR_3 vs. SGR_1, and SGR_3 vs. SGR_2 were plotted in Figure 3.4 using the same data from Figure 3.3. From Eq. 3.2, these plots are expected to follow the ratios $3c_0/c_w$, $5c_0/c_w$, and 5/3 for SGR_2/SGR_1, SGR_3/SGR_1, and SGR_3/SGR_2, respectively. These ratio lines are also plotted in Figure 3.4. Note that the SGR_3 vs. SGR_2 values are under the modeling line at higher frequencies. Similar to that which was observed in Figure 3.3, the SGR_3 values were overestimated.

A possible explanation for this phenomenon can be found in [FM78]. That study found that high frequencies penetrate deeper into the bronchi and lungs compared to lower fre-

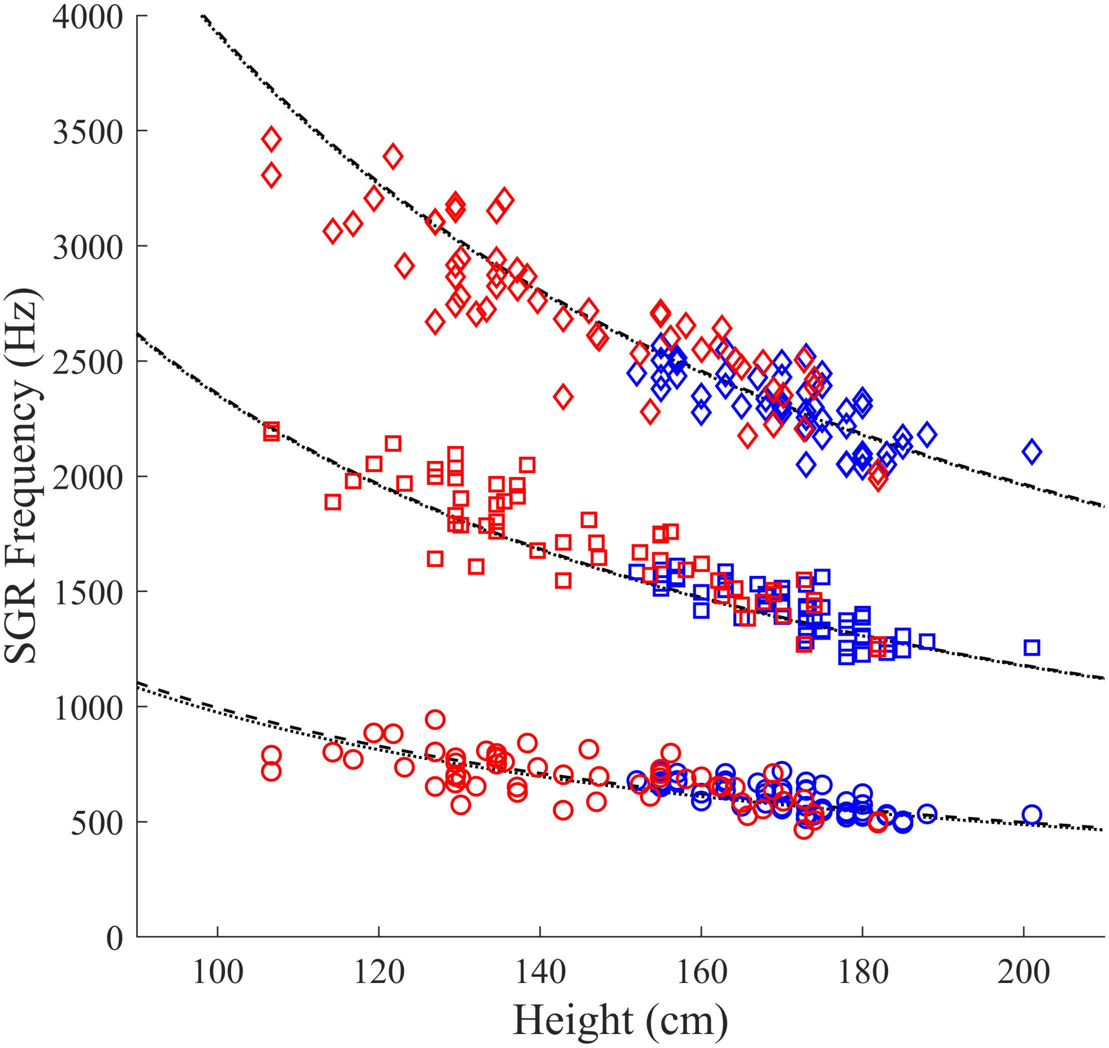


Figure 3.3: Measured SGR_1 (○), SGR_2 (□), and SGR_3 (◊) vs. height for 50 adult speakers (blue) and 55 child speakers (red). The modeling curves were estimated using Eq. 3.2 along with the minimum RMS error criterion with all data shown (dashed) and just the child data (dotted). At the higher frequencies, the modeling curves overestimate SGR_3 and underestimate SGR_2.

quencies. This is also consistent with similar frequency-dependent penetration depths documented in the acoustics of horns [Ben90, Pyl75]. With respect to the tube model of Eq. 3.2, this implies that l is larger than the adult approximation of $l = h/k_a$ as SGR_3 increases in frequency. Furthermore, the curving of the data in Figures 3.3 and 3.4 suggests that the

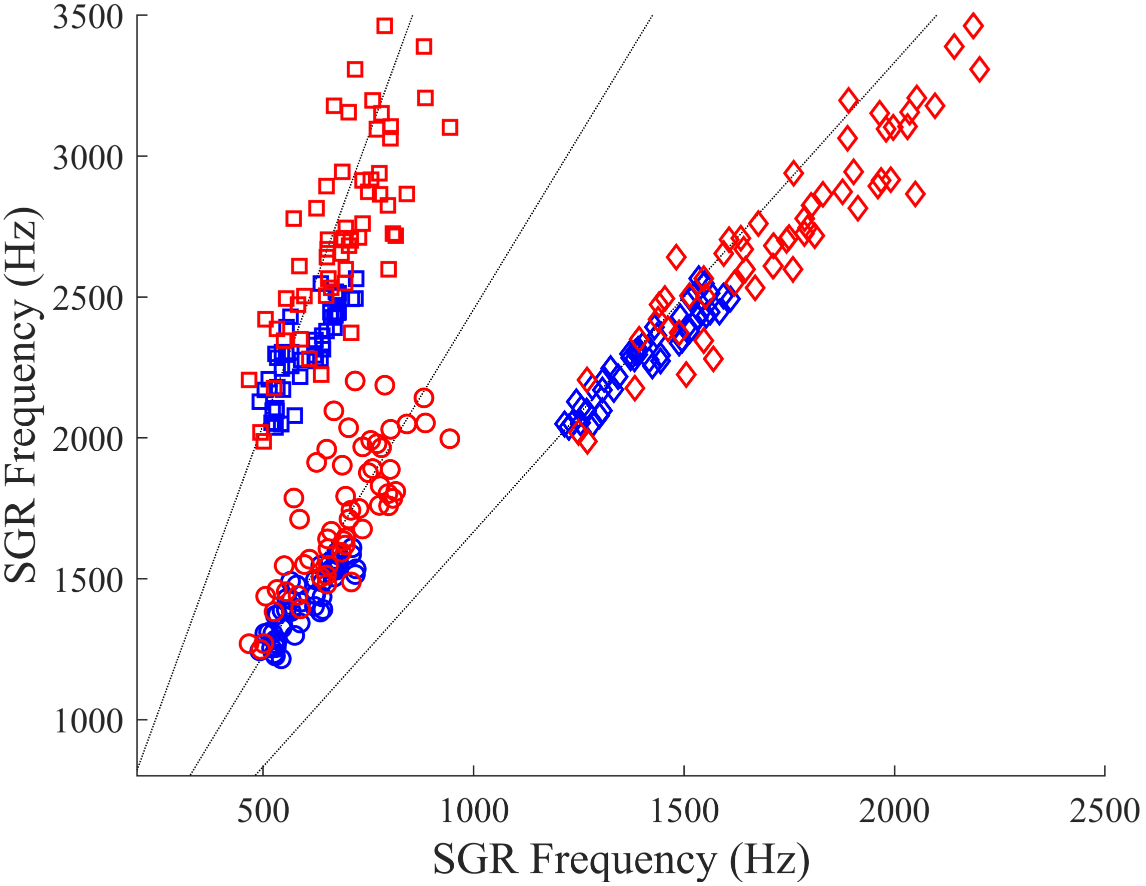


Figure 3.4: Measured SGR_2 vs. SGR_1 (○), SGR_3 vs. SGR_1 (□), and SGR_3 vs. SGR_2 (◊) for 50 adult speakers (blue) and 55 child speakers (red). The expected ratio lines derived from Eq. 3.2 are also displayed. The SGR_3 vs. SGR_2 values are lower than the modeling line, suggesting that SGR_3 increases slower than the model given by Eq. 3.2.

scaling of the effective tube length l with respect to h is non-linear.

To address this problem, we propose a refined tube model for SGR_3, in which the effective acoustic length of the subglottal system is modified, thus incorporating the penetration depth findings of [FM78]. The term l_a is introduced as the acoustic length of the subglottal system for SGR_1 and SGR_2, previously defined as $l_a = h/k_a$. To maintain consistency, the refined tube model replaces l_a with l_{a,SGR_3}, which represents the acoustic length for the SGR_3 model. Specifically, as height decreases, and SGR_3 correspondingly increases, the effective acoustic length of the tube model, l_{a,SGR_3} should decrease slower than l_a as higher frequencies penetrate further into the subglottal system than lower frequencies. This is implemented by defining the acoustic length of SGR_3 as a function of l_a. The model is implemented as follows:

$$SGR_3 = \frac{5c_0}{4l_{a,SGR_3}(l_a)} \tag{3.3}$$

$$l_{a,SGR_3}(l_a) = l_a + f(l_a) \tag{3.4}$$

where the effective acoustic length for SGR_3 can be derived from the acoustic length used to model SGR_1 and SGR_2, $l_a = h/k_a$, and some additive function representing the additional penetration depth of the higher frequencies. Such a function must follow certain assumptions within reasonable values of l_a:

- $f(l_a)$ is positive
- $f(l_a)$ is smaller than l_a
- $f(l_a)$ approaches 0 as l_a becomes large
- $\frac{\partial f(l_a)}{\partial l_a}$ is greater than -1 to ensure l_{a,SGR_3} is monotonically increasing

The proposed candidate function is as follows:

$$f(l_a) = \frac{l_a}{1 + e^{(\alpha l_a - \beta)}} \tag{3.5}$$

where additional parameters α and β are determined by minimizing the RMS error over a set of child SGR_3 data.

While the original model required two parameters, k_a and c_w, to be determined using SGR data, this new model requires the estimation of two additional parameters for a total of four parameters, k_a, c_w, α, and β. With sufficient SGR and height data from adults and children, the minimum RMS error criterion can be used as follows:

Table 3.2: Minimum RMS estimates of k_a, c_w, α, and β using Eq. 3.2 for SGR_1 and SGR_2 and Eq. 3.3-3.5 for SGR_3, along with the RMS errors (ϵ_{RMS}) computed using either all adult and child data or only child data.

					Child ϵ_{RMS} (Hz)			Adult ϵ_{RMS} (Hz)		
Data	c_w (cm/s)	k_a	α (1/cm)	β	Sg1	Sg2	Sg3	Sg1	Sg2	Sg3
Child and Adult	43,849	9.070	0.235	0.805	88	122	147	49	68	107
Child	42,735	9.126	0.298	1.704	87	123	147	55	71	116

1. Determine k_a using Eq. 3.2 and SGR_2 vs. height data.

2. Determine c_w using Eq. 3.2, k_a, and SGR_1 vs. height data.

3. Determine α and β using Eq. 3.3–3.5, k_a, and SGR_3 vs. height data.

To verify this new model, the estimation procedure above was followed using all the data from Figure 3.3 to produce regression curves with the new model using two sets of data. The first set included all 55 children and 50 adults. The second set only included the 55 children. The resulting model parameters and RMS errors for each of the SGRs is shown in Table 3.2. The new modeling curves are shown along with the SGR vs. height data in Figure 3.5.

While the RMS error did not decrease for SGR_1, the RMS errors for SGR_2 and SGR_3 decreased by at least 9.5% and 21.4%, respectively, for children. Similarly, the RMS error for SGR_2 decreased by at least 5.3% for adults, thus further demonstrating the improvement of the model defined by Eq. 3.2–3.5 rather than just Eq. 3.2. Similarly, the modeling ratios of SGR_2/SGR_1, SGR_3/SGR_1, and SGR_3/SGR_2 are shown along with the SGR_2 vs. SGR_1, SGR_3 vs. SGR_1, and SGR_3 vs. SGR_2 data in Figure 3.6. In both Figures 3.5 and 3.6, the modeling curves clearly fit the data better than Figures 3.3 and 3.4, respectively.

3.3 The Normalization Properties of Subglottal Resonances

One of the relevant properties of SGRs in speech production is the SGRs ability to divide the vowel space, as noted by several past studies [ALL13, CS07, MLW08, CBG09]. Due

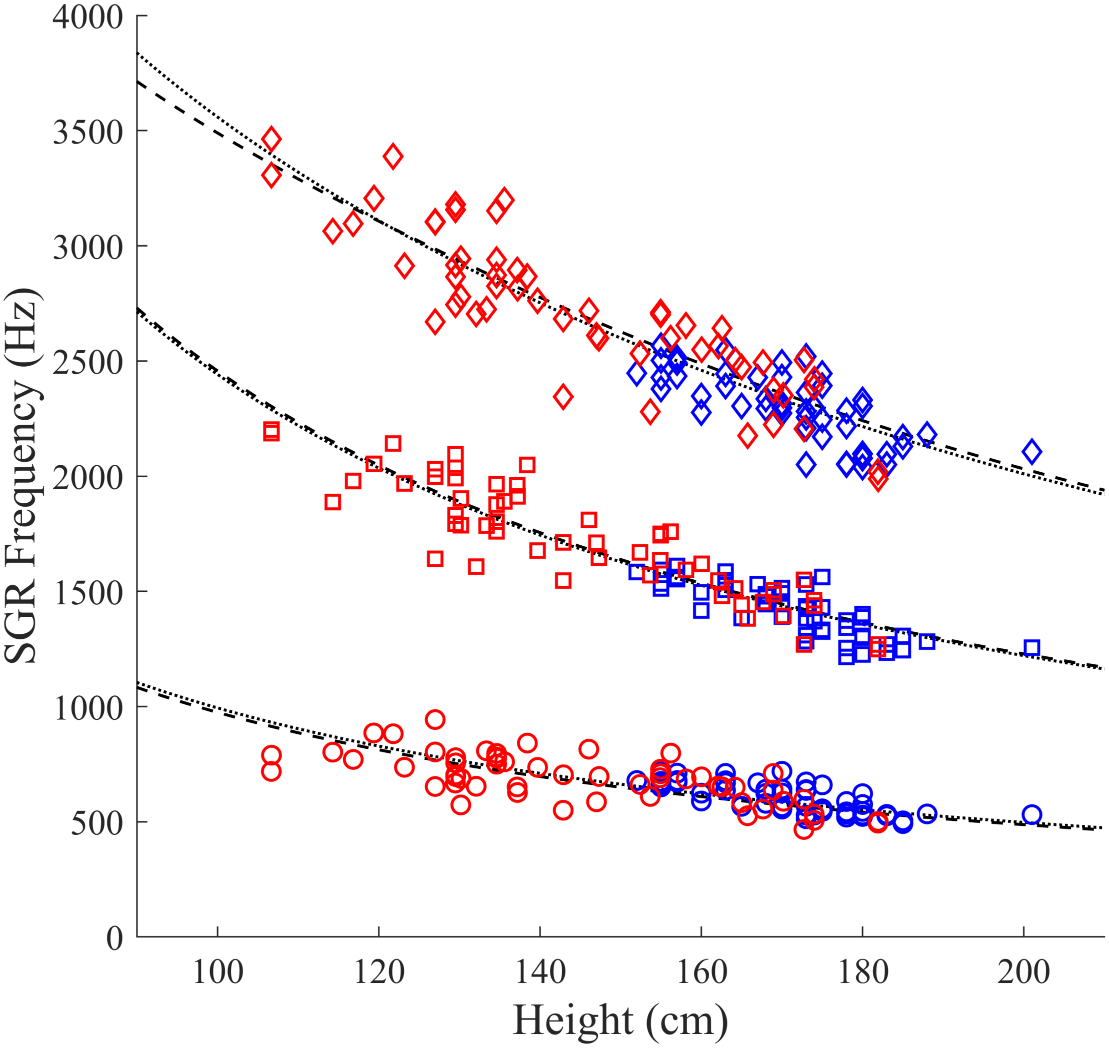


Figure 3.5: Measured SGR_1 (○), SGR_2 (□), and SGR_3 (◊) vs. height for 50 adult speakers (blue) and 55 child speakers (red). The modeling curves for SGR_1 and SGR_2 were estimated using Eq. 3.2, and the modeling curve for SGR_3 was measured using Eq. 3.3-3.5. All models were derived using the minimum RMS error criterion with all data shown (dotted) and just the child data (dashed). Compared to Figure 3.3, Eq. 3.3-3.5 are more effective at modeling SGR_3.

to the coupling of the subglottal system with the vocal tract, resonances in the subglottal acoustic space manifest themselves as zeros in the speech acoustic space, often interfering with formants, which act like poles in the speech acoustic space [RS11]. As such, it has been

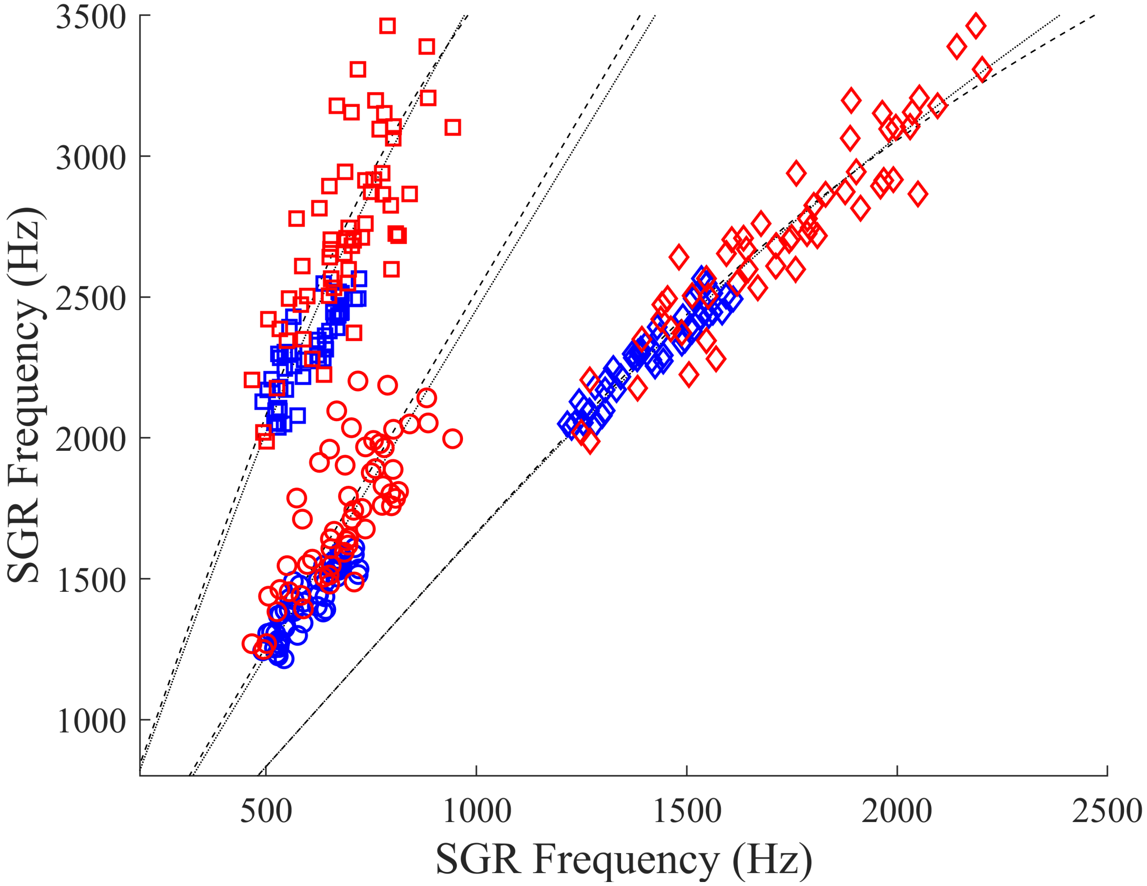


Figure 3.6: Measured SGR_2 vs. SGR_1 (○), SGR_3 vs. SGR_1 (□), and SGR_3 vs. SGR_2 (◊) for 50 adult speakers (blue) and 55 child speakers (red). The expected ratio lines with SGR_1 and SGR_2 were derived from Eq. 3.2, and the expected ratio lines with SGR_3 were derived using Eq. 3.3-3.5. All models were derived using the minimum RMS error criterion with all data shown (dotted) and just the child data (dashed). Compared to Figure 3.4, the ratio line for SGR_3 vs. SGR_2 displayed in this figure better fits the data.

hypothesized that speakers actively avoid these zeros when producing vowels to prevent a reduction in formant energy [Lul10].

Further investigations into SGRs have revealed that SGR locations have a quantal effect [Ste98, LBM07], dividing the vowel space into regions that separate front and back vowels as well as high and low vowels. This was demonstrated in [ALL13], which showed that a speaker's SGR_1 and SGR_2 divide the speaker's F_2 vs. F_1 space into four quadrants, each containing the F_2 vs. F_1 location of one of the four corner vowels (/ɑ/, /i/, /u/, /æ/). Specifically, SGR_1 serves as a threshold for the value of F_1 when determining high and low vowels, and SGR_2 serves as a threshold for F_2 when determining front and back vowels.

To examine the effectiveness of the SGRs in normalizing the vowel space of child speech, the first and second formants (F_1, F_2) of the vowels /ɑ/, /i/, /u/, and /æ/ were measured from each of the corresponding utterances from the 43 children in The Child Subglottal Resonances Database. The formants were normalized by dividing F_1 by SGR_1 and F_2 by SGR_2 such that all speakers have the same vowel space threshold locations. The scatterplot of F_2/SGR_2 vs. F_1/SGR_1 is shown in Figure 3.7. The dividing lines in the figure are located at $F_1/SGR_1 = 1$ and $F_2/SGR_2 = 1$.

Observing Figure 3.7, the formants of child vowel utterances have a similar relationship with SGRs as that of adults. For the child formant data in Figure 3.7, when considering typical F_1 and F_2 placements with respect to the SGR boundaries, the vowels followed the quantal patterns for /ɑ/, /i/, /u/, /æ/ 81.7%, 100.0%, 84.4%, and 86.8% of the time, respectively. The utterances that were atypical can be explained by mispronunciations or variability due to the acoustic instability of child speech. However, the quantal trend still remains, suggesting the effectiveness of SGRs for speech normalization.

3.4 Chapter Summary

In this chapter, we proposed a modification to the existing quarter-wavelength resonator model for SGRs to better accommodate child SGRs without sacrificing the modeling of adult SGRs. Furthermore, this modification demonstrates the importance of considering the penetration depth of higher frequencies when modeling child SGRs. Additionally, normalization using SGR_1 and SGR_2 was shown to be reasonably effective at dividing the vowel space of children, justifying the use of SGRs as normalization factors in feature normalization procedures [GPY15]. The next chapter will propose the use of fundamental frequency

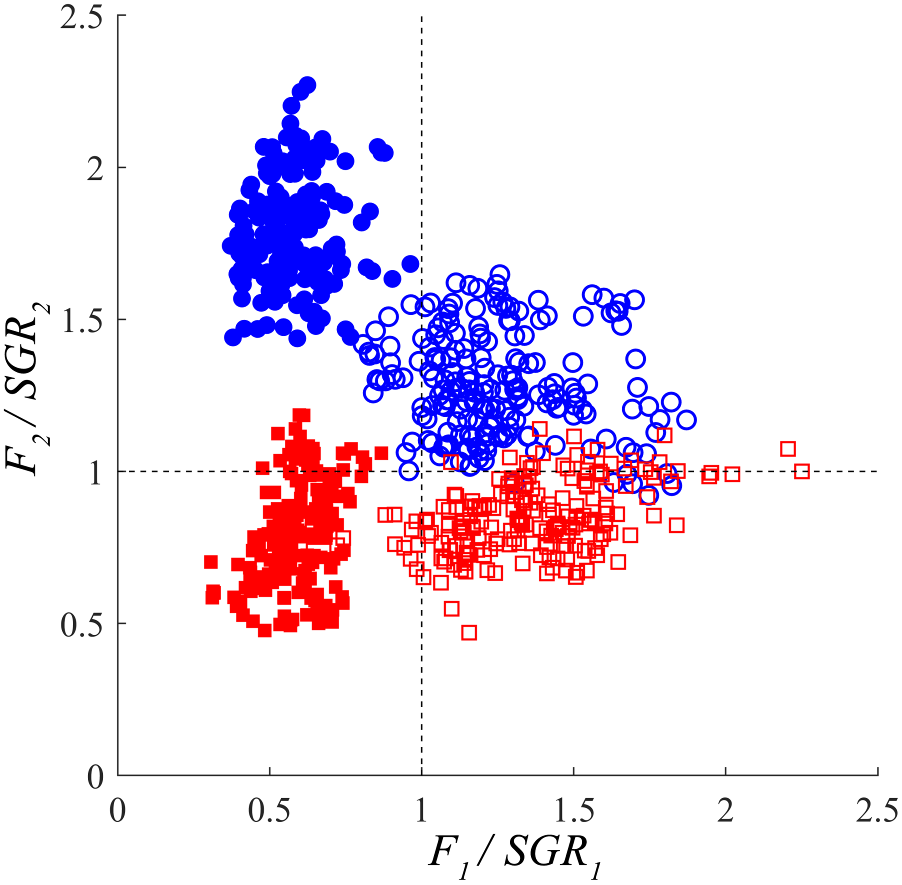


Figure 3.7: F_2/SGR_2 vs. F_1/SGR_1 for the vowels /u/ (■), /ɑ/ (□), /i/ (●), and /æ/ (○) from utterances by 43 children. The vowel boundaries as defined by SGR_1 and SGR_2 are normalized to be 1. In general, each vowel is mainly located in a single quadrant.

(f_o), another physical parameter of speech, as an alternative normalization factor.

CHAPTER 4

Tonotopic Distances and f_o-based Warping

In this chapter, we will review the relationship between the fundamental frequency (f_o) and formants using the tonotopic distances, a metric developed to model human speech perception [SG86, Tra81, FDT96, CL79, Chi85]. These tonotopic distances are reformulated such that a warping method for child speech can be derived using only a speaker's median f_o. This warping method can be used for both feature normalization and data augmentation when applied to child ASR systems.

4.1 The Tonotopic Distances Between Formants and f_o

4.1.1 Modeling

The tonotopic distances are a set of features given by the distance between adjacent formant frequencies ($F_{n+1} - F_n$ for $n \in \{1, 2, 3, \ldots\}$), along with the distance between the first formant and fundamental frequency ($F_1 - f_o$), in some perceptual frequency scale such as the Mel or Bark scale. For consistency, the rest of this dissertation will use the Mel scale. A number of previous studies have found success in modeling human vowel perception using the tonotopic distances [SG86, Tra81, FDT96, CL79, Chi85]. That is, vowel utterances are more likely to be perceived as the same vowel if the vowel utterances' tonotopic distances are similar.

An equivalent feature representation of this set of tonotopic distances can be formulated as the differences between each formant and f_o ($F_n - f_o$ for $n \in \{1, 2, 3, \ldots\}$). This can be shown by noting that the tonotpic distance $F_{n+1} - F_n$ can be obtained using $F_{n+1} -$

f_o and $F_n - f_o$ by taking the difference $(F_{n+1} - f_o) - (F_n - f_o)$. This reformulation has several implications for vowel space modeling. Notably, this implies that f_o, which is often relegated to the voice source and separated from the vocal tract analysis of speech production [RS11], plays a fundamental role in vowel perception. That is, a linear relationship (with a slope of 1) exists between a vowel's formants and f_o in the perceptual frequency scale. It is known that children have much higher formant and f_o values than adults [LPN97, LPN99]. As such, the relationship between formant and f_o values in child or adult speech may be perceptually causal rather than simply correlated. Furthermore, this reformulation removes the dependence of formant frequencies on adjacent formant frequencies, instead being dependent on f_o. This property will be used in Section 4.2 to formulate a normalization technique dependent solely on f_o.

4.1.2 Analysis

To examine this linear relationship in the perceptual space, the vowel productions of hVd utterances from all 43 speakers in The Child Subglottal Resonances Database were analyzed. Of the 14 vowels, four tense vowels, /ɑ/, /i/, /u/, and /æ/, and four lax vowels, /ʌ/, /ɪ/, /ʊ/, and /ɛ/, were chosen for analysis. The values of f_o, F_1, F_2, and F_3 were measured. All measurements were done with Praat [BW17] with manual corrections as needed. Least-squares linear regression lines relating F_1, F_2, and F_3, to f_o for each vowel were computed. Of all the regressions, the slopes of 19 of the 24 regression lines were between 0.70 and 1.30, reasonably close to the expected value of 1. Furthermore, these slopes contributed significantly to the regression ($p < 0.001$) with Pearson's correlation coefficients greater than $r > 0.5$.

The formant and f_o data of the 8 vowels are displayed in Figure 4.1, along with least-squares regression lines fixed to have a slope of 1 to compare the data against the expected modeling line. Clearly, the data follow the regression lines with an upward trend in formant values as f_o increases, demonstrating the validity of the reformulation. It should be noted that as f_o increases, the variability in formants for some of the vowels also increases such as for F_1 of the vowel /i/. This is likely due to mispronunciation or the speech variability of children. Regardless, even with such variability, the linear trend is still apparent.

4.2 f_o Normalization

Based on the reformulation of the tonotopic distances, we can derive a frequency normalization technique in the spectral domain that relies solely on the value of a speaker's f_o. First, we note that according to the tonotopic distance model of speech perception, when f_o and the formants (F_1, F_2, F_3, ...) are measured on some perceptual scale, the difference between the formants and f_o ($F_1 - f_o$, $F_2 - f_o$, ...) should be constant across different utterances of the same vowel. Thus, by measuring a speaker's f_o, we can normalize the formants of a vowel to a default value as follows:

$$F_{n,norm} = F_{n,orig} - (f_{o,utt} - f_{o,def}) \tag{4.1}$$

for $n \in \{1, 2, 3, \ldots\}$ where $f_{o,utt}$ is the f_o of the utterance, $f_{o,def}$ is a predetermined value of f_o to represent a default speaker, $F_{n,orig}$ is the n-th formant in the original utterance, $F_{n,norm}$ is the n-th formant after normalizing to $f_{o,def}$, and all frequencies are measured in the perceptual Mel scale.

Although Eq. 4.1 has been formulated specifically for formant normalization, this comes with the complication of formant estimation, which is unreliable for children or speakers with high f_o and formant values. Instead, we can avoid direct manipulation of formants by further generalizing Eq. 4.1 to normalize the entire spectrum as follows:

$$f_{norm} = f_{orig} - (f_{o,utt} - f_{o,def}) \tag{4.2}$$

where f_{orig} is some frequency in the original spectrum and f_{norm} is the corresponding fre-

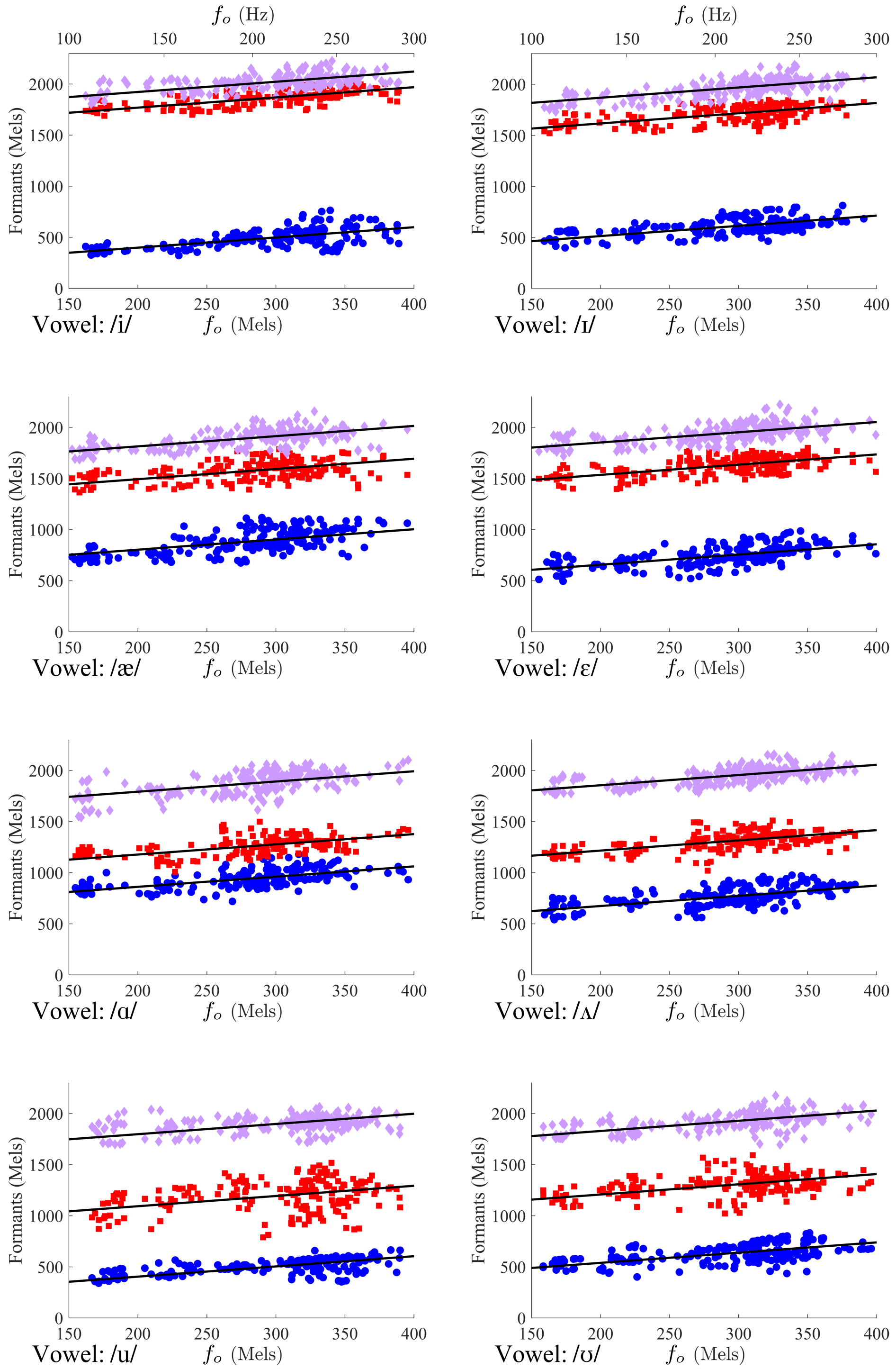


Figure 4.1: F_1 vs. f_o (blue), F_2 vs. f_o (red), and F_3 vs. f_o (purple) for the vowels /i/, /æ/, /ɑ/,/u/, /ɪ/, /ɛ/, /ʌ/, and /ʊ/ from corresponding hVd words of children in The Child Subglottal Resonances Database. Also shown are the least-squares linear regression lines, fixed to have a slope of 1. The data follow the linear relationship implied by the reformulation of the tonotopic distances.

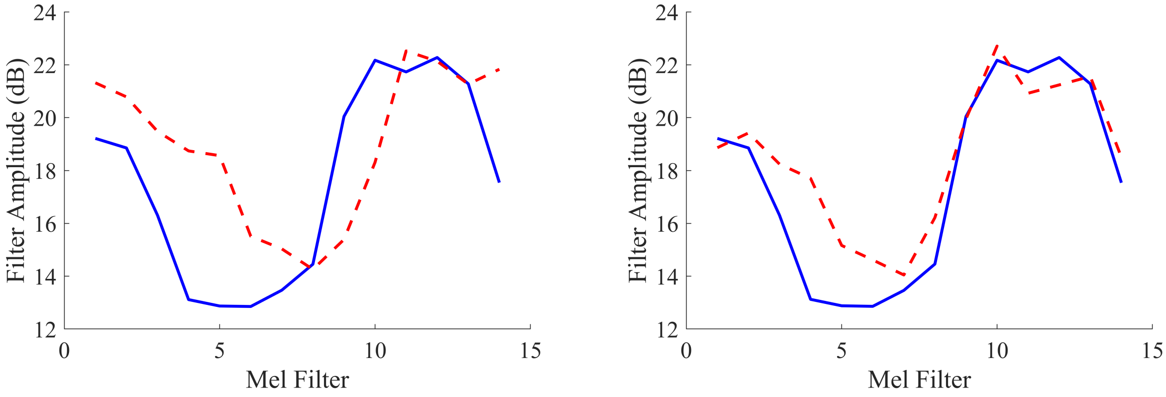


Figure 4.2: Mel filter bank outputs of the vowel /i/ spoken by an 18-year-old male (solid) and a 7-year-old male (dashed) computed with 15 filter banks using a frequency range of 20 Hz to 6 kHz. The filter bank outputs are computed both without (left) and with (right) f_o normalization. When normalization is applied, default f_o is chosen to be $f_{o,def} = 100$ Hz. The 18 year old male had $f_{o,utt} = 106$ Hz, and the 7 year old male had $f_{o,utt} = 270$ Hz. The filter bank outputs computed with f_o normalization are better aligned than the outputs without f_o normalization. Note that this normalization technique warps every speaker to the same default space ($f_o = 100$ Hz) rather than warping one speaker to another speaker's speech space.

quency in the normalized spectrum. That is, the frequency content in f_{orig} is shifted to f_{norm}. In the case of a discrete spectrum, f_{norm} can be reinterpreted as the normalized frequency corresponding to some discrete Fourier transform (DFT) index and f_{orig} as the frequency from the original spectrum mapped to the index of f_{norm}. Note that the value of $f_{o,def}$ need not be speaker nor vowel dependent and can be empirically chosen; hence $f_{o,def}$ acts as a correction factor. We have named this method f_o normalization.

An example of f_o normalization is shown in Figure 4.2. The Mel filter bank outputs of an 18-year-old male and 7-year-old male saying the vowel /i/ are displayed both with and without f_o normalization. When f_o normalization is applied, the filter bank outputs for the two speakers become more similar.

4.3 f_o Perturbation

While f_o normalization uses Eq. 4.2 to reduce variability between speakers by fixing $f_{o,def}$ to some default value and adjusting $f_{o,utt}$ based on the speaker, an alternative procedure is to use Eq. 4.2 as a data augmentation method to create variability in the data. To perform data augmentation, features are extracted several times from a single speech utterance while changing $f_{o,def}$. The resulting set of features is consistent with the structure of speech as defined by the tonotopic distances and can be used to augment training data. This is particularly useful for training large neural-network-based systems, which often requires large amounts of training data. We refer to this method as f_o perturbation.

Furthermore, f_o normalization and f_o perturbation can be used simultaneously. This is done by setting $f_{o,utt}$ to be the f_o of the utterance and choosing multiple values for $f_{o,def}$. This procedure allows us to remove large inter-speaker variabilities while generating speech-like features to further augment the training data.

4.4 f_o Warping Parameter Considerations

For f_o warping to be effective, several parameter considerations must be taken into account. For f_o normalization, the default value of $f_{o,def}$ must be chosen beforehand so features can be normalized to the same default speech space. After initial experimentation, we found that $f_{o,def} = 100$ Hz or 150 Mels is a reasonable choice for optimizing ASR word error rate.

A further consideration is the approximation of f_o when performing f_o normalization. It is desirable to use the median f_o over the speech utterance as the value of $f_{o,utt}$. Using a median can normalize outliers from the f_o estimation process and compensate for intra-speaker variability. Additionally, storage of a single number to be applied to a speaker's speech is an efficient way to quickly adapt an ASR system as opposed to requiring multiple values of f_o during feature extraction for every individual speaker. DFT computation over an utterance can also be implemented much more efficiently when only one value of $f_{o,utt}$ is used, resulting in faster computation times during feature extraction.

For f_o perturbation, additional values for $f_{o,def}$ must be chosen. This can be done by perturbing the initial choice of $f_{o,def}$. In our experiments, we chose to increase the amount of data by 7-fold. After preliminary experimentation, we found that adding ±20, ±40, and ±60 Mels to the initial value of $f_{o,def}$ is effective. For example, when choosing an initial $f_{o,def} =$ 100 Hz, the f_o perturbation method would use $f_{o,def} \in \{58.52, 72.10, 85.93, 100.00, 114.32, 128.90, 143.74\}$ Hz, using all values for every utterance when extracting features.

4.5 Chapter Summary

In this chapter, we examined the tonotopic distances between formants and f_o, and reformulated these distances as linear relationships between formants and f_o in the perceptual frequency scale. This reformulation motivated the f_o warping technique, which is capable of warping the frequency spectrum of an utterance while maintaining the speech-like structure. The warping technique can be used for normalization, data augmentation, or both simultaneously. Additional parameter considerations were also discussed. The next chapter will analyze the performance of ASR systems for young children, examine the effectiveness of f_o warping, and compare the warping to SGR-based normalization and other state-of-the-art normalization and data augmentation techniques.

CHAPTER 5

ASR Experiments and Results

In this chapter, we report on ASR experiments to assess their performance on child speech in several scenarios. The f_o-based warping techniques are applied to child ASR training and testing and compared with SGR-based normalization and other state-of-the-art normalization and data augmentation techniques. Both single word and continuous speech ASR systems are analyzed. We focus on kindergarten-aged children as many of these children are pre-literate, and educational applications and child human-robot interactions (HRI) can benefit greatly from an improved child ASR system for accessibility [KLM17].

5.1 The Effect of Age on ASR Performance

5.1.1 Database

The first set of experiments analyzed the effect of age on ASR performance using the OGI Kids' Speech Corpus. To eliminate the confounding factor of a child language model, only single words from the scripted speech task were used to perform single word recognition. Utterances contained a total of 208 possible words. These words ranged from short, easy words (e.g., "chair") to longer, more difficult words (e.g., "organization"). The data were separated into 11 subsets by grade (kindergarten to 10th grade), and 1,654 word utterances were randomly chosen from each grade to ensure a fair comparison between grades.

5.1.2 Matched-Grade Experimental Setup

For the matched-grade experiments, a ten-fold cross-validation ASR experiment was performed for each grade where 90% of the data was used for training a triphone-based ASR system and the remaining 10% was used as testing data. The language model was based on single words with all words being equally probable.

For each system, 13-dimensional Mel-frequency cepstral coefficients (MFCCs) were extracted with a window size of 25 ms, frame shift of 10 ms, 23 filters, a lifter coefficient of 22, and bandwidth of 4 kHz. Cepstral mean normalization (CMN) was used. An additional 7-frame linear discriminant analysis (LDA) was also used for a final 40-dimensional feature input. While we also tested the MFCC features with first derivatives (delta) and second derivatives (delta-delta), these features did not perform as well as the the system with a 7-frame concatenation and will not be reported.

Due to the small size of the training data and scale of the word recognition task, ASR systems were trained on 250, 500, or 1000 triphones to examine the effect of increasing the number of triphones on ASR performance. Both Gaussian mixture model Hidden Markov model-based (GMM-HMM) and deep neural network Hidden Markov model-based (DNN-HMM) ASR systems were evaluated. DNNs were trained on an additional 9-frame LDA with 2 hidden layers and 2-norm non-linearities with input dimension of 500 and output dimension of 100. Feature space maximum likelihood linear regression (fMLLR) speaker adaptive training was also used. However, we did not use vocal tract length normalization (VTLN) as it was found not to be helpful. All systems were trained using the Kaldi ASR toolkit [PGB11].

5.1.3 Kindergarten Mismatched-Grade Experimental Setup

For the mismatched-grade experiments, the same systems that were trained using the DNN-HMM 250 triphone systems for 1st to 10th grade from the matched-grade experiments were

Table 5.1: Word error rates (WERs) (%) of ASR systems for the single word matched-grade experiments. Features were extracted with a 4 kHz bandwidth. Systems were trained with fMLLR speaker adaptive training. Both GMM and DNN-based acoustic models are shown with the number of triphones used in parentheses. The kindergarten ASR performed dramatically worse than older grades and was more affected by the number of triphones.

System	**Grade**										
	K	1	2	3	4	5	6	7	8	9	10
GMM(250)	28.32	15.39	11.76	11.12	5.98	6.30	4.88	6.24	3.75	4.25	3.85
GMM(500)	30.08	17.28	12.99	11.18	5.61	7.34	5.42	5.89	6.16	4.57	4.23
GMM(1000)	35.93	19.55	14.57	12.86	6.34	7.10	5.85	6.54	6.90	5.28	4.88
DNN(250)	26.91	14.64	10.50	10.39	4.65	4.64	4.78	5.39	3.34	3.58	3.56
DNN(500)	26.34	16.18	9.51	9.54	4.50	5.46	4.15	5.42	3.57	3.39	3.80
DNN(1000)	30.30	16.06	10.69	10.06	5.22	5.09	5.20	5.14	3.65	3.57	4.05

used. These ASR system were chosen as they performed the best in the experiment explained in Section 5.1.2. The systems were tested with kindergarten speech. Additionally, VTLN was used to reduce the mismatch between age groups.

5.1.4 Results and Discussion

The results of the matched-grade experiment for both the GMM-HMM and DNN-HMM systems are shown in Table 5.1. Notice three major performance jumps between neighboring grades. Between kindergarten and 1st grade, the word error rate (WER) had an absolute decrease of 10%, resulting in a relative decrease of over 38%. Between 1st and 2nd grade, the WER had an absolute decrease of between 3-7%, resulting in a relative decrease of over 23%. Finally, between 3rd and 4th grade, the WER had an absolute decrease of 5%, resulting in a relative decrease of almost 50%. After 4th grade, the WER levels stabilized to approximately 5%, which is similar to the performance on adult speech. This indicates four major age groups in terms of ASR performance: kindergarten, 1st grade, 2nd and 3rd grade, and 4th grade and above.

Observing the GMM-HMM ASR performance in Table 5.1, varying the number of triphones in the acoustic model seems to have different effects depending on grade. For the kindergarten ASR system, WER decreased significantly when the number of triphones was decreased from 1000 to 500 ($p < 0.001$). The 1st grade ASR system also showed a similar performance improvement when the number of triphones was decreased from 1000 to 500, albeit at a lesser significance level ($p < 0.05$). For the 2nd grade and older ASR systems, changing the number of triphones did not have a significant effect except for the 8th grade system, which is likely an outlier.

All DNN-HMM models seemed to perform comparably or better than the GMM-HMM models. However, the significant effect of changing the number of triphones for 1st grade and 8th grade systems disappeared. The kindergarten ASR system still improved significantly when the number of triphones decreased from 1000 to 500 ($p < 0.01$).

Past studies on child speech suggest that young children do not have the ability to articulate speech in a consistent manner [GGN09, SBM87, ZHG15]. Thus, the inclusion of additional triphones does not provide additional benefit when training ASR systems for young children. In fact, as the number of triphones increases, the amount of data available to train each triphone decreases. As the dataset was small and young children are inconsistent in pronunciation, increasing the number of triphones results in poor training conditions for young child ASR.

It is important to note that a single age difference can cause a dramatic degradation in performance, such as the difference between the kindergarten and 1st grade systems. This is particularly important as this implies that age groups may need to be specifically targeted and evaluated when training ASR systems, rather than the conventional method of grouping a large age range together as "child" speech data.

The results of the mismatched-grade experiments are shown in Table 5.2. The testing data used were always kindergarten speech. Unsurprisingly, the systems trained on older child speech (6th-10th grade) performed poorly on kindergarten speech. While VTLN was effective at improving the performance of these systems, the performance still failed to reach the level of the systems trained on younger child speech (1st-5th grade).

Table 5.2: Word error rates (WERs) (%) of ASR systems for the single word mismatched-grade experiments. Features were extracted with a 4 kHz bandwidth. Each ASR was trained on a single grade level and tested on kindergarten speech. The systems tested were equivalent to the DNN-based ASR systems with 250 triphones and fMLLR in the matched-grade experiments. Additionally, VTLN and SGR feature normalization were used and found to be effective on systems trained on older children. The best performing system (in **boldface**) was trained on 1st grade speech with no feature normalization.

Feature	**Training Grade**										
Normalization	K	1	2	3	4	5	6	7	8	9	10
None	26.91	**23.11**	24.80	25.38	26.45	24.83	28.64	31.64	36.58	39.00	43.62
VTLN	28.49	24.04	26.75	25.40	26.00	24.17	26.11	29.63	31.65	32.25	34.85
SGR	28.07	26.53	26.96	25.79	26.85	25.82	28.02	29.54	32.33	33.36	35.02

Notably, systems trained on younger child speech did not show any benefit from VTLN. This is likely due to the fact that these younger children are more physiologically similar to kindergarten children than the older children. Thus, the normalization seems to only have an effect when the difference in age is large.

Interestingly, the ASR system trained on 1st grade speech performed significantly better than the system trained on kindergarten speech when tested on kindergarten speech. This may suggest that 1st grade speech is more suitable for training a kindergarten ASR than even kindergarten speech. This is likely due to the large reductions in speech variability the children experience as they age [LPN97, LPN99]. Furthermore, some educators hypothesize that child speech becomes less variable as the children learn to read (which typically occurs in 1st grade), further justifying the results.

5.2 The Effect of SGR and f_o Normalization on Single-Word Child ASR

5.2.1 Database

This experiment examined the effects of f_o normalization using the OGI Kids' Speech Corpus. The data subsets were setup in the same way as described in Section 5.1. That is, the data were separated by grade, and 1,654 single-word utterances were randomly chosen from each grade.

5.2.2 Experimental Setup

The features used were 13-dimensional MFCCs extracted with a window size of 25 ms, frame shift of 10 ms, and a lifter coefficient of 22. When extracting MFCCs, a bandwidth of 5.2 kHz was chosen such that F3 was contained in the signal for all children. During feature extraction, several spectral warping normalization strategies were used including no normalization (baseline), VTLN, F3-based normalization, SGR normalization, and the proposed f_o normalization with $f_{o,def} = 100$ Hz. The number of filters used for extraction varied, with features normalized with f_o normalization using 15 filters and all other features using 19 filters. The number of filters used was empirically chosen. Additionally, CMN was also applied to all features.

For the mismatched-grade experiments, the data from the 1st to 10th grade children were used as training data, separated by grade. For each grade, all 1,654 word utterances were used to train a DNN-HMM ASR system with 250 triphones. After feature extraction, a 7-frame LDA and fMLLR were applied for a final 40-dimensional feature input. DNNs were trained on a 9-frame LDA, 2 hidden layers, and 2-norm non-linearities with an input dimension of 500 and output dimension of 100. All ASR systems were trained using the Kaldi ASR Toolkit. Each system was tested using all 1,654 word utterances from the kindergarten

Table 5.3: Word error rates (WERs) (%) of DNN-HMM ASR systems for the single word mismatched-grade experiments. Features were extracted with a bandwidth of 5.2 kHz. Each ASR system was trained on a single grade level (1st-10th grade) and tested on kindergarten speech. MFCCs were extracted with no normalization, f_o normalization, VTLN, $F3$-based normalization, and SGR normalization. All WERs that are not significantly different ($p > 0.05$) from f_o normalization are in **bold**.

Feature	**Training Grade**									
Normalization	1	2	3	4	5	6	7	8	9	10
None	**23.58**	**22.49**	**25.03**	**25.03**	**26.42**	**30.17**	**33.43**	37.91	41.05	47.28
f_o	**25.70**	**24.55**	**25.76**	**26.96**	**26.06**	**29.63**	**32.59**	**34.28**	**34.76**	**37.76**
VTLN	**24.43**	**26.60**	**27.45**	**25.94**	**25.09**	**29.32**	**31.38**	**36.15**	38.27	41.72
$F3$	**26.90**	27.45	28.90	**26.48**	**26.48**	**30.53**	**30.59**	**31.80**	**34.95**	**37.36**
SGR	**25.63**	**23.10**	**28.05**	**26.48**	**25.70**	**28.05**	**29.99**	**32.41**	**34.82**	**38.75**

speech data.

5.2.3 Results and Discussion

The results of the mismatched-grade experiment are shown in Table 5.3. The first row shows the WERs of the baseline system with no normalization, followed by the WERs of the systems trained with f_o normalization, VTLN, F3-based normalization, and SGR normalization.

VTLN showed significant improvement over the baseline when training on 9th and 10th grade speech. However, f_o normalization provided a further significant improvement over VTLN for these grades. Additionally, f_o normalization provided a significant improvement over F3-based normalization when training on 2nd and 3rd grade speech. The f_o normalization did not significantly improve over SGR normalization and did not perform significantly worse either.

Notably, besides performance improvement, the f_o normalization method also has additional computational benefits over some of the other normalization techniques. When using VTLN in an ASR system, speech data must be passed through the system multiple times to compute the maximum likelihood (ML) warping factor. As f_o normalization also performed significantly better than VTLN in heavily mismatched systems, f_o normalization should be used in favor of VTLN for young child ASR.

While the f_o and SGR normalization methods had similar performance, f_o normalization may provide additional computational benefits over SGR normalization as well. The SGR computation requires a reliable estimation of the subglottal resonances using only a microphone signal. The SGR computation used in SGR normalization requires the computation of formants [GPY15, ALL13] which are unreliable to estimate, especially for children with high f_o values. Furthermore, the SGR normalization algorithm also requires a rough approximation of age, further complicating the technique. However, f_o normalization only requires f_o, which is a commonly computed parameter.

5.3 The Effect of f_o Normalization and Perturbation on Continuous Child ASR

5.3.1 Databases

This set of experiments examined the effects of f_o normalization and f_o perturbation on child ASR using several databases containing both adult and child speech. For adult speech data, the LibriSpeech ASR Corpus was used. All 960 hours of clean and noisy speech were used for training data.

The first child speech database used was the OGI Kids' Speech Corpus. All multi-word read speech utterances (both full sentences and phrases) were used, which consists of a total of 10,072 utterances from children in kindergarten to 5th grade. Approximately 70% of the utterances from each grade were used as adaptation data for a total of 7,051 utterances. The remaining utterances were used for testing.

The second child speech database used was the CMU Kids Corpus. Of the 5,180 utterances, 70% of the utterances were used for adaptation data for a total of 3,626 utterances. The remaining utterances were used for testing.

5.3.2 Experimental Setup

The features used were 13-dimensional MFCCs with a window size of 25 ms, shift of 10 ms, 23 filters, and a lifter coefficient of 22. The full bandwidth of 8 kHz was used for feature extraction. The features were extracted both with and without f_o normalization. Additionally, when extracting features for child speech adaptation data, features were extracted both with and without f_o perturbation. The f_o normalization and perturbation procedures are similar to those reported earlier. That is, when applying f_o normalization, $f_{o,def} = 100$ Hz and $f_{o,utt}$ was chosen to be the median f_o of the utterance. Otherwise, without f_o normalization, we simply set $f_{o,utt} = f_{o,def}$. Similarly, when applying f_o perturbation, we replace $f_{o,def} = 100$ Hz with $f_{o,def} \in \{58.52, 72.10, 85.93, 100.00, 114.32, 128.90, 143.74\}$ Hz and extract features for each $f_{o,def}$ value in the set, effectively multiplying the adaptation data by 7. Finally, we also extracted the child speech adaptation data features with vocal tract length perturbation (VTLP) for comparison with f_o perturbation. For a fair comparison, the VTLP was also set to multiply the adaptation data by 7 by setting the warping factors for the VTLP to be from the set of $\{0.94, 0.96, 0.98, 1.00, 1.02, 1.04, 1.06\}$.

The acoustic model for the ASR system was a bidirectional long-short term memory Hidden Markov model (BLSTM-HMM)-based acoustic model. Compared to the previous experiments, the BLSTM-HMM acoustic model was chosen due to its larger size and ability to model time information. The system was first trained using the LibriSpeech data, either with or without f_o normalization and with or without f_o perturbation. Triphone alignments were first extracted by training GMM-HMM ASR systems using the LibriSpeech tri6b recipe from the Kaldi ASR Toolkit. These alignments were used to train the BLSTM-HMM system using the PyKaldi2 toolbox [LXC19]. The input to the BLSTM was a 7-frame concatenation (3 frames left, 3 frames right) for a 91-dimensional feature input. The BLSTM had 3 layers with 512 cells in each direction followed by a feed-forward softmax layer that mapped the output to approximately 5,700 triphone probabilities.

Table 5.4: Word error rates (WERs) (%) of the continuous speech child ASR experiment using a BLSTM-based acoustic model adapted from adult speech. Features were extracted with an 8 kHz bandwidth. ASR systems either used no data augmentation, VTLP, or f_o perturbation. WERs for both the CMU Kids Corpus and OGI Kids' Speech Corpus are reported. The system using f_o perturbation performed the best on both datasets. The best performing system that performed significantly better than the baseline ($p < 0.05$) is in **bold**.

Augmentation	**CMU Kids**	**OGI Sent.**
None	16.88	6.84
VTLP	17.05	6.22
f_o Per.	16.63	**5.85**

After training the BLSTM-HMM system on adult speech, the ASR systems were adapted to child speech. Two separate systems were trained, one adapted using the OGI Kids' Speech Corpus while the other was adapted using the CMU Kids Corpus. All the parameters of the acoustic model trained on adult speech were used as an initialization for training the child model, and the same procedure for training was applied using the child speech data for parameter fine-tuning.

The OGI Kids' Speech Corpus was used to test the ASR system adapted using the OGI Kids' Speech Corpus, and the testing data from the CMU Kids Corpus was used to test the ASR system adapted using the CMU Kids Corpus. A 4-gram language model (LM) trained on approximately 14,500 Project Gutenberg books was used for decoding. This LM is one of the LMs included in Kaldi's LibriSpeech recipe [PCP15].

5.3.3 Results and Discussion

The results of the system trained using f_o perturbation for data augmentation, along with the system using VTLP for comparison to another standard data augmentation technique, are shown in Table 5.4. When using the OGI Kids' Speech Corpus as adaptation data, the use of f_o perturbation results in a significant improvement from 6.84% to 5.85% ($p < 0.001$). However, the system using the CMU Kids Corpus as adaptation data shows a much smaller

Table 5.5: Word error rates (WERs) (%) of the continuous speech child ASR experiment using a BLSTM-based acoustic model adapted from adult speech. Features were extracted with an 8 kHz bandwidth. The left two columns indicate whether f_o normalization ("f_o Norm?") and data augmentation using f_o perturbation ("f_o Per?") were used. WERs for both the CMU Kids Corpus and OGI Kids' Speech Corpus are reported in the latter columns. The system using both f_o normalization and f_o perturbation performed the best on both datasets. The best performing system that performed significantly better than the baseline ($p < 0.05$) is in **bold**.

f_o **Norm?**	f_o **Per?**	**CMU Kids**	**OGI Sent.**
No	No	16.88	6.84
Yes	No	16.93	6.50
No	Yes	16.63	5.85
Yes	Yes	16.47	**5.52**

improvement. When applying VTLP, the OGI Kids' Speech Corpus system also shows an improvement over the baseline, but the improvement is smaller than when using f_o perturbation. Furthermore, the use of VTLP results in worse performance than the baseline for the CMU Kids Corpus system. These results suggest that f_o perturbation is superior to VTLP as a data augmentation technique. This is likely due to the fact that VTLP uses a simple linear warping function while f_o perturbation uses a non-linear warping function based on speech perception.

The results of the systems trained both with and without f_o normalization and f_o perturbation are shown in Table 5.5. The left two columns indicate whether f_o normalization or f_o perturbation was used. The right two columns display the WERs of the systems trained and tested on either the OGI Kids' Speech Corpus or the CMU Kids Corpus.

Applying f_o normalization to the child ASR system results in a slight improvement for the OGI Kids' Speech Corpus system. Replacing f_o normalization with f_o perturbation results in a more substantial improvement as examined previously. Applying both f_o normalization and f_o perturbation together provides a further improvement to 5.52%, a relative improvement of 19.3% over the baseline. Applying both f_o normalization and f_o perturbation on the CMU Kids Speech system results in a relative improvement of 2.4% over the baseline with the

Table 5.6: Word error rates (WERs) (%) of the continuous speech child ASR experiment using a BLSTM-based acoustic model adapted from adult speech. Features were extracted with an 8 kHz bandwidth. The left two columns indicate whether f_o normalization ("f_o Norm?") and data augmentation using f_o perturbation ("f_o Per?") were used. WERs on the OGI Kids' Speech Corpus, separated by testing grade, are reported in the latter columns. The system using both f_o normalization and f_o perturbation performed the best for all grades. The best performing systems that performed significantly better than the baseline ($p < 0.05$) are in **bold**.

		Testing Grade					
f_o **Norm?**	f_o **Per?**	K	1	2	3	4	5
No	No	16.97	9.17	6.73	5.71	4.15	4.99
Yes	No	17.44	9.17	6.19	4.80	3.47	4.93
No	Yes	13.97	7.89	5.27	5.11	3.72	4.35
Yes	Yes	**12.87**	**7.38**	**4.88**	4.78	3.35	4.24

WER decreasing to 16.47%. Both systems result in improvements over recently reported child ASR systems such as [WGP19], which reported WERs of 10.8% when using the OGI Kids' Speech Corpus and 17.3% when using the CMU Kids Corpus.

The difference in improvement between the two systems is likely due to the differences in age range between the two child speech databases. The OGI Kids' dataset contained children from kindergarten to 5th grade, while the CMU Kids Corpus only contained children from 1st to 3rd grade. As shown in the previous section, f_o normalization is more effective when the age range between training and testing speakers is large. This may also be true of f_o perturbation. That is, since there is less variability in the CMU Kids Corpus, additional variability in the training data through f_o perturbation is unnecessary and unhelpful when training the ASR system. On the other hand, the OGI Kids' Speech Corpus has a larger age range and includes kindergarten speech, which is more difficult to recognize, thus resulting in ASR performance improvements when applying both f_o normalization and f_o perturbation.

To further examine the effectiveness of these techniques on the OGI Kids' Speech Corpus ASR system, the testing data were separated by grade, as shown in Table 5.6. Similar to the results in Table 5.5, the use of both f_o normalization and f_o perturbation resulted in the best performance for all grades. These systems performed significantly better than the baseline for 2nd grade and younger at $p < 0.05$. Additionally, these systems performed significantly better for 3rd and 4th grade at $p < 0.1$. Thus, both f_o normalization and f_o perturbation provide improvements over the full age range of the OGI Kids' Speech Corpus.

5.4 Chapter Summary

In this chapter, we performed several child ASR experiments using various frequency normalization and data augmentation techniques. The first several experiments examined the effect of educational grade level (K-10th grade) on single word ASR performance. When the training and testing data were matched, WER was dramatically higher for the younger grades. When testing on kindergarten speech, WER increased significantly when the training data were from older children. Additionally, when applying normalization, f_o normalization performed equal to or better than all other normalization techniques.

The next experiments examined the use of f_o normalization and f_o perturbation on continuous speech child ASR. When the age range of the system was large, f_o normalization and f_o perturbation were more effective than VTLN and VTLP, respectively. Additionally, the system that used both f_o normalization and f_o perturbation performed best in all cases and grades. The next chapter will summarize the dissertation and discuss the applications of this research for educational technology and future work.

CHAPTER 6

Conclusion

6.1 Summary

This dissertation examines aspects of child speech, specifically child subglottal resonances (SGRs) and fundamental frequency, and their use in novel frequency normalization and data augmentation methods for child ASR. In Chapter 2, several databases and software that were used as part of this dissertation were introduced and explained. In particular, The Child Subglottal Resonances Database, containing approximately 15.5 hours of simultaneous microphone and subglottal accelerometer recordings of children, was published as part of this work. Additionally, The GFTA-JIBO Kids Corpus, a database consisting of microphone recordings of children from pre-kindergarten through 2nd grade playing educational games with the social robot JIBO, was introduced. Other publicly available databases used in this dissertation include the LibriSpeech ASR Corpus, The OGI Kids' Speech Corpus, and the CMU Kids Corpus. Speech software tools used in this dissertation include Praat, WaveSurfer, VOICEBOX, the MBSC pitch detection algorithm, and the Kaldi ASR Toolkit.

In Chapter 3, the existing quarter-wavelength resonator SGR model was examined using child speech. This examination revealed that this model, which was developed using adult data, did not model child SGRs accurately due to the variable penetration depth of higher frequencies in the subglottal system. An additional logistic function was used to more accurately model the increasing penetration depth of the higher frequency SGRs, which is particularly relevant for young children. Additionally, SGRs were demonstrated to be effective at normalizing the vowel space, with vowels effectively separated by SGR_1 and SGR_2 at least 81.7% of the time.

In Chapter 4, the tonotopic distance model of vowel perception was introduced. This model was reformulated and used as the basis for f_o normalization, a frequency normalization technique for ASR feature extraction that only requires the median f_o of a speaker or utterance as a parameter. The f_o normalization function was reformulated into f_o perturbation, a data augmentation technique for ASR feature extraction capable of generating additional speech-like features for training data.

In Chapter 5, several experiments were performed evaluating child ASR, as well as the proposed normalization and augmentation techniques. When evaluating the state-of-the-art ASR systems in a single-word child ASR experiment, the ASR systems using young child speech performed significantly worse than the systems using older child speech, with kindergarten speech achieving a WER of only 26.91%, while 10th grade speech achieved a WER of less than 4%. Furthermore, when ASR systems were mismatched by age, kindergarten speech performed even worse, with a 34.95% WER using systems trained on older child speech.

To improve the performance of these systems, SGR normalization and f_o normalization were evaluated against VTLN and $F3$-based normalization in a single-word child speech ASR experiment. Both SGR normalization and f_o normalization performed equal to or significantly better than VTLN and the baseline regardless of the grade of the training data speakers. The f_o normalization and f_o perturbation techniques were also evaluated on a continuous speech child ASR experiment. f_o perturbation was also compared against VTLP and performed significantly better, with the combination of f_o normalization and f_o perturbation performing the best out of all systems tested.

6.2 Educational Applications

This work makes several contributions to the fields of speech science, early childhood education, and educational technology. These contributions include the data collection, methodology of the data collection, and the experimental results.

Both The Child Subglottal Resonances Database and The GFTA-JIBO Kids Corpus that were collected and published as part of this work (see Chapter 2), provide new child speech data from children, as young as 4 years old and as old as 16 years old, speaking in various contexts. The data in both databases were collected using specific educational and clinical tasks. The Child Subglottal Resonances Database contains pronunciations of consonant-vowel-consonant (CVC) words, which are useful for evaluating child pronunciations and common errors across various ages. It also provides the subglottal signals of these pronunciations, which will allow future research on the subglottal system. Similarly, The GFTA-JIBO Kids Corpus contains recordings of children taking the 3rd Goldman Fristoe Test of Articulation (GFTA-3), a common clinical assessment for speech disorder diagnosis. Furthermore, the nature of these databases can be used to target improvements in educational speech technology and specific age groups rather than using more general child speech databases.

Additionally, for the GFTA-JIBO Kids Corpus, data collection methods, best practices, and setups were documented in Chapter 2. As this database recorded children interacting with a robotic learning companion, an interaction not very common in speech databases, these insights may prove to be valuable for future data collection of child-robot interactions. For instance, the positioning of the robot and microphone, effective methods to maintain child engagement, and effective question-and-answer procedures using the robot were noted, which can help future researchers recreate our setup and produce similar high-quality data.

Finally, the results of the experiments in Chapter 5 outline effective methods to train child ASR systems that take advantage of the acoustic properties of child speech. While child

ASR systems still require a number of improvements to be effective for general usage, the use of normalization and data augmentation methods can likely provide further performance improvement for task specific ASR systems such as pronunciation assessments and learning exercises. This is a logical direction for the applied usage of child ASR systems while further research is done to understand child speech in a more general setting.

6.3 Future Work

While the proposed normalization and data augmentation techniques provide significant improvements to child ASR systems, the resulting ASR systems still do not come close to the ASR performance of adult speech. In fact, it is doubtful that without significant training data, child ASR will ever reach the performance level of adult ASR. As such, there are still many additional research directions to explore.

Of particular importance is a universal ASR system for both adults and children. This study only adapts an adult ASR system to children, but this is known to degrade performance on adult speech. It is important to acknowledge many applications of ASR are used by both children and adults. Whether warping can effectively reduce feature variability in ASR systems designed for simultaneous use by children and adults remains to be investigated.

It is worth noting that child ASR systems are often used in adverse and noisy environments such as grade-school classrooms or playgrounds. To ensure the effectiveness of child ASR in multiple applications, it is essential to consider how these techniques and others are affected by the acoustic environment. This may require new robust ASR techniques.

REFERENCES

[ALL13] H. Arsikere, G. K. F. Leung, S. M. Lulich, and A. Alwan. "Automatic estimation of the first three subglottal resonances from adults' speech signals with application to speaker height estimation." *Speech Communication*, **55**(1):51–70, 2013.

[AN07] P. F. Assmann and T. M. Nearey. "Relationship Between Fundamental and Formant Frequencies in Voice Preference." *The Journal of the Acoustical Society of America*, **122**(2):EL35–EL43, 2007.

[BBD05] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russel, S. Steidl, and M. Wong. "The PF-STAR Children's Speech Corpus." In *Proc. of INTERSPEECH*, pp. 2761–2764, 2005.

[Ben90] A. H. Benade. *Fundamentals of Music Acoustics.* Dover, Mineola, NY, 1990.

[BN12] S. Barreda and T. M. Nearey. "The Direct and Indirect Roles of Fundamental Frequency in Vowel Perception." *The Journal of the Acoustical Society of America*, **131**(1):466–477, 2012.

[BN13] S. Barreda and T. M. Nearey. "The Perception of Formant-Frequency Range is Affected by Veridical and Judged Fundamental Frequency." In *Proc. of Meetings on Acoustics*, volume 19, p. 060197, 2013.

[Bro06] M. Brookes. "Voicebox: A Speech Processing Toolbox for MATLAB.", 2006.

[BW17] P. Boersma and D. Weenink. "Praat: doing phonetics by computer.", 2017.

[BYP00] H. T. Bunnell, D. M. Yarrington, and J. B. Polikoff. "STAR: Articulation Training for Young Children." In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pp. 85–88, 2000.

[CA05] X. Cui and A. Alwan. "MLLR-Like Speaker Adaptation Based on Linearization of VTLN with MFCC Features." In *Proc. of INTERSPEECH*, pp. 273–276, 2005.

[CA06] X. Cui and A. Alwan. "Adaptation of Children's Speech with Limited Data Based on Formant-Like Peak Alignment." *Computer Speech and Language*, **20**(4):400–419, 2006.

[CB87] B. Cranen and L. Boves. "On subglottal formant analysis." *The Journal of the Acoustical Society of America*, **81**(3):734–746, 1987.

[CBG09] T. G. Csapó, Z. Bárkányi, T. E. Gráczi, T. Bohm, and S. M. Lulich. "Relation of formants and subglottal resonances in Hungarian vowels." In *Proc. of INTERSPEECH*, pp. 484–487, 2009.

[CGK14] X. Cui, V. Goel, and B. Kingsbury. "Data augmentation for deep neural network acoustic modeling." In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5582–5586, 2014.

[Chi85] L. A. Chistovich. "Central auditory processing of peripheral vowel spectra." *The Journal of the Acoustical Society of America*, **77**(3):789–805, 1985.

[CL79] L. A. Chistovich and V. V. Lublinskaya. "The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli." *Hearing Research*, **1**:185–195, 1979.

[CS07] X. Chi and M. Sonderegger. "Subglottal coupling and its influence on vowel formants." *The Journal of the Acoustical Society of America*, **122**(3):1735–1745, 2007.

[DLM11] G. Dogil, S. M. Lulich, A. Madsack, and W. Wokurek. "Crossing the Quantal Boundaries of Features: Subglottal Resonances and Swabian Diphthongs." In *Tones and Features: Phonetic and Phonological Perspectives*, pp. 137–148. De Gruyter Mouton, The Hague, Netherlands, 2011.

[EMG97] M. Eskenazi, J. Mostow, and D. Graff. "The CMU Kids Speech Corpus LDC97S63.", 1997.

[ESW97] J. Epps, J. R. Smith, and J. Wolfe. "A novel instrument to measure acoustic resonances of the vocal tract during phonation." *Measurement Science and Technology*, **8**:1112–1121, 1997.

[Fan60] G. Fant. *Acoustic Theory of Speech Production.* Mouton, The Hague, Netherlands, 1960.

[FBL16] J. Fainberg, P. Bell, M. Lincoln, and S. Renals. "Improving children's speech recognition through out-of-domain data augmentation." In *Proc. of INTERSPEECH*, pp. 1598–1602, 2016.

[FDT96] R. P. Fahey, R. L. Diehl, and H. Traunmüller. "Perception of Back Vowels: Effects of Varying F1-F0 Bark Distance." *The Journal of the Acoustical Society of America*, **99**(4):2350–2357, 1996.

[FG05] A. Faria and D. Gelbart. "Efficient Pitch-Based Estimation of VTLN Warp Factors." In *Proc. of INTERSPEECH*, pp. 213–216, 2005.

[FIL72] G. Fant, K. Ishizaka, J. Lindqvist, and J. Sundberg. "Subglottal Formants." *KTH Speech Transmission Laboratory Quarterly Progress and Status Report 1*, pp. 1–12, 1972.

[FM78] J. J. Fredberg and J. A. Moore. "The distributed response of complex branching duct networks." *The Journal of the Acoustical Society of America*, **63**(3):954–961, 1978.

[FNS01] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama. "Multiple-Regression Hidden Markov Model." In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 513–516, 2001.

[GGN09] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos. "A Review of ASR Technologies for Children's Speech." In *Proc. of the Workshop on Child, Computer and Interaction (WOCCI)*, pp. 7.1–7.8, 2009.

[GLC11] T. E. Gráczi, S. M. Lulich, T. G. Csapó, and A. Beke. "Context and speaker dependency in the relation of vowel formants and subglottal resonances – Evidence from Hungarian." In *Proc. of INTERSPEECH*, pp. 1901–1904, 2011.

[GPY15] J. Guo, R. Paturi, G. Yeung, S. M. Lulich, H. Arsikere, and A. Alwan. "Age-Dependent Height Estimation and Speaker Normalization for Children's Speech Using the First Three Subglottal Resonances." In *Proc. of INTERSPEECH*, pp. 1665–1669, 2015.

[GWL14] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner, and N. Bodenstab. "Child Automatic Speech Recognition for US English: Child Interaction with Living-Room-Electronic-Devices." In *Proc. of the Workshop on Child Computer Interaction (WOCCI)*, pp. 21–26, 2014.

[GYM16] J. Guo, G. Yeung, D. Muralidharan, H. Arsikere, A. Afshan, and A. Alwan. "Speaker verification using short utterances with DNN-based estimation of subglottal acoustic features." In *Proc. of INTERSPEECH*, pp. 2219–2222, 2016.

[HCC14] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. "Deep Speech: Scaling Up End-to-end Speech Recognition." 2014.

[IMK76] K. Ishizaka, M. Matsudaira, and T. Kaneko. "Input acoustic-impedance measurement of the subglottal system." *The Journal of the Acoustical Society of America*, **60**(1):190–197, 1976.

[JH13] N. Jaitly and G. E. Hinton. "Vocal tract length perturbation (VTLP) improves speech recognition." In *Proc. of the International Conference on Machine Learning (ICML)*, 2013.

[Jun09] Y. Jung. *Acoustic articulatory evidence for quantal vowel categories: The features [low] and [back].* PhD thesis, 2009.

[KK90] D. H. Klatt and L. C. Klatt. "Analysis, synthesis, and perception of voice quality variations among female and male talkers." *Journal of the Acoustical Society of America*, **87**(2):820–857, 1990.

[KL08] L. L. Koenig and J. C. Lucero. "Stop Consonant Voicing and Intraoral Pressure Contours in Women and Children." *The Journal of the Acoustical Society of America*, **123**(2):1077–1088, 2008.

[KLM17] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme. "Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations." In *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 82–90, 2017.

[KLP08] L. L. Koenig, J. Cc Lucero, and E. Perlman. "Speech Production Variability in Fricatives of Children and Adults: Results of Functional Data Analysis." *The Journal of the Acoustical Society of America*, **124**(5):3158–3170, 2008.

[KPP17] T. Ko, V. Peddinti, D. Povel, and S. Khudanpur. "Audio Augmentation for Speech Recognition." In *Proc. of INTERSPEECH*, pp. 3586–3589, 2017.

[KWE91] D. Kewley-Port, C. S. Watson, M. Elbert, D. Maki, and D. Reed. "The Indiana Speech Training Aid (ISTRA) II: Training Curriculum and Selected Case Studies." *Clinical Linguistics and Phonetics*, **5**(1):13–38, 1991.

[LAA11] S. M. Lulich, A. Alwan, H. Arsikere, J. R. Morton, and M. S. Sommers. "Resonances and wave propagation velocity in the subglottal airways." *The Journal of the Acoustical Society of America*, **130**(4):2108–2115, 2011.

[LAM11] S. M. Lulich, H. Arsikere, J. R. Morton, G. K. F. Leung, A. Alwan, and M. S. Sommers. "Analysis and automatic estimation of children's subglottal resonances." In *Proc. of INTERSPEECH*, pp. 2817–2820, 2011.

[LBM07] S. M. Lulich, A. Bachrach, and N. Malyska. "A role for the second subglottal resonance in lexical access." *The Journal of the Acoustical Society of America*, **122**(4):2320–2327, 2007.

[LD93] R. G. Leonard and G. Doddington. *TIDIGITS LDC93S10.* Linguistic Data Consortium, Philadelphia, PA, 1993.

[Ljo02] A. Ljolje. "Speech Recognition Using Fundamental Frequency and Voicing in Acoustic Modeling." In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pp. 2137–2140, 2002.

[LMA12] S. M. Lulich, J. R. Morton, H. Arsikere, M. S. Sommers, G. K. F. Leung, and A. Alwan. "Subglottal resonances of adult male and female native speakers of American English." *The Journal of the Acoustical Society of America*, **132**(4):2592–2602, 2012.

[LNB14] J. Lilley, S. Nittrouer, and H. T. Bunnell. "Automating an Objective Measure of Pediatric Speech Intelligibility." In *Proc. of INTERSPEECH*, pp. 1578–1582, 2014.

[LPN97] S. Lee, A. Potamianos, and S. Narayanan. "Analysis of Children's Speech: Duration, Pitch and Formants." In *Proc. of EUROSPEECH*, pp. 473–476, 1997.

[LPN99] S. Lee, A. Potamianos, and S. Narayanan. "Acoustics of Children's Speech: Developmental Changes of Temporal and Spectral Parameters." *The Journal of the Acoustical Society of America*, **105**(3):1455–1468, 1999.

[LR98] L. Lee and R. Rose. "A Frequency Warping Approach to Speaker Normalization." *IEEE Transactions on Speech and Audio Processing*, **6**(1):49–60, 1998.

[Lul10] S. M. Lulich. "Subglottal resonances and distinctive features." *Journal of Phonetics*, **38**(1):20–32, 2010.

[Lul13] S. M. Lulich. "Estimation of lumped vocal fold mechanical properties from non-invasive microphone recordings." In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 8041–8045, 2013.

[LXC19] L. Lu, X. Xiao, Z. Chen, and Y. Gong. "PyKaldi2: Yet another speech toolkit based on Kaldi and PyTorch." 2019.

[LZM09] S. M. Lulich, M. Zañartu, D. D. Mehta, and R. E. Hillman. "Source-filter interaction in the opposite direction: Subglottal coupling and the influence of vocal fold mechanics on vowel spectra during the closed phase." In *Proc. of Meetings on Acoustics*, volume 6, p. 060007, 2009.

[Mer76] P. Mermelstein. "Distance Measures for Speech Recognition, Psychological and Instrumental." *Pattern Recognition and Artificial Intelligence*, **116**:374–388, 1976.

[MLW08] A. Madsack, S. M. Lulich, W. Wokurek, and G. Dogil. "Subglottal resonances and vowel formant variability: A case study of high German monophthongs and Swabian diphthongs." In *Proc. of Laboratory Phonology 11*, pp. 91–92, 2008.

[MSB03] M. Magimai-Doss, T. A. Stephenson, and H. Bourlard. "Using Pitch Frequency Information in Speech Recognition." In *Proc. of EUROSPEECH*, pp. 2525–2528, 2003.

[OS87] D. O'Shaughnessy. *Speech Communication: Human and Machine.* Addison-Wesley, Boston, MA, 1987.

[PA06] S. Panchapagesan and A. Alwan. "Multi-Parameter Frequency Warping for VTLN by Gradient Search." In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1181–1184, 2006.

[PB52] G. E. Peterson and H. L. Barney. "Control Methods Used in a Study of the Vowels." *The Journal of the Acoustical Society of America*, **24**(2):175 – 184, 1952.

[PCP15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. "Librispeech: An ASR corpus based on public domain audio books." In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[PCZ19] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. "SpecAugment: A simple data augmentation method for automatic speech recognition." In *Proc. of INTERSPEECH*, pp. 2613–2617, 2019.

[PGB11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý. "The Kaldi Speech Recognition Toolkit." In *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[Pyl75] R. W. Pyle. "Effective Length of Horns." *The Journal of the Acoustical Society of America*, **57**(6):1309–1317, 1975.

[RBS96] M. Russell, C. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker. "Applications of automatic speech recognition to speech and language development in young children." In *Proc. of the Fourth International Conference on Spoken Language Processing (ICSLP)*, pp. 176–179, 1996.

[RS11] L. R. Rabiner and R. W. Schafer. *Theory and Application of Digital Speech Processing.* Pearson Higher Education, Upper Saddle, River, NJ, 2011.

[SB00] K. Sjölander and J. Beskow. "Wavesurfer – an open source speech tool." In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, volume 4, pp. 464–467, 2000.

[SBM87] J. A. Sereno, S. R. Baum, G. C. Marean, and P. Lieberman. "Acoustic Analyses and Perceptual Data on Anticipatory Labial Coarticulation in Adults and Children." *The Journal of the Acoustical Society of America*, **81**(2):512–519, 1987.

[SCA18] S. Spaulding, H. Chen, S. Ali, M. Kulinski, and C. Breazeal. "A Social Robot System for Modeling Children's Word Pronunciation." In *Proc. of the International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 1658–1666, 2018.

[SDS16] S. Shahnawazuddin, A. Dey, and R. Sinha. "Pitch-Adaptive Front-End Features for Robust Children's ASR." In *Proc. of INTERSPEECH*, pp. 3459–3463, 2016.

[SG86] A. K. Syrdal and H. S. Gopal. "A Perceptual Model of Vowel Recognition Based on the Auditory Representation of American English Vowels." *The Journal of the Acoustical Society of America*, **79**(4):1086–1100, 1986.

[SG14] R. Serizel and D. Giuliani. "Vocal Tract Length Normalization Approaches to DNN-Based Children's and Adults' Speech Recognition." In *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, pp. 135–140, 2014.

[SHC00] K. Shobaki, J.-P. Hosom, and R. A. Cole. "The OGI Kids' Speech Corpus and Recognizers." In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pp. 258–261, 2000.

[SHS03] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth. "Acoustic Normalization of Children's Speech." In *Proc. of EUROSPEECH*, pp. 1313–1316, 2003.

[Sjo97] K. Sjölander. "The Snack Sound Toolkit.", 1997.

[SLR09] O. Saz, E. Lleida, and W.-R. Rodríguez. "Avoiding Speaker Variability in Pronunciation Verification of Children's Disordered Speech." In *Proc. of the 2nd Workshop on Child, Computer and Interaction (WOCCI)*, pp. 11.1–11.5, 2009.

[Smi92] B. L. Smith. "Relationships Between Duration and Temporal Variability in Children's Speech." *The Journal of the Acoustical Society of America*, **91**(4):2165–2174, 1992.

[SPL14] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan. "Improving Speech Recognition for Children Using Acoustic Adaptation and Pronunciation Modeling." In *Proc. of the Workshop on Child Computer Interaction (WOCCI)*, pp. 15–19, 2014.

[Ste98] K. N. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, MA, 1998.

[SZ15] R. Sadeghian and S. A. Zahorian. "Towards an Automated Screening Tool for Pediatric Speech Delay." In *Proc. of INTERSPEECH*, pp. 1650–1654, 2015.

[TA13] L. N. Tan and A. Alwan. "Multi-Band Summary Correlogram-Based Pitch Detection for Noisy Speech." *Speech Communication*, **55**(7-8):841–856, 2013.

[Tit06] I. R. Titze. *The Myoelastic Aerodynamic Theory of Phonation*. National Center for Voice and Speech, Denver, CO, 2006.

[Tit08] I. R. Titze. "Nonlinear source–filter coupling in phonation: Theory." *The Journal of the Acoustical Society of America*, **123**(5):2733–2749, 2008.

[Tra81] H. Traunmüller. "Perceptual Dimension of Openness in Vowels." *The Journal of the Acoustical Society of America*, **69**(5):1465–1475, 1981.

[TSK06] J. Tepperman, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan. “Pronunciation Verification of Children’s Speech for Automatic Literacy Assessment.” In *Proc. of INTERSPEECH*, pp. 845–848, 2006.

[VK07] H. K. Vorperian and R. D. Kent. “Vowel Acoustic Space Development in Children: A Synthesis of Acoustic and Anatomic Data.” *Journal of Speech, Language, and Hearing Research*, **50**(6):1510–1545, 2007.

[WGP19] F. Wu, L. P. Garcia, D. Povey, and S. Khudanpur. “Advances in automatic speech recognition for child speech using factored time delay neural network.” In *Proc. of INTERSPEECH*, pp. 1–5, 2019.

[WLA09a] S. Wang, Y.-H. Lee, and A. Alwan. “Bark-shift based nonlinear speaker normalization using the second subglottal resonance.” In *Proc. of INTERSPEECH*, pp. 1619–1622, 2009.

[WLA09b] S. Wang, S. M. Lulich, and A. Alwan. “Automatic detection of the second subglottal resonance and its application to speaker normalization.” *The Journal of the Acoustical Society of America*, **126**(6):3268–3277, 2009.

[YA18] G. Yeung and A. Alwan. “On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children.” In *Proc. of INTERSPEECH*, pp. 1661–1665, 2018.

[YAO17] G. Yeung, A. Afshan, K. E. Ozgun, K. Kaewtip, S. M. Lulich, and A. Alwan. “Predicting Clinical Evaluations of Children’s Speech with Limited Data Using Exemplar Word Template References.” In *Proc. of the Workshop on Speech and Language Technology in Education (SLaTE)*, pp. 161–166, 2017.

[YBA19a] G. Yeung, A. L. Bailey, A. Afshan, M. Q. Pérez, A. Martin, S. Spaulding, H. W. Park, A. Alwan, and C. Breazeal. “Towards the Development of Personalized Learning Companion Robots for Early Speech and Language Assessment.” In *Proc. of the Annual Meeting of the American Educational Research Association (AERA)*, 2019.

[YBA19b] G. Yeung, A. L. Bailey, A. Afshan, M. Tinkler, M. Q. Pérez, A. Martin, A. A. Pogossian, S. Spaulding, H. W. Park, M. Muco, A. Alwan, and C. Breazeal. “A Robotic Interface for the Administration of Language, Literacy, and Speech Pathology Assessments for Children.” In *Proc. of the Workshop on Speech and Language Technology in Education (SLaTE)*, pp. 41–42, 2019.

[YD15] D. Yu and L. Deng. *Automatic Speech Recognition: A Deep Learning Approach.* Springer, New York, NY, 2015.

[YLG18] G. Yeung, S. M. Lulich, J. Guo, M. S. Sommers, and A. Alwan. "Subglottal Resonances of American English Speaking Children." *The Journal of the Acoustical Society of America*, **144**(6):3437–3449, 2018.

[ZHG15] M. Zharkova, W. J. Hardcastle, F. E. Gibbon, and R. J. Lickley. "Development of Lingual Motor Control in Children and Adolescents." In *Proc. of the International Congress of Phonetic Sciences (ICPhS)*, 2015.

[ZMH11] M. Zañartu, D. D. Mehta, J. C. Ho, G. R. Wodicka, and R. E. Hillman. "Observation and analysis of in vivo vocal fold tissue instabilities produced by nonlinear source-filter coupling: A case study." *The Journal of the Acoustical Society of America*, **129**(1):326–339, 2011.

[ZNB06] Z. Zhang, J. Neubauer, and D. A. Berry. "The Influence of Subglottal Acoustics on Laboratory Models of Phonation." *The Journal of the Acoustical Society of America*, **120**(3):1558–1569, 2006.