ABSTRACT

Title of dissertation: ARTICULATORY REPRESENTATIONS TO ADDRESS ACOUSTIC VARIABILITY IN SPEECH

Ganesh Sivaraman, Doctor of Philosophy, 2017

Dissertation directed by: Professor Carol Espy-Wilson
Department of Electrical and Computer Engineering

The past decade has seen phenomenal improvement in the performance of Automatic Speech Recognition (ASR) systems. In spite of this vast improvement in performance, the state-of-the-art still lags significantly behind human speech recognition. Even though certain systems claim super-human performance, this performance often is sub-par across domains and across datasets. This gap is predominantly due to the lack of robustness against speech variability. Even clean speech is extremely variable due to a large number of factors such as voice characteristics, speaking style, speaking rate, accents, casualness, emotions and more. The goal of this thesis is to investigate the variability of speech from the perspective of speech production, put forth robust articulatory features to address this variability, and to incorporate these features in state-of-the-art ASR systems in the best way possible. ASR systems model speech as a sequence of distinctive phone units like beads on a string. Although phonemes are distinctive units in the cognitive domain, their physical realizations are extremely varied due to coarticulation and

lenition which are commonly observed in conversational speech. The traditional approaches deal with this issue by performing di-, tri- or quin-phone based acoustic modeling but are insufficient to model longer contextual dependencies. Articulatory phonology analyzes speech as a constellation of coordinated articulatory gestures performed by the articulators in the vocal tract (lips, tongue tip, tongue body, jaw, glottis and velum). In this framework, acoustic variability is explained by the temporal overlap of gestures and their reduction in space. In order to analyze speech in terms of articulatory gestures, the gestures need to be estimated from the speech signal. The first part of the thesis focuses on a speaker independent acoustic-to-articulatory inversion system that was developed to estimate vocal tract constriction variables (TVs) from speech. The mapping from acoustics to TVs was learned from the multi-speaker X-ray Microbeam (XRMB) articulatory dataset. Constriction regions from TV trajectories were defined as articulatory gestures using articulatory kinematics. The speech inversion system combined with the TV kinematics based gesture annotation provided a system to estimate articulatory gestures from speech. The second part of this thesis deals with the analysis of the articulatory trajectories under different types of variability such as multiple speakers, speaking rate, and accents. It was observed that speaker variation degraded the performance of the speech inversion system. A Vocal Tract Length Normalization (VTLN) based speaker normalization technique was therefore developed to address the speaker variability in the acoustic and articulatory domains. The performance of speech inversion systems was analyzed on an articulatory dataset containing speaking rate variations to assess if the model

was able to reliably predict the TVs in challenging coarticulatory scenarios. The performance of the speech inversion system was analyzed in cross accent and cross language scenarios through experiments on a Dutch and British English articulatory dataset. These experiments provide a quantitative measure of the robustness of the speech inversion systems to different speech variability. The final part of this thesis deals with the incorporation of articulatory features in state-of-the-art medium vocabulary ASR systems. A hybrid convolutional neural network (CNN) architecture was developed to fuse the acoustic and articulatory feature streams in an ASR system. ASR experiments were performed on the Wall Street Journal (WSJ) corpus. Several articulatory feature combinations were explored to determine the best feature combination. Cross-corpus evaluations were carried out to evaluate the WSJ trained ASR system on the TIMIT and another dataset containing speaking rate variability. Results showed that combining articulatory features with acoustic features through the hybrid CNN improved the performance of the ASR system in matched and mismatched evaluation conditions. The findings based on this dissertation indicate that articulatory representations extracted from acoustics can be used to address acoustic variability in speech observed due to speakers, accents, and speaking rates and further be used to improve the performance of Automatic Speech Recognition systems.

ARTICULATORY REPRESENTATIONS TO ADDRESS ACOUSTIC VARIABILITY IN SPEECH

by

Ganesh Sivaraman

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Carol Espy-Wilson, Chair/Advisor
Professor Shihab Shamma,
Professor Behtash Babadi
Professor Jonathan Z. Simon
Dr. Vikramjit Mitra
Professor William J. Idsardi

Dedication

करोमि यद्यत् सकलम् परस्मै नारायणायेति समर्पयामि ||

Whatever work I do I offer it all to Bhagawan Narayana

To

Dear Thatha Shri V. Venkataraman - the man who planted and nurtured the seeds of curiosity and scientific inquiry in the mind of a young boy;

& Paati Smt. V. Jayalakshmi who has a special place in her heart for me.

In fond memory of Thatha Shri T.G. Padmanabhan who could not be with us to witness this occasion.

To Paati Smt. Seethalakshmi who always showers me with love and blessings.

Acknowledgments

I would like to express my sincere gratitude to all the people who have been part of my PhD journey and helped me realize this dissertation.

First of all, I would like to thank my advisor Prof. Carol Espy-Wilson for believing in me since day one and motivating me to gain more knowledge and explore new ideas for my research. I have learned a lot about speech science and signal processing from her which will guide me for the rest of my career. Her intuition and insights have been my guideposts throughout my dissertation. Whenever I felt stuck or frustrated I have always found some way forward after talking to her. I have learned a great deal from her about perfection in presentation, writing, and attention to detail. I am thankful to her for always encouraging me to seek fellowship opportunities, new collaborations, and teaching experiences. I have thoroughly enjoyed research and teaching with Carol over the past six years.

Next, I would like to thank Dr. Vikramjit Mitra who has been a great mentor throughout my thesis. Vikram was instrumental in helping me make the decision to pursue my PhD as an extension of his dissertation work. I am thankful to him for always encouraging me to learn the latest techniques and tools for speech recognition and deep learning. It would have been very difficult for me to perform some of my core experiments if not for his help. I am also deeply grateful to him for giving me the opportunity to intern at SRI which helped me tremendously in taking my thesis to a whole new level. I thank Vikram for being available to discuss ideas and help me with experiments in spite of his busy schedule. I admire his creativity in

research and his ability to perform thorough experimentation. I have enjoyed my collaboration with him over the years and hope to work with him in the future.

I thank Prof. Shihab Shamma, Prof. Behtash Babadi, Prof. Jonathan Simon, Prof. William J. Idsardi and Dr. Vikramjit Mitra for being in my dissertation committee and providing insightful comments and suggestions to improve this thesis.

Next, I would like to express my gratitude for my collaborators Dr. Mark Tiede and Dr. Hosung Nam for teaching me concepts of phonology and providing tools and articulatory datasets needed for my research. I thank Dr. Martijn Wieling for giving me the opportunity to intern with his group at University of Groningen in the Netherlands. I thoroughly enjoyed my interactions with Dr. Wieling and look back fondly on my time in the Netherlands. I hope to continue this collaboration in the future.

The Speech Technology and Research Laboratory at SRI International has been instrumental in shaping me as a researcher. I thank Dr. Horacio Franco, Dr. Dimitra Vergyri, and Dr. Andreas Kathol for providing me with valuable insights and support during the two times I was an intern at SRI. I thank the lab for providing me access to their compute clusters which helped me perform my speech recognition experiments.

As a graduate student, one realizes that funding is a very crucial aspect of academic research. I thank the National Science Foundation for supporting this research with the grants IIS-1162046, and BCS-1436600. I thank the Graduate school of the University of Maryland for awarding me the International Graduate Research Fellowship. I thank the the A. James Clark School of Engineering for

awarding me the future faculty fellowship which supported my travel to conferences to present my work. I would also like to thank Nvidia for their device grant which enabled me to perform my research.

I thank all the staff in ECE and ISR for helping me do all the required paperwork over these years and answering all my questions related to department regulations.

A vibrant research group is essential to remain motivated and develop research ideas as a PhD student. I am grateful to my wonderful lab mates Saurabh, Nadee, Vasudha, Yi Chun, Ayanah, and Xinhui for being great colleagues over these years. I have greatly enjoyed all the discussions with them and am thankful for their feedback for my practice talks.

With my homeland so far away, friends have been like family to me in the US. I have been fortunate to have made wonderful friends over these years in Maryland. I thank Jayanand, Arun, and Kunvar for all the fun we had during the initial years as Masters students. I thank Monika for all her help with editing my research statements and cover letters. I thank her for all the amazing food that she generously shared with me and my roommates. I am grateful to have wonderful friends like Dev, Swami, Vidya, Harsha, and Subhashini who were also a great support group through the difficult journey of doctoral research. I would like to thank Develop Empower and Synergize India (DESI-UMD) and all the members of the student group over these years for the wonderful memories celebrating Indian festivals, Indian Independence day, and temple trips.

Finally, I would like to express my deepest gratitude to my parents and my

sister who have been very supportive and encouraging of my endeavors. I thank them for all their sacrifices to help me accomplish this dissertation.

Contents

List of Figures

List of Abbreviations

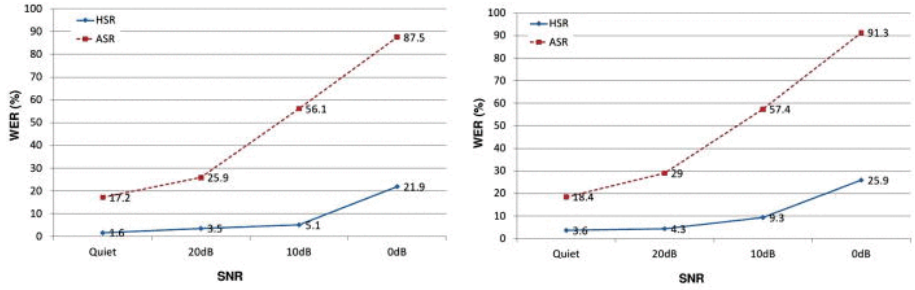
ASR	Automatic Speech Recognition
HSR	Human Speech Recognition
HCI	Human Computer Interaction
DNN	Deep Neural Network
DBN	Deep Belief Network
CNN	Convolutional Neural Network
HCNN	Hybrid Convolutional Neural Network
HMM	Hidden Markov Models
DyBN	Dynamic Bayesian Network
XRMB	X-ray Microbeam dataset
WSJ	Wall Street Journal
EMA	Electromagnetic Articulaometry
MRI	Magnetic Resonance Imaging
TV	Tract Variable
PPMC	Pearson Product Moment Correlation

Chapter 1

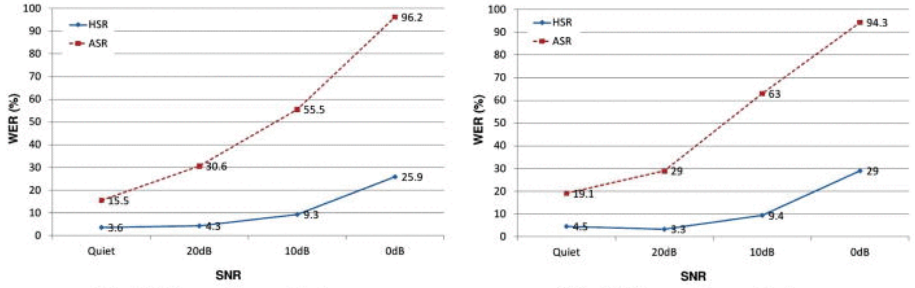
Introduction

Automatic Speech Recognition (ASR) is considered to be one of the major challenges in Human Computer Interaction (HCI) that can revolutionize human lives. ASR systems enable us to interact and operate their computers, handheld devices like smartphones, wearable technology like smart watches, and even cars. Enabling us to easily operate our devices in a hands free manner can make our lives much easier. Voice based applications can also enable the physically disabled and uneducated people to make use of the modern technology and connect with the world. Simple tasks such as searching the web, creating a reminder or a list of things to do, making a phone call, sending a text message etc. can be done today through voice commands, thanks to the advancements in personal assistant systems in smartphones and computers which have ASR as a major part of their technology. However, these technologies make a lot of mistakes due to the errors made by the ASR systems. Most systems are adept at handling dictated speech or short commands or queries in American English. But if in case of conversational or casual speech, the ASR systems perform poorly. The ASR systems are still far behind Human Speech Recognition (HSR) both in terms of accuracy and robustness. Shinozaki and Furui (2003) compared the performance of human speech recognition against state-of-the-art Hidden Markov Model (HMM) (Rabiner, 1989)

based ASR, using a corpus of spontaneous Japanese speech. They found that the recognition error rate of humans was roughly half that of the ASR system. Lippmann (1997) compared the performance of humans against HMM based ASR for six modern speech corpora with vocabularies ranging from 10 to 65000 and content ranging from isolated words to spontaneous speech. He found that the error rates for machines were often more than an order of magnitude greater than those for humans in quiet, wideband read speech. The paper also points out further degradation of ASR performance for spontaneous speech. The performances of machines have improved by leaps and bounds with the advent of Deep Neural Network based acoustic modeling (Hinton et al., 2012). The current best ASR performance on the Switchboard corpus is 8% Word Error Rate by "The IBM 2015 English Conversational Telephone Speech Recognition System" (Saon et al., 2015). The most recent study comparing ASR and HSR performance was performed by Juneja (2012). In this study, Juneja compared the performance of HSR and ASR in a null grammar setting. He argues that the null grammar test scenario is ideal for comparing the acoustic modeling performance of HSR and ASR without any context awareness or grammar. He constructed a read-speech corpus of nonsensical sentences (NullG-Eval0) containing 4 to 8 words per sentence selected randomly from vocabularies of different sizes (1000, 2000, 3000 and 4000). The final corpus had 40 sentences from each of the vocabulary sizes. These 160 sentences were recorded from 9 speakers (6 males and 3 females) to create the dataset containing 1440 null grammar utterances. He created noisy versions of the dataset from the clean corpus by adding white Gaussian noise. This created noisy datasets with 3 different SNR



(a) 1000 word vocabulary (b) 2000 word vocabulary



(a) 4000 word vocabulary (b) 8000 word vocabulary

Figure 1.1: Comparison of Word error rates for humans and a high performance HMM based ASR on sentences from a null grammar corpus (NullG-Eval0). (Figure taken from (Juneja, 2012))

levels – 0, 10, and 20dB. He presented sentences from this null grammar corpus to human subjects and a HMM based ASR system trained on read speech corpora (TIMIT, WSJ0, and WSJ1). He computed the Word Error Rates (WER) for both humans and the ASR system at different SNRs and different vocabulary sizes. He observed that the ASR system exhibits as much as an order of magnitude more errors than HSR. Figure 1.1 shows the plot of WERs for HSR and ASR across different perplexities and noise conditions. This result shows that there is still a large difference in the acoustic modeling performance of ASR and HSR.

Several efforts are being made by the research community to bridge the gap between HSR and ASR, which when achieved for spontaneous speech, will revolutionize the field of HCI. Speech variability is one of the major challenges

limiting the performance of ASR systems. Shinozaki and Furui (2003) in their work comparing human and machine speech recognition pointed out that the gap between the error rates of HSR and ASR is due to insufficient model accuracy and lack of robustness of the ASR system against "vague and variable pronunciations". There are several sources of variability in speech that severely affect ASR systems (Benzeghiba et al., 2007). Speaker's age, gender, speaker voice characteristics, speaker nativity and accent are some sources of variability associated with the identity of speakers. Speaking rate, style, loudness, emotions and pronunciation variations are other sources of variability associated with the manner of speech. Speaking rate and casualness in conversational speech result in coarticulation and lenition phenomena leading to significant variation in speech. Coarticulation is a commonly observed phenomenon in continuous speech where phonemes are influenced by the neighboring phonemes resulting in different variants based on the context (Hardcastle and Hewlett, 2006). Coarticulation is caused due to the overlap of the articulator movements (articulatory gestures) from neighboring phonemes resulting in substitution of phonemes. For example, when the phrase "did you" is uttered casually, the final /d/ in "did" and the initial /y/ in "you" can overlap and become /j/ thus sounding like "dijyu". Lenition results in reduction or deletion of phonemes. For example, Figure 1.2 shows the phrase "perfect memory" uttered at slow and fast rates. In the fast spoken utterance, we can see that the burst for the /t/ in "perfect" is missing in the acoustical waveform. This apparent deletion of /t/ has occurred due to the overlap in production with the adjacent /m/ sound, and the fact that there is undershoot in the production of the /t/ gesture. The

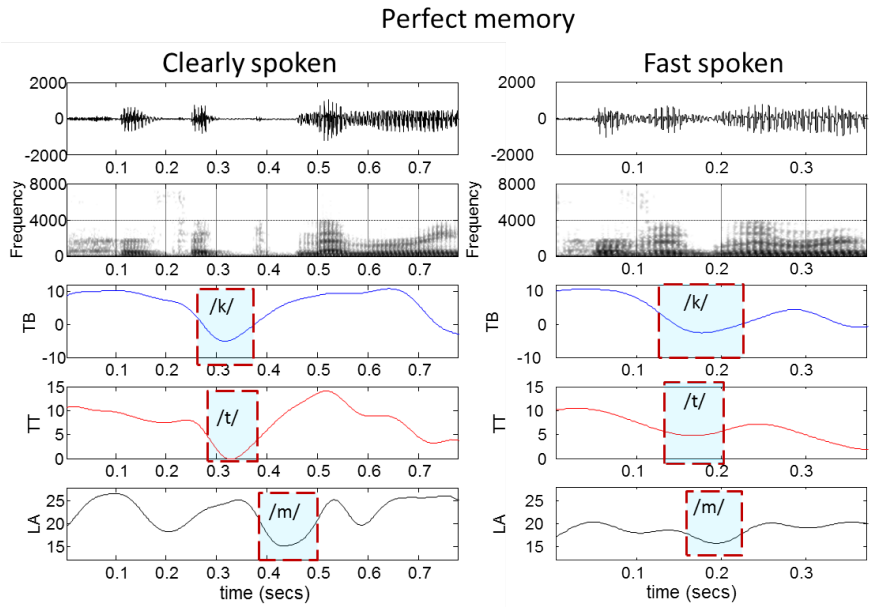


Figure 1.2: Acoustics and articulatory trajectories for the phrase "Pefect memory" uttered at two different speaking rates. TB = Tongue Body constriction, TT = Tongue Tip constriction, LA= Lip Aperture. Blue boxes show the time windows of the constrictions for /k/, /t/, and /m/ in "perfect memory" (Tiede et al., 2001)

plot also shows the recorded movements of the articulators. The TB plot shows the distance between the Tongue Body and the palate, the TT plot shows the distance between Tongue Tip and the palate. The LA plot shows the Lip Aperture i.e., the distance between the upper and lower lip. When we look at the articulator movements, we realize that in reality, the tongue tip constriction for /t/ was made but it was overlapped by the closure of the lips for /m/ thus resulting in what looks like deletion of the /t/ sound.

A major difficulty faced by conventional state-of-the-art phone based ASR systems is due to the fact that they have encoded such contextual variations as independent, piecewise, stationary units such that speech is modeled as concatenated strings of independent constant regions. To address the effect of coarticulation

in DNN and HMM-based ASR, di-phone and tri-phone based models are used, where models are created for each phone with all possible contexts. Unfortunately these di-phone or tri-phone based ASR systems limit the contextual influence to only immediately close neighbors. They require a large training data to combinatorically generate all possible di-phone or tri-phone units. In their investigation of the capability of tri-phones to model pronunciation variability, Jurafsky et al., (2001) stated that the amount of training data each tri-phone receives plays an important role in modeling pronunciation variability. They pointed out that although tri-phones are capable of modeling variabilities like phone reduction and substitution, they fail to model phone deletion. Current state-of-the-art DNN based ASR systems form phonemic categories by grouping different instances of the variations into subcategories that explicitly represent these variations (Nagamine et al., 2015). Thus, the more the variability in the labelled training data, the better it is for the ASR systems to become robust. This is particularly a challenge because correct annotation of thousands of hours of training data is an infeasible task. Considering these challenges that current ASR systems face, some fundamental questions arise – Does the human brain require such large amounts of labelled data in order to learn speech recognition? Is there a way to model speech in a way that improves the capability of modeling the variability? Most of the variability that we observe in speech is caused due to the variations and imprecision in the speech production mechanism. Coarticulation and lenition, which are commonly observed in conversational speech, are effects of contextual constraints in the articulatory movements. The variations due to rate of speech (Sivaraman et al., 2015a) and

non-native accents (Sangwan and Hansen, 2012) can also be analyzed from the articulatory features. The acoustics of all the possible triphones for a particular phoneme can be explained as consequences of overlap and lenition of the articulatory gestures for that phoneme with the gestures of its neighbors. Analyzing speech in terms of articulatory features also allows for modeling contextual effects beyond just the neighboring phonemes as is the case with triphones. The acoustic variability is a consequence of the variability in articulatory dynamics through a non-linear relationship; it is possible to model the acoustic variability more directly and parsimoniously through the articulatory parameters. Thus, articulatory feature based approaches promise better modeling of speech variability. The theory of Articulatory phonology (Browman and Goldstein, 1992) which analyzes speech as a series of coordinated actions provides us a robust framework to represent speech in terms of action units that are closely related to the physical movements in the vocal tract. This thesis proposes to borrow ideas from Articulatory phonology to develop ASR architectures that are robust against speech variability.

In the neuroscience literature there are two contending theories of speech perception (Diehl et al., 2004), namely - (1) General Auditory theory and (2) Motor theory of speech perception. The auditory theory argues that speech perception happens solely due to representations of speech in the auditory cortical regions. The motor theory says that speech production information is essential in recognizing speech and contribute to the robustness of the perception. This dissertation is inspired by the motor theory of speech perception and aims to develop a robust architecture for ASR combining both speech production and auditory information.

1.1 Objectives of this study

The objective of this thesis is to analyze the variability of speech using articulatory representation and develop an ASR system on clean speech using a combination of acoustic and articulatory features. Towards this objective, we develop a speaker independent acoustic to articulatory speech inversion system to estimate the articulatory features from speech. We analyze the performance of speech inversion systems to different kinds of variability like rate of speech, multiple speakers, accent, and language. We explore state-of-the-art convolutional neural network (CNN) architectures to optimally combine acoustic and articulatory features for ASR.

The specific contributions of this thesis are as follows: • Development of a speaker independent speech inversion system on the multi-speaker X-ray Microbeam dataset. Chapter 3 describes the design and development of the speech inversion system.

• A speaker adaptation algorithm for speaker independent speech inversion. Section 4.2.2 explains the speaker adaptation algorithm.

• Analysis of speech inversion across different speech variability. We consider the speaker variabilities, speaking rate, and accent. The analyses are presented in Chapter 4.

• Hybrid convolutional neural network architecture for acoustic and articulatory feature based ASR. Chapter 7 presents the ASr experiments performed on the Wall Street Journal (WSJ) dataset. We developed a hynrid CNN architecture

to best combine the acoustic and articulatory feature streams for ASR. We present the ASR experiments and results in chapter 7.

• Cross-corpus phone recognition. ASR systems tend to perform very well on the domain that they are trained on. However they perform poorly across domains. We evaluated our WSJ dataset trained ASR systems on the TIMIT dataset and another dataset containing normal and fast rate speech that we collected. The cross-corpus phone recognition experiments and results are presented in section 7.5.

Chapter 2

Background: Articulatory features and their application to ASR

2.1 Relevance of articulatory features in today's state-of-the-art ASR

Automatic Speech Recognition has been an active area of research for the past five decades. Early attempts at ASR were rule based models that were formed using phonetic feature theory (Lieberman, 1970) and the spectral properties of phonemes (Halle and Stevens, 1962) (Zue and Lamel, 1986) (Espy-Wilson, 1994). With increasing complexity of the speech corpora (going from isolated phonemes to continuous read speech), such rule based systems became too complicated to model the variabilities of continuous speech. It is then, that statistical generative models such as Hidden Markov Models (HMMs) (Rabiner et al., 1985) were introduced to ASR research. HMMs did not rely on knowledge based rules but rather formed implicit rules based on statistics derived from the data. Homomorphic signal processing (Oppenheim, 1969), combined with HMMs revolutionized the ASR technology and grew to be the state-of-the-art for ASR since the late 80s through the early 2010. The next big leap came from the introduction of Deep learning techniques where the state-of-the-art ASR acoustic models changed from generative GMM-HMM systems to a hybrid DNN-HMMs (Hinton et al., 2012) system that have the power of discriminative models and the flexibility of the generative models. After all these machine learning advancements for ASR, a recent study (Nagamine

et al., 2015) exploring the inner learned structures of DNN acoustic models has shown that the hidden layers of the DNN form phonemic categories by selective tuning of the nodes to various phonetic features (manner and place of articulations of phonemes) as outlined in the phonetic feature theory (Sim, 2016). This brings back the attention of ASR research on phonetic features, articulatory features and acoustic phonetics. These fields of study could play a great role in improving the DNN acoustic models thus pushing the limits of state-of-the-art ASR. Hence, it is essential to look back at speech analysis based on these theories and integrate them with the new models like DNNs, and convolutional Neural Networks (CNNs) to make improvements to the current state-of-the-art ASR.

2.2 Types of articulatory features

Articulatory features are a parametric representation of the vocal tract configurations and movements during speech production. As gleaned from the literature, there are mainly two forms of representation of articulatory features – (1) Continuous articulatory features and (2) Discrete articulatory features. Several studies have been performed in the literature by analyzing speech in the form of both types of articulatory features. Both types of articulatory features have been extensively applied to ASR research.

2.2.1 Continuous articulatory features

This class of articulatory features consists of different parameterizations of real or simulated movements of the vocal tract. The most direct way to capture articulatory information from speech is by placing transducers on the articulators and recording their movements while speech is produced.

2.2.1.1 Methods of measuring or synthesizing articulatory data

The earliest attempts used to measure the contact of the tongue with the hard palate involved Electropalatography (EPG) (Hardcastle, 1972). Although EPG provided the instances and locations when the tongue came in contact with the palate, it provided no information of the shape of the tongue when there were no constrictions or the constriction targets were not reached.

The X-ray microbeam (XRMB) (Westbury, 1994) technique was a method of measuring the movement of different points along the vocal tract. The XRMB technique recorded flesh point trajectories of gold pellets placed at different points along the vocal tract obtained using X-ray photographs of subjects.

Electromagnetic Articulography (EMA) (Schönle et al., 1987) is a more modern and commonly used method in speech production research. EMA tracks the movement of electromagnetic pellets placed at different points along the vocal tract as a subject speaks. EMA is known to have a good time resolution of tracking the articulator movements. In both the EMA and XRMB methods, the dataset consists of speech signals with simultaneously recorded trajectories (X-Y positions on the

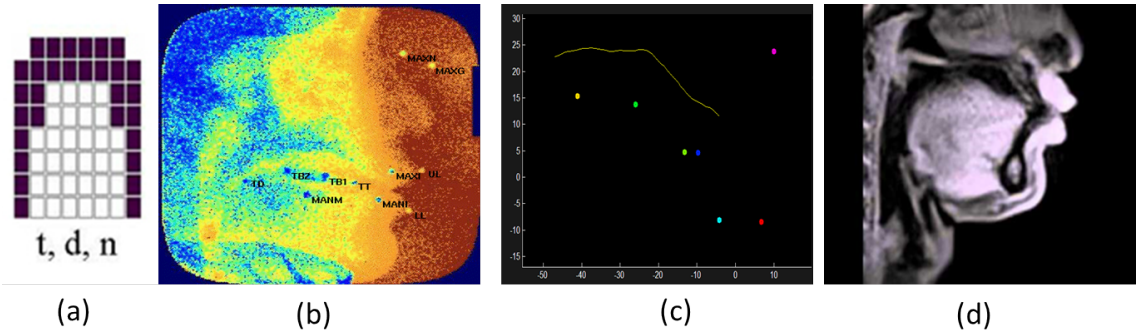


Figure 2.1: A frame of articulatory recording from different measurement techniques - (a) Electro Palatography, (b) X-ray microbeam, (c) Electromagnetic Articulometry (EMA), (d) Real Time Magnetic Resonance Imaging (rt-MRI)

mid-sagittal plane) of the pellets placed along the vocal tract. These trajectories are referred to as pellet trajectories.

A more recent method using real time Magnetic Resonance Imaging (rt-MRI) (Narayanan et al., 2004) was developed to image the complete mid-sagittal view of the vocal tract. Rt-MRI has a high spatial resolution but a low temporal resolution. The rt-MRI databases (Narayanan et al., 2011) consist of speech signals along with video of MRI images recorded simultaneously with speech.

Figure 2.1 shows a frame of articulatory recording from each of the measurement techniques discussed above - EPG, XRMB, EMA, and rt-MRI.

Articulatory data can also be generated using synthetic speech production models. The Task Dynamics and Applications (TADA) model (Nam et al., 2004) from Haskins Laboratories is an articulatory speech production model that includes a task dynamic model and a vocal tract model. The task dynamic model of speech production (Saltzman and Munhall, 1989) (Nam et al., 2004) models speech as a constellation of gestures with dynamically specified parameters as model input

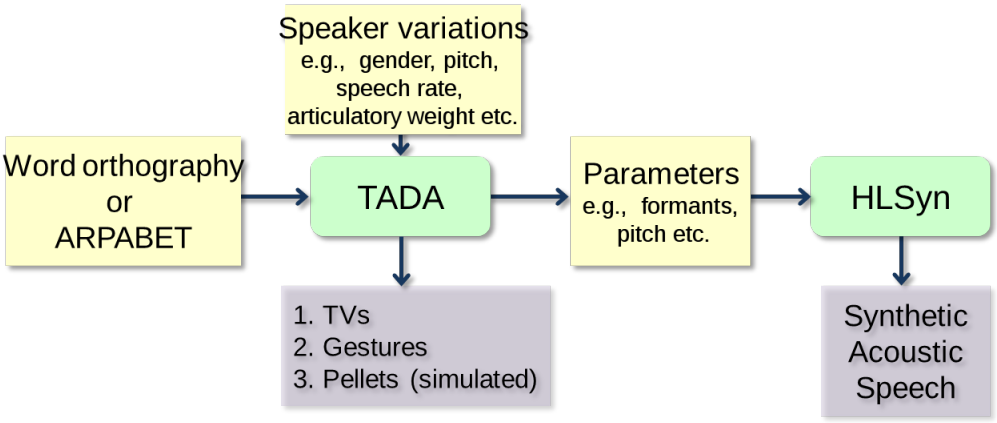


Figure 2.2: Synthetic speech and articulatory information generation using TADA and HLSyn. Figure taken from (Mitra, 2010)

for an utterance. The gestures and the inter-gestural co-ordinations parametrize the constriction actions performed by the articulators and the inter articulator coordinations. Given the orthographic transcription of an utterance, the model computes the series of gestures and inter-gestural coordinations required to produce the utterance. This is performed using a lookup table and predefined rules. The gestural patterns are input to a second order dynamical system which produces time functions of the physical trajectories for each vocal tract variable (TV). The time functions of model articulators are input to a vocal tract model which computes the area function and the corresponding formants. The formants and pitch information are used to generate a synthetic speech waveform using the HLSyn^{TM} toolkit (Hanson and Stevens, 2002). Figure 2.2 shows the schematic of the TADA system that generates synthetic speech and articulatory trajectories.

The TADA model is a theoretically sound model of speech production. However the model is limited in the amount of articulatory variability that it can simulate. There is much greater variability in real speech compared to what TADA

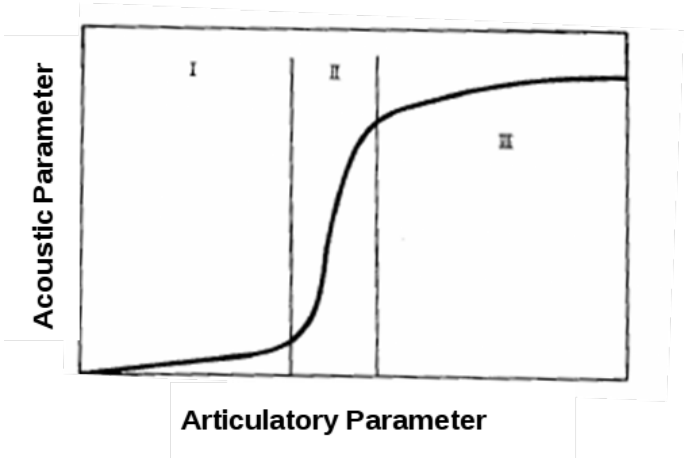


Figure 2.3: Schematic showing a change in a relevant acoustic parameter as an articulatory parameter specifying some aspect of the state or configuration of the speech production system is manipulated. The curve can be divided into regions I, II, and III. In regions I and III the acoustic parameter remains relatively stable when small modifications are made in the articulation. In region II there are large changes in acoustics for small shifts in articulation. There is a significant acoustic contrast between regions I and III. Figure taken from (Stevens, 1989)

can simulate. The synthetic speech generated by HLSyn also sounds machine like quite different from real speech. In order to study the variability of speech real articulatory data is more preferable.

Recording of articulatory data is an expensive and time consuming process which involves sophisticated equipment housed in a laboratory. This makes it impractical to create large speech databases containing articulatory data. Hence it is essential to train models from the available articulatory data to learn a mapping from acoustics to articulations, known as acoustic to articulatory speech inversion.

The mapping from the acoustics to articulatory space has been found to be nonlinear and non-unique based on empirical studies (Stevens, 1989) (Qin and Carreira-Perpiñán, 2007) (Neiberg et al., 2008).

The non-linearity of the acoustic to articulatory mapping is evident from the

quantal nature (Stevens, 1989) of the articulatory-acoustic relation. The change in acoustic parameters corresponding to the manipulation of articulatory structure through a range of values is non-monotonic. There are ranges of articulatory parameter for which there is very little change in the acoustic parameter and other ranges where the acoustic parameter is more sensitive to changes in articulation (Stevens, 1989). The quantal nature of the acoustic-articulatory relation is shown in the Figure 2.3

Similar acoustic consequences can result from completely different articulatory configurations, thus making the problem of one-to-many mapping a challenge. One of the reasons for this non-uniqueness is the coordinated compensatory movements of the articulators to achieve acoustic targets in different contexts (Maeda, 1990) (Guenther et al., 1999). The existence of two distinct vocal tract configurations (bunched and retroflexed) for the American English /r/ sound (Zhou et al., 2008) is a perfect example of the non-unique mapping if we consider only the first three formants of speech. [1] Several machine learning techniques have been developed by various researchers to perform this challenging task.

2.2.1.2 Speech inversion: Estimating articulatory features from speech

Speech inversion or acoustic-to-articulatory inversion of speech has been a widely researched topic in the last 35 years. One of the earliest works in this

[1] Zhou et al. (2008) show that the higher formants do distinguish between these vocal tract shapes.

area was by Atal et al. (1978). Their model used four articulatory parameters: length of the vocal tract, distance of the maximum constriction region from the glottis, cross sectional area at the maximum constriction region and the area of the mouth opening. For each articulatory configuration, the corresponding acoustic space was defined by the frequency, bandwidth and the amplitudes of the first five formants. They stored these articulatory acoustic pairs in the computer memory as codebooks. Thus given information in acoustic space, their approach would yield the corresponding vocal tract configuration by scanning through the codebook. Following a similar codebook based approach, Rahim et al. (1991) used an articulatory synthesis model to generate a database of articulatory-acoustic vector pairs. They characterized the acoustic space using 18 FFT-derived cepstral coefficients, and the articulatory space using 10 vocal tract areas and a nasalization parameter. They trained Multi-Layered Perceptrons (MLP) to map from acoustic data to the vocal tract area functions. The articulatory-acoustic data pairs were obtained by random sampling over the manifold of reasonable vocal tract shapes within the articulatory parameter space of Mermelstein's articulatory model (Mermelstein, 1973). However their random sampling approach selected many uncommon but physiologically-plausible articulatory configurations. To address this fact, Ouni and Laprie (2005) sampled an articulatory space such that the inversion mapping is locally linearized. They sampled more aggressively in regions where the inversion mapping is complex and less aggressively elsewhere. Codebook based reconstruction of continuous articulatory data suffer from noisy estimates of the articulator movements. However, the movements of the articulators are

smooth and low pass signals. To ensure the smoothness, codebook based techniques apply a low pass filter after the estimation. Some approaches have tried to impose the smoothness criterion within the optimization problem for reconstruction such that the estimates have smooth trajectories. One of the earliest such approaches was introduced by Schroeter and Sondhi (1994) in which they used Dynamic Programming to search the articulatory codebooks by imposing a penalty on fast changes in the articulatory trajectories so that the estimated articulator trajectories were smooth. More recently, Ghosh and Narayanan (2010) introduced a generalized smoothness criterion to recover optimal smooth trajectories using a codebook based technique. They incorporated the smoothness criterion in the estimation problem by minimizing the energy of the output of high pass filters applied to the target articulatory trajectories. The high pass filters were designed separately for each articulator based on the frequency content of each articulatory trajectory in the MOCHA TIMIT database (Wrench, 2000). They showed that having articulator specific optimal smoothing filters in the smoothness criterion was better than using a fixed smoothing filter. Artificial Neural Network (ANN) based techniques were also widely explored to perform speech inversion. Kobayashi et al. (1991) proposed a feed-forward MLP architecture with two hidden layers to predict the articulatory parameters. Their approach was found to be 10 times faster than Shirai and Kobayashi (1986) and also offered better estimation accuracy than analysis-by-synthesis techniques. Neural network based speech inversion became very popular after the work of Papcun et al. (1992) in which they used articulatory information obtained from XRMB and used Multiple Layer Perceptrons (MLPs) to

perform speech inversion to obtain three articulatory motions for six English stop consonants. The three articulatory features used were the y-coordinates for the lower lip, tongue tip and tongue dorsum. They used data recorded from three male native American English speakers, who uttered six non-sense words. The words had repeated [-Cə-] syllables, where 'C' belonged to one of the six English oral stop consonants /p,b,t,d,k,g/. The MLP topology was decided based upon trial-and-error and the optimization of the topology was based upon minimizing training time and maximizing estimation performance. The network was trained using a standard backpropagation algorithm. An important observation that they made in their study was that the trajectories of articulators considered critical for the production of a given consonant demonstrated higher correlation coefficients than for those which were considered non-critical to the production of that consonant. This observation was termed as the 'Critical articulator phenomenon'. Due to this phenomenon they observed that for a given consonant, the critical articulator dynamics were found to be much more constrained than that of the non-critical ones. This observation was supported by Richmond et al. (2003) who used Mixture Density Networks (MDN) to obtain the articulator trajectories as conditional probability densities of the input acoustic parameters. They observed that the conditional probability density functions of the critical articulators show very small variance as compared to the non-critical articulator trajectories. They also showed that MDNs performed better than ANNs in estimating the articulatory trajectories. Their experiments were based on single speaker articulatory data from the MOCHA TIMIT dataset.

Mitra et al. (2010a) used the TADA model to generate synthetic datasets containing spech with simultaneous Tract Variables (TVs) for the sentences in the X-ray microbeam (XRMB) dataset (Westbury, 1994). He then used the synthetic data to train speech inversion systems using different machine learning techniques. He performed a comparison of different machine learning techniques (Mitra et al., 2010b) for estimating TVs from speech and found that feed forward neural networks performed the best for estimating the TVs. He used contextualized MFCCs as the input acoustic features for estimating the TVs. He obtained estimation accuracies of up to 90% correlation between estimated and actual TVs. He also showed that the neural network models estimated TVs with greater accuracy than the pellet trajectories. Later, Mitra et al., (2014) (Mitra et al., 2014a) developed a Deep neural network based speech inversion system trained on synthetic data generated from 111,929 words in the CMU dictionary using the TADA model. This inversion system achieved correlation values of 95% on an average across the 8 TVs. Their work also compared various acoustic features for the speech inversion system and found MFCCs to be the best performing feature.

Deep neural networks have been used for estimating the flesh point trajectories from real articulatory datasets. Uria et al. (2011) implemented a Deep Belief Network based speech inversion system on a single speaker from the MOCHA TIMIT dataset. They obtained improved results with average root mean square error of 0.95mm which was a significant error reduction from the performance obtained using TMDNs (Richmond et al., 2003) (0.99mm mse). They observed that the unsupervised layer wise pre-training of the deep networks using Gaussian Bernoulli

RBMs greatly improved the performance of the ANN based approach. Recently, Liu et al. (2015) developed deep recurrent neural networks using Bidirectional Long Short Term Memories (BLSTM). They reported best speech inversion results with an average root mean squared error of 0.816 beating the Deep Belief network model. They observed that the trajectories estimated by the recurrent neural networks were smoother than those predicted by non-recurrent networks without any smoothing.

Gaussian mixture modeling (GMM) is another popular approach to speech inversion. Toda et al. (2004) modeled the mapping from articulatory space to acoustic space using GMMs. They performed speaker dependent experiments on the MOCHA TIMIT dataset. The found that the estimation accuracy improved steadily by increasing the number of Gaussians up to a certain point. However, they found that the estimates were noisy and had discontinuities. In order to obtain smooth trajectories, they developed Maximum Likelihood Estimation (MLE) technique by including articulatory dynamic feature to obtain smooth estimates of the articulatory trajectories. They found that the MLE based inversion system gave more accurate articulatory movements compared to the GMM based mapping with low pass filtering. A Hidden Markov Model based speech inversion technique was developed by Hiroya and Honda (2004). They modeled phone specific HMMs with an articulatory-to-acoustic mapping for each hidden state. Given a test utterance, the inverse mapping was performed by estimating the optimal state sequence followed by a Maximum-a-posteriori (MAP) estimate of the articulatory parameters using dynamic features to obtain smooth estimates. They performed experiments with and without prior phonemic information. They obtained an rms error of 1.50mm

with phone information and 1.73 without the phone labels. HMM and GMM based modeling have the potential to develop speaker independent speech inversion and speaker adaptation using maximum likelihood and MAP model adaptation as they have explicit conditional probability models for the acoustic and articulatory spaces.

Efforts have also been made in implementing dynamic models for performing speech inversion. Dusan and Deng (2000) used Extended Kalman Filter (EKF) to perform speech inversion by imposing high-level phonological constraints on the articulatory estimation process. In their approach Dusan and Deng (2000) segmented the speech signal into phonological units and construct the trajectories based on the recognized phonological units and a Kalman smoothing step is used to perform the final estimate. Dynamic model based approaches typically work well for vowel sounds, but have failed to show promise for consonantal sounds.

Speaker variability is a common challenge that is encountered by speech inversion systems. Speech inversion systems trained on one particular speaker perform very poorly when tested on another speaker. This is because of the speaker variability in the acoustic as well as articulatory domains. There have been few attempts in the research community to address the challenge of speaker adaptation for speech inversion systems. Among the most recent attempts at speaker normalization for speaker independent speech inversion is the work by Ghosh and Narayanan (2011). In their work, they developed a Generalized Acoustic Space (GAS) consisting of unsupervised GMM models trained on acoustic data (MFCC features) from the TIMIT database which consists of a large number of speakers. The GAS is a probability distribution of the acoustic features from a

large pool of English speakers. This GAS formed the acoustic model to perform acoustic feature matching for their codebook based speech inversion system using the generalized smoothness criterion. They performed their experiments on the MOCHA TIMIT corpus with male-female cross test and showed performance close to a speaker dependent system for lip aperture, tongue tip and tongue body. Following up on this work, Afshan and Ghosh (2015) developed various supervised and unsupervised training techniques to refine the GAS to enhance the performance of speaker independent speech inversion. They observed that clustering the generalized acoustic space based on phone identities gave best speech inversion results. Hueber et al. (2015) developed a Cascaded Gaussian Mixture Regression technique to perform speaker adaptation for a GMM based speech inversion system. Their work exploited the state-of-the-art speaker adaptation techniques like MAP and MLLR to perform acoustic model adaptation before performing the inversion. They also developed an improved technique by combining the adaptation problem with the inversion problem to show better adaptation performance compared to MAP and MLLR.

2.2.2 Discrete articulatory features

Discrete articulatory features are binary features related to the state of articulators during speech production. In the literature, there are two types of discrete articulatory features - (1) distinctive features, and (2) articulatory gestures. The distinctive features (DF) as defined in phonological theory (Chomsky and Halle,

1968) are a set of binary features that group phonemes into categories that are discriminative. On the other hand, articulatory gestures are closely related to the physiological movements of the articulators during speech production.

DFs have both articulator bound features related to place of articulation and articulator free features related to manner of articulation (Stevens, 2002). DFs can be obtained from acoustics by detection of acoustic landmarks from speech (Stevens, 2002). This formed the basis of landmark based speech recognition systems which performed speech recognition by detecting acoustic landmarks from speech using rules based on the speech spectrum (Espy-Wilson, 1994). The acoustic landmarks were then used to infer DFs and word or phone recognition was performed based on the decoded DFs for a given utterance. A probabilistic framework to detect DFs from speech using an ensemble of Support Vector Machines (SVMs) was developed by Juneja and Espy-Wilson (2008). Kirchhoff (1999) used a set of heuristically defined articulatory features inspired by the DFs. She used rule based mappings from phones to the articulatory features to generate the groundtruth for small and large vocabulary speech datasets. She estimated the articulatory features using Multiple Layer Perceptron (MLP).

As mentioned earlier, Articulatory phonology (Browman and Goldstein, 1992) analyzes speech as a series of constriction gestures performed by the articulators in the vocal tract. Gestures of two different articulators can overlap and influence each other based on the coupling between the articulator movements. These gestures are control parameters for the Task dynamical system that models the movement of the articulators based on second order dynamics. The TADA system implemented this

model of speech production as a Matlab toolkit (Nam et al., 2004). The TADA model defines gestures by the following parameters: (1) gestural score, (2) the mass parameter, which is assumed to be uniformly equal to 1 in all gestures, (3) the stiffness parameter, which represents the elasticity of the gesture and is proportional to gestural "speed", (4) the damping parameter, which is typically set to "critical" in the gestural model to signify that there is no oscillatory overshoot or undershoot of the TVs and when the gesture moves closer to its target, this parameter gives the TV its inherent smoothness, (5) the target parameter, which defines the constriction location or degree for that particular TV on which that gesture is defined and (6) the blending parameter, which defines how two overlapping gestures corresponding to the same TV should be blended with one another. Nam et al. (2012) created a synthetic dataset corresponding to the XRMB dataset and developed an algorithm to annotate real speech data with gestures defined by TADA.

Mitra et al. (2011) used the TADA model to generate synthetic gestures, TVs and synthetic speech for the Aurora2 database. They developed three different ANN based architectures to estimate gestures from speech. They found that combining both TVs and MFCCs as input features, performed the best for gesture recognition. They performed word recognition experiments on the Aurora 2 dataset and observed that the estimated gestures improved the noise robustness of the HMM based word recognition systems.

In an effort to represent continuous real articulatory trajectories as a combination of a fixed number of basis vectors, Ramanarayanan et al., (2015) (Ramanarayanan et al., 2015) developed a convolutive Non-negative Matrix

Factorization (xNMF) algorithm to represent articulatory trajectories as gesture like primitives. They performed an interval based phone classification task and found that such data derived primitives retained the discriminatory information about phone categories.

2.3 Articulatory feature based ASR systems

Articulatory feature based ASR has been an active research area for the past two decades. Several methods have been developed to incorporate various forms of articulatory features in ASR systems. Broadly, the approaches can be classified into two categories. In the first category, articulatory features extracted from speech are appended to acoustic features in a standard ASR architecture. In the second kind of approach, the ASR architecture is modified in order to efficiently combine acoustic and articulatory features.

In the first category, the earliest approaches were by Zlokarnik (1995) and Wrench and Richmond (2000). Zlokarnik, (1995) performed isolated word recognition by appending measured articulatory data with acoustic features in a HMM based word recognition system. With actual articulatory data, he reported a 60% relative improvement in word error rate whereas when he appended the articulatory features estimated using a MLP model, the improvement was 18-25%. Wrench and Richmond (2000) however did not report significant improvement to the phone recognition performance by appending estimated articulatory features (AFs). Kirchhoff, (2000) performed Large Vocabulary Conversational Speech Recognition

on the German Vermobil corpus using discrete articulatory features estimated from the acoustics using MLPs. She combined acoustic and estimated AFs in three different ways in a HMM based LVCSR system: HMM state level combination, word level combination and input feature level combination. Among the three methods, the HMM state level combination gave the best WER improvement over the MFCC baseline system. She also observed that adding articulatory features to acoustic features improves the noise robustness of the ASR system.

Livescu et al. (2007) created a database of spontaneous speech which was manually labeled at the articulatory feature level. They considered a small subset of the Switchboard corpus and transcribed it with eight tiers of AFs. One of the most important attributes of this database was that it allowed some inter-AF overlapping, which was not used in any of the AF based systems or databases proposed before. In a different study, Çetin et al. (2007) proposed a tandem model of MLP and HMM as an ASR system. The MLPs were used for AF classification and the HMM outputs used a factored observation model. Their proposed tandem model using AFs was found to be as effective as the phone-based model. The factored observation model used in their research was found to outperform the feature concatenation approach, which indicated that the acoustic features and tandem features yield better results when considered independently rather than jointly.

It was shown by Arora and Livescu (2013) that the multi-view learning based approach of Kernel Canonical Correlation Analysis (KCCA) can be used to obtain a feature set that learns a joint representation of both the acoustic and articulatory spaces. They showed significant improvement in phone recognition results from

features learned through KCCA for cross-speaker and cross corpus settings. Their results proved that the KCCA based features learned a speaker and corpus invariant representation of the joint acoustic and articulatory spaces.

Articulatory features derived from synthetic speech based inversion models, have been shown to reduce the Word Error Rate (WER) of an LVCSR system on the Aurora 4 dataset (Mitra et al., 2014a). Badino et al. (2016) developed methods to integrate articulatory features in a DNN-HMM based phone recognition system. They developed three different DNN architectures for acoustic to articulatory inversion. They also performed experiments with autoencoder transformed articulatory trajectories. They argued that the autoencoder transformed features encode the inter-articulator coordination. They showed higher mutual information between autoencoder features and phone labels than that between articulatory trajectories and phones. They incorporated the articulatory features in two different ways in the DNN-HMM phone recognition system. The first method simply concatenated the acoustic and the reconstructed AFs to form the input vectors for the DNN acoustic models. They obtained a 10% reduction in phone error rates by this method for a speaker dependent system, but the system performed poor than the baseline in the cross speaker case. In the second method, they initialized the DNNs for the acoustic modeling with the weights of the articulatory inversion DNN. With this method, they obtained 2% relative reduction in phone error rate in the cross speaker case without appending the articulatory features with acoustic features. They claim that the speech inversion based pretraining is a promising method to use measured articulatory data for training ASR systems

on other datasets. Dynamic Bayesian Networks (DyBN) has also been explored for the purpose of ASR using articulatory features. The major advantage of DyBN is its capability to model explicitly the inter-dependencies between the AFs. DBNs can be used to perform both AF recognition and word recognition simultaneously. One of the earlier works incorporating DBNs for the task of AF recognition was performed by Frankel et al., (2007) (Frankel et al., 2007). It was observed that modeling inter-feature dependencies improved the AF recognition accuracy. In their work, they created phone-derived AFs and set that as the standard, by modeling inter-feature dependencies; they observed an improvement in overall frame-wise percentage feature classification from 80.8% to 81.5%. Mitra et al., (2012) (Mitra et al., 2012) developed a neural network based architecture for estimating articulatory gestures from speech. The gestures were based on the TADA model. The gesture recognition system was trained on synthetic speech and was used to test on real speech. They developed gesture based DyBN architectures for performing word recognition on the Aurora-2 corpus. They modeled the gestures as discrete hidden random variables and the acoustic features as the observations in the gesture based DyBN. Word recognition results showed that incorporating gesture information improved the ASR performance compared to acoustic only systems in the noisy test conditions.

Thus, based on the literature survey we observe that there has been considerable effort to perform speech inversion and articulatory feature based ASR in the speaker dependent setting. Most speaker independent ASR experiments have been performed with articulatory features from speech inversion systems trained

on synthetic data. In this thesis we propose to perform a careful analysis of the variability in the acoustic and articulatory space. We plan to explore the speaker variability, by performing speech inversion experiments on the XRMB data (Westbury, 1994) which consists of speech and articulatory data from 46 different speakers of American English. We then plan to explore the variability in speech that occurs due to varying rate of speech. We will analyze a recently recorded EMA articulatory database consisting of utterances at fast and normal speaking rates. Articulatory gestures have so far been defined theoretically based on phone identities or from the TADA model. To the best of our knowledge, there has been no effort to define gestures for real articulatory trajectories. We propose to define articulatory gestures using the kinematics of the real articulatory data and perform gesture recognition. Finally, we will evaluate our articulatory features by performing phone recognition experiments on the Wall Street Journal and TIMIT datasets.

Chapter 3

Acoustic to articulatory speech inversion

3.1 Overview

The lungs, glottis, velum, tongue, lips, teeth etc. are the major organs that are actively involved in the speech production process. These organs are called articulators. The vibration of the vocal folds or the lack of it determines if the sound is periodic or not. The vocal tract acts like an acoustic tube that modulates the glottal source waveform leading to the resonances that are characteristic of any phoneme. The various phonemes of any language are the outcome of the different vocal tract shapes along with the state of the vocal folds. Thus, speech is produced by the movement of these articulators, molding the shape of the vocal tract to produce the series of phonemes which make up the building blocks of language. Figure 3.1 shows a schematic of the physical processes that are involved in speech production.

3.2 Modes of measuring articulatory data

There are different ways to measure real articulatory data. The earliest attempts measured the contact of the tongue with the hard palate using Electropalatography (EPG) (Hardcastle, 1972). This technique shows just an array

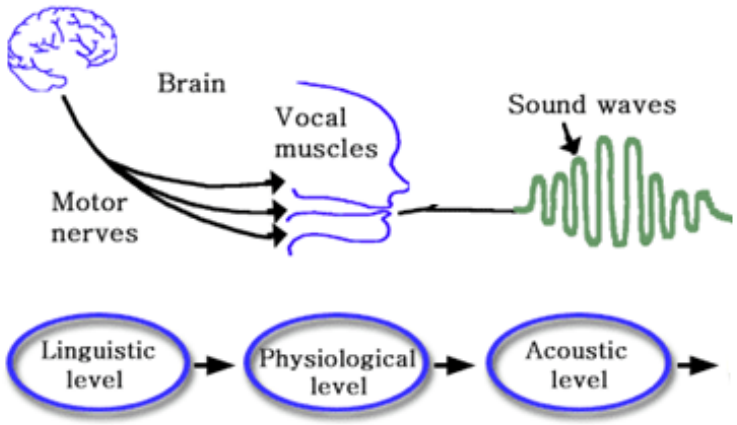


Figure 3.1: Schematic of physical processes involved in speech production (Denes and Pinson, 2015)

of points on the hard palate that are in contact with the tongue. Without tongue constriction, we don't get much information about the shape of the tongue or the vocal tract. A mid-saggital view of the vocal tract gives a better picture of its shape. X-ray Microbeam (Westbury, 1994) is a technique that gives a midsaggital view of the vocal tract. In this technique, gold pellets are placed at various points along the vocal tract of a subject and the motions of these pellets are tracked using X-ray photographs as the subject is speaking. One of the earliest datasets measuring articualtory movement at multiple points of various articulators along the vocal tract is the University of Wisconsin X-ray microbeam (XRMB) database (Westbury, 1994).

Electromagnetic Articulography (EMA) (Schönle et al., 1987) is a more commonly used method in speech production research. EMA tracks the movement of electromagnetic pellets placed at different points along the vocal tract as a subject speaks. EMA is known to have a good time resolution of tracking the articulator

movements. EMA articulatory data is typically measured at 200Hz sampling rate and downsampled to 100Hz for analysis. The XRMB dataset samples different sensors at different sampling rates because of their limit of 700 samples per second of aggregate sampling rate across all pellets. They sample the T1 (Tongue tip) pellet with the highest sampling rate of 160Hz and the UL (Upper Lip) pellet with the lowest sampling rate of 40Hz (Westbury, 1994). All the sampled pellet data are resampled to 200Hz. In both the EMA and XRMB methods, the dataset consists of speech signals with simultaneously recorded trajectories (X-Y positions on the mid-sagittal plane) of the pellets placed along the vocal tract. These trajectories are referred to as pellet trajectories. A more recent method using real time Magnetic Resonance Imaging (rt-MRI) (Narayanan et al., 2004) was developed to image the complete mid-sagittal view of the vocal tract. Rt-MRI has a high spatial resolution but a low temporal resolution. The rt-MRI databases consist of video of MRI images recorded simultaneously with speech. Figure 2.1 shows a frame from each of these different methods of measuring articulatory information. Apart from these real articulatory data, a discrete manner and place based articulatory features based on phonetics was defined by (Kirchhoff, 1999). This thesis does not deal with discrete articulatory representations.

3.3 Tract Variables and Gestures

Tract variables are another form of representation of articulatory data that is derived from the TAsk Dynamics and Applications (TADA) model (Nam et al.,

2004) of speech production. The TADA model is based on the theory of Articulatory Phonology (Browman and Goldstein, 1992) that defines speech as a constellation of coordinated articulatory gestures. Speech gestures can be defined as constricting actions for distinct organs/constrictors (lips, tongue tip, tongue body, velum and glottis) along the vocal tract. Each gesture is dynamically coordinated with a set of appropriate articulators. A word can be defined as a constellation of distinct gestures (gestural scores). Given the ARPABET transcription of an English word, the TADA model computes the gestural scores along with the inter-articulatory gestural coordination to produce the word and outputs the time functions of the vocal tract variables (TVs: degree and location variables of the constrictors) and model articulator variables. The Matlab implementation of TADA, interfaces with the HLSyn (Hanson and Stevens, 2002) speech synthesis toolkit to synthesize speech from the parameters generated by TADA. Thus, the TADA model provides a working theoretical framework for speech production. The vocal tract time functions or Tract Variables (TVs) are time-varying physical realizations of gestural constellations at the distinct vocal tract sites for a given utterance. These TVs describe geometric features of the shape of the vocal tract tube in terms of constriction degree and location. Each TV has its corresponding gestural score in the gestural space. There are eight TVs as defined by the TADA model. They relate to the lips, tongue, jaw, glottis, and velum. Table 3.1 lists the different TVs along with their associated articulators. Figure 3.2 shows the theoretical definition of TVs plotted on a model of the vocal tract defined by TADA.

Figure 3.3 shows the gestural activations and TVs for the utterance "miss you"

Table 3.1: Constriction organ, tract variables and their associated model articulators

Constriction organ	**Tract variables**	**Articulators**
Lip	Lip Aperture (LA)	Upper lip, lower lip, jaw
	Lip Protrusion (LP)	
Tongue Body	Tongue body constriction degree (TBCD)	Tongue body, jaw
	Tongue body constriction location (TBCL)	
Tongue Tip	Tongue tip constriction degree (TTCD)	Tongue body, tip, jaw
	Tongue tip constriction location (TTCL)	
Velum	Velum (VEL)	Velum
Glottis	Glottis (GLO)	Glottis

obtained from TADA. A gestural score is a binary parameter which defines whether a gesture is active or not at a given time instant. The gestural scores are shown as shaded regions of the TVs, and the active gestures during consonants are outlined by a green rectangle, and the active gestures during vowels are outlined by a black rectangle in Figure 3.3. The TVs are shown as continuous curves in the background of Figure 3.3.

Although the TADA model is a sound theoretical model for speech production, it is currently not possible to produce the amount of variability observed in real speech using TADA. Both the synthetic speech and TVs do not exhibit the variability observed in real speech. As a result, the focus of this thesis is on real speech and articulatory data collected from subjects. We have developed methods to convert the X-Y pellet trajectories from XRMB and EMA to Tract Variables (Nam et al., 2012) (Sivaraman et al., 2015b)(Sivaraman et al., 2017). Absolute positions of the points

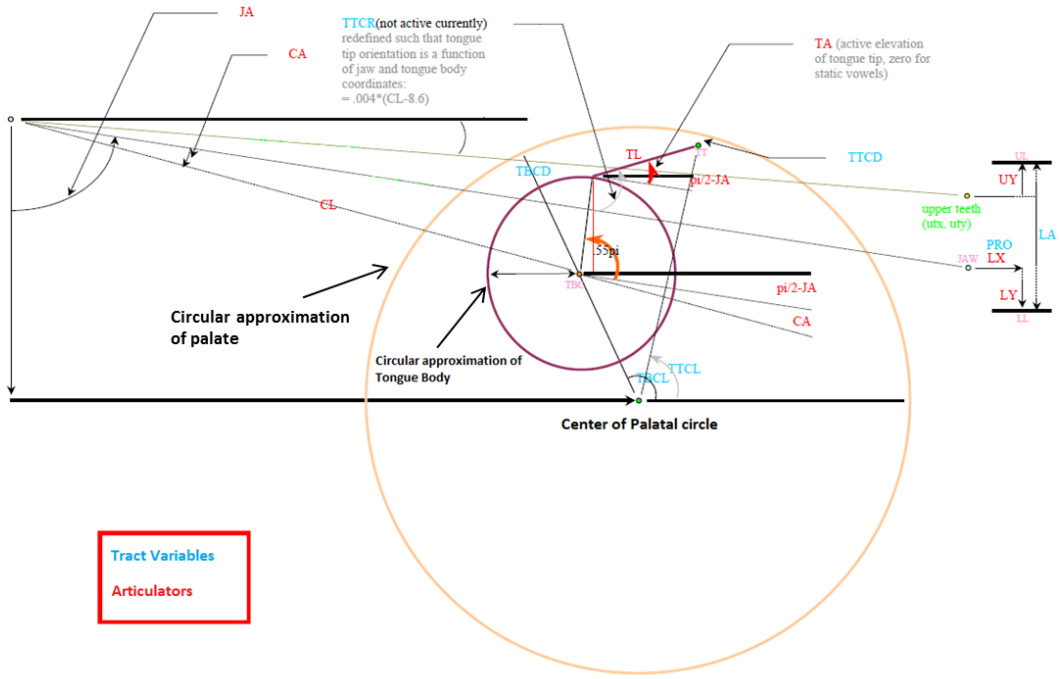


Figure 3.2: Schematic showing TADA model's definition of TVs

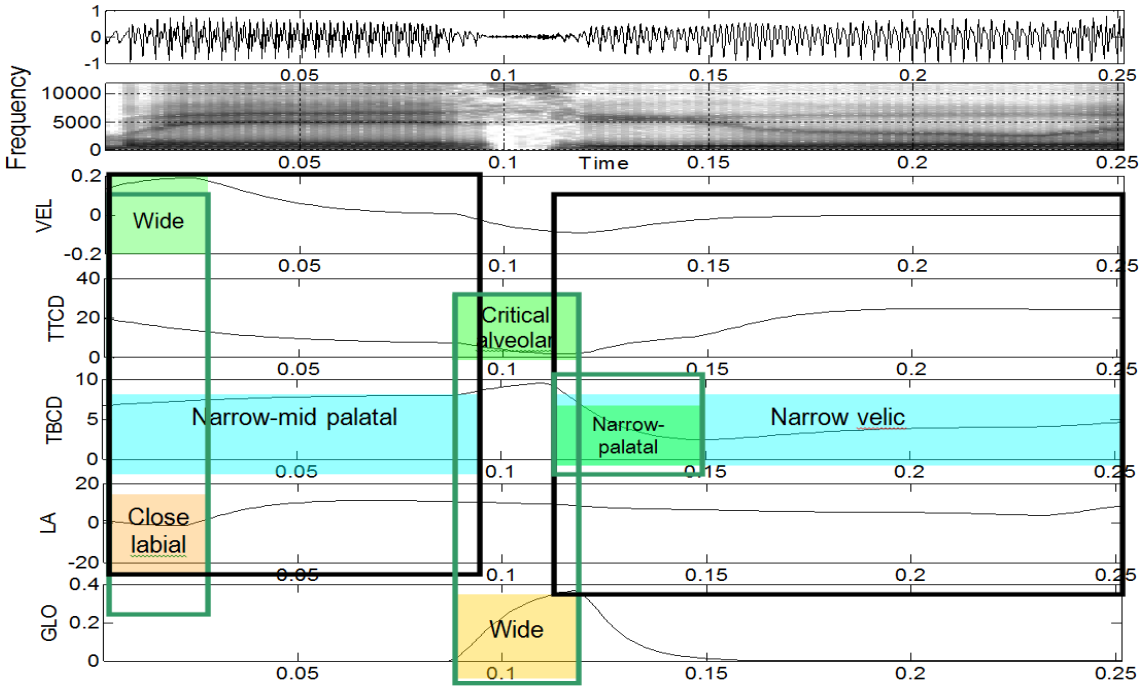


Figure 3.3: Gestures and TVs for the utterance "miss you". Active gestures are marked by colored blocks and the corresponding TVs are smooth curves

along the vocal tract are sensitive to the anatomy of the speaker and head pose variations. Although head poise variations can be corrected in post processing, the anatomical differences like tongue length and shape, vocal tract length, lip aperture etc. persist in the pellet trajectories. TVs are relative measures that determine the location and degree of constrictions performed by the articulators. Such relative measures are robust to anatomical variations and are more directly related to the targets of speech production tasks. Another advantage of converting pellet data to TVs is that we can define articulatory gestures as described in Articulatory Phonology (Browman and Goldstein, 1992). Although, the TVs obtained from real articulatory data look different from synthetic TVs, we can still observe the same phenomena of coarticulation as temporal overlap of consecutive gestures and deletion and reduction as gestures with unfulfilled articulatory targets.

3.4 The X-ray microbeam (XRMB) articulatory dataset

The Wisconsin X-ray Microbeam (XRMB) database (Westbury, 1994) consists of naturally spoken utterances – isolated sentences and short paragraphs. The speech data was collected from 32 males and 25 female subjects along with X-ray microbeam cinematography of the mid-saggital plane of the vocal tract with pellets placed at points as shown in Figure 3.4. The trajectory data are recorded for the individual articulators: Upper Lip, Lower Lip, Tongue Tip, Tongue Blade, Tongue Dorsum, Tongue Root, Lower Front Tooth (Mandible Incisor), Lower Back Tooth (Mandible Molar). We call these trajectories as pellet trajectories. A common

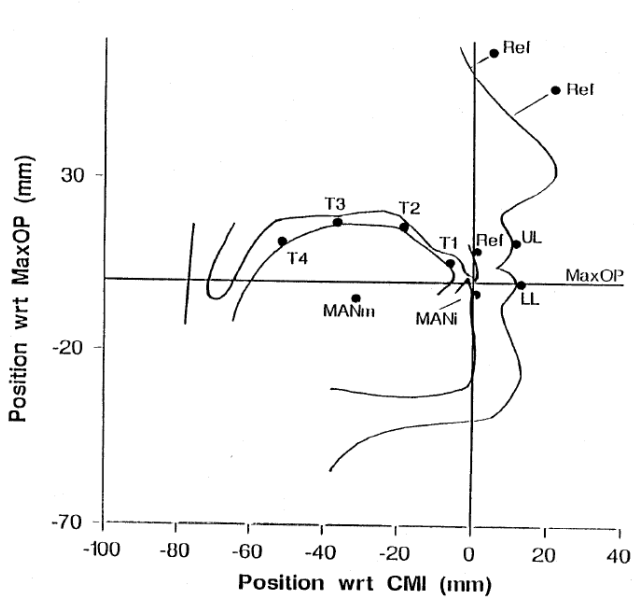
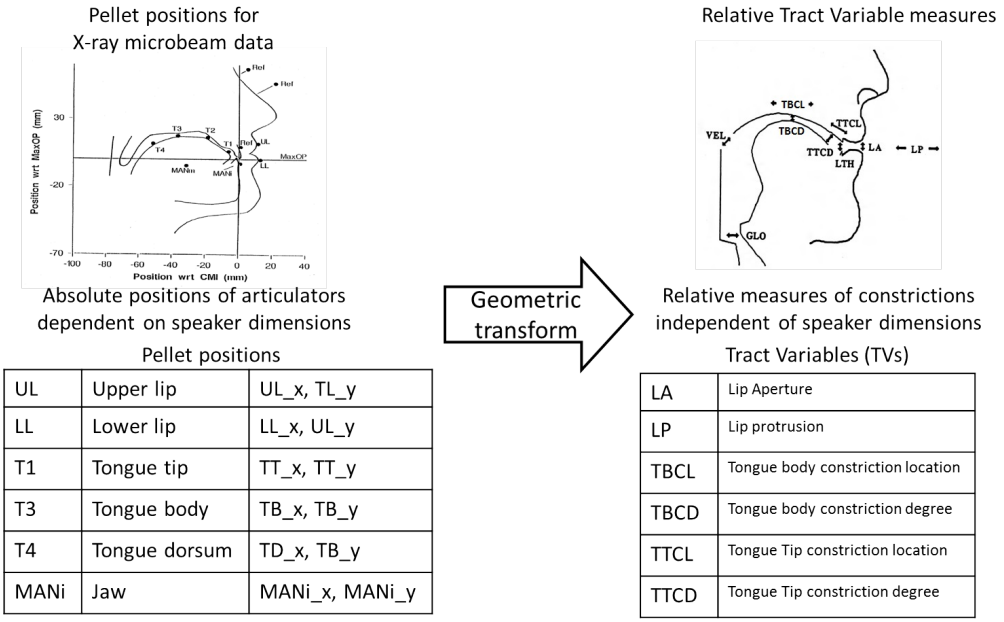


Figure 3.4: Positions of pellets in the XRMB database (Westbury, 1994)

problem with articulatory recordings of this type is the mistracking of pellets or the pellets falling off while recording. Such problems were encountered in the XRMB recordings and were marked as mistracked segments. These segments were removed from the database before using it for our analysis.

3.4.1 Converting XRMB pellets to Tract Variables (TVs)

The X-Y positions of the pellets are closely tied to the anatomy of the speakers. Speech production involves the shaping of the vocal tract by producing constrictions at different places along the vocal tract using the articulators. Hence, the quantification of the vocal tract shape is better performed by the location and degree of these constrictions than the X-Y positions of the pellets. The TVs are a relatively speaker independent representation of articulations. They also provide us a theoretical framework to analyze speech production with the theoretical framework



Pellet positions

UL	Upper lip	UL_x, TL_y
LL	Lower lip	LL_x, UL_y
T1	Tongue tip	TT_x, TT_y
T3	Tongue body	TB_x, TB_y
T4	Tongue dorsum	TD_x, TB_y
MANi	Jaw	MANi_x, MANi_y

Tract Variables (TVs)

LA	Lip Aperture
LP	Lip protrusion
TBCL	Tongue body constriction location
TBCD	Tongue body constriction degree
TTCL	Tongue Tip constriction location
TTCD	Tongue Tip constriction degree

Figure 3.5: Schematic of transformation of XRMB database from pellets to TV trajectories.

of articulatory phonology. Hence, the pellet trajectories were converted to TV trajectories using geometric transformations as outlined in (Mitra et al., 2012). Thus the transformed XRMB database consists of 21 males and 25 females, with a total of 4 hour hours of speech data with corresponding 6 TV trajectories. The TVs obtained from the seven pellet trajectories were – Lip Aperture (LA), Lip Protrusion (LP), Tongue Body Constriction Location (TBCL), Tongue Body Constriction Degree (TBCD), Tongue Tip Constriction Location (TTCL) and, Tongue Tip Constriction Degree (TTCD). A rough schematic of the transformation is shown in Figure 3.5

3.5 Deep Neural Network based approach to speech inversion

Section outlined different approaches to acoustic-to-articulatory speech inversion. Based on the comparison of the different machine learning algorithms

(Mitra et al., 2010a), we chose Artificial Neural Networks (ANN) to be the best suited approach for estimating TVs from speech. This is a function mapping approach to speech inversion where the frame wise input acoustic features are mapped to frame wise measurements of TVs which represent the instantaneous configuration of the vocal tract. With the advent of Deep Neural Networks (DNN), faster learning strategies and higher computational power, it has been shown that deep architectures can represent certain families of functions more efficiently than shallow ones (Bengio and Lecun, 2007). Hence we explore feedforward DNNs for learning the mapping from acoustics to TVs.

A DNN can have M inputs and N outputs; hence, a nonlinear complex mapping of M vectors into N different functions can be achieved. In such an architecture, the same hidden layers are shared by all N outputs, giving the DNN the implicit capability to exploit any correlation that the N outputs may have amongst themselves. The feed-forward DNN used in our study to estimate the TVs from speech were trained with back propagation using a stochastic gradient descent algorithm.

The system shown in Figure 3.6 outlines the blocks involved in the speech inversion system design. The details of the speech inversion system are given in the next few subsections.

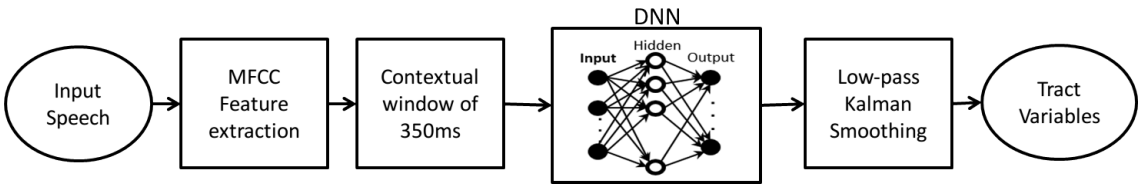


Figure 3.6: Block diagram of the speech inversion system

3.5.1 Data preparation

The XRMB data was used to train the neural networks for the speech inversion. The dataset used for this thesis consists of the XRMB utterances with the time aligned TVs namely – Lip Aperture (LA), Lip Protrusion (LP), Tongue Body Constriction Location (TBCL), Tongue Body Constriction Degree (TBCD), Tongue Tip Constriction Location (TTCL) and the Tongue Tip Constriction Degree (TTCD). The TVs were obtained from the pellet trajectories using the method described (Nam et al., 2012). The Glottis (GLO) and Velum (VEL) tract variables are not present in the dataset because the XRMB data does not contain any measurement of the positions of vocal cords and velum. All non-speech silences were removed from the XRMB data using the phone transcriptions obtained from phone alignments. The XRMB data has many mistracked segments due to pellet fall and tracking errors. All files containing mistracked segments in the middle of utterances were removed. Files having mistracking in the beginning and/or end of utterances were truncated to remove the mistracked segments. After all the preparation, the usable XRMB dataset contained 4 hours of data from 46 different speakers.

3.5.2 Feature extraction

The utterances were downsampled to 8 kHz. The input features to the neural network were varied and compared. We experimented with different acoustic features – MFCC, Perceptual Linear Prediction (PLP) and mel-spectrum (MELSPECT). Single hidden layer neural networks to estimate TVs were trained for each feature type and the best performing feature was chosen for fine tuning. For MFCCs, 13 cepstral coefficients were extracted using a Hamming analysis window of 20ms with a frame rate of 10ms. The TVs and MFCCs were mean and variance normalized to have zero mean and a variance of 0.25. Two different methods of mean and variance normalization were performed and compared. The mean and variance normalization was performed separately for every speaker in the database. This ensured some normalization of inter-speaker variations in measurements of acoustics and articulations. The MFCCs were then contextualized by concatenating every other feature frame within a 350ms window. This amounted to 8 frames of MFCCs on either side of each frame being concatenated to form the contextualized MFCC features. While splicing the frames, we skipped two frames, thus concatenating every other frame within a 35 frame window centered at the current analysis frame. The experiments with other features were performed by adding the same amount of context as for MFCCs.

3.5.3 DNN Training

For the ANN-based TV estimator, the input dimension was 221 for MFCC features (= 13 MFCCs x 17 frames) and the output dimension was 6 (= 6 TVs). The speakers in the dataset were split into train, development and test sets. 36 speakers were assigned for training, and 5 each for development and test sets. The splitting of speakers was random such that the training set consisted of no more than 80% of the utterances and the test and development sets contained nearly an equal number of utterances because the number of utterances from each speaker is not the same due to mistracked segments. A three hidden layer neural network was trained. First, a DNN with 1024 neurons in each hidden layer was trained with different acoustic features as inputs. The best performing feature on the XRMB cross-validation set was selected and then the network parameters like number of hidden layers and number of neurons in each layer were tuned. Networks with different numbers of hidden-layer neurons (128 to 1024) were trained, and among them the best performing network on the cross-validation set was chosen. It was observed that the outputs of the neural network were not as smooth as the original TVs. TVs being vocal tract movements are necessarily smooth signals. Hence, a low-pass Kalman smoothing was performed to remove estimation noise by the neural network. The performance of the TV estimator was measured by computing the Pearson Product Moment Correlations (PPMC) of the estimated TVs with the groundtruth TVs on the test set.

$$r = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_1^n (x_i - \bar{x})^2}\ \sqrt{\sum (y_i - \bar{y})^2}} \tag{3.1}$$

The Kalman smoothed TVs showed high correlation with the original TVs and lower mean squared error (MSE).

3.5.4 Results of speaker independent speech inversion

As described in the previous section, 3 hidden layer neural networks with 1024 neurons each were trained to estimate TVs using three different types of acoustic features. The acoustic features we considered for our experiment were Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), and Mel-Spectrogram (MELSPECT) features. The MFCC and PLP features were 13 dimensional cepstral coefficients per frame. The MELSPECT feature contained 40 mel-filterbank energies for every frame. For each of these features, the analysis frame width was 20ms and the shift was 10ms. The input features were contextualized by concatenating 8 frames on either side. The results on the XRMB cross validation set from these experiments are presented in Table 3.2. The results are Pearson correlations between actual and estimated TVs.

Based on the results shown in Table 3.2, the TV estimator performed best with MFCCs. As a result, MFCCs were used for all further experimentation. We next focused on tuning the DNN parameters for the MFCC feature based speech inversion system. We trained DNNs with 1, 2, 3, 4, and 5 hidden layers with 128, 256, 512, 1024, and 2048 neurons in each layer. Thus we trained 25 such DNNs for mapping contextualized MFCCs to TVs. We computed the correlation between actual and estimated TVs for the cross-validation set and selected the best

Table 3.2: TV estimation Correlation results for different input features

	MFCC	**PLP**	**MELSPECT**
LA	0.799	0.790	0.748
LP	0.670	0.654	0.633
TBCL	0.874	0.866	0.826
TBCD	0.749	0.746	0.671
TTCL	0.765	0.753	0.681
TTCD	0.864	0.867	0.809
Average	**0.787**	**0.779**	**0.728**

performing configuration. Figure 3.7 shows the plot of the correlations for different network configurations. Based on the plot, we can see that a 5-layer DNN with 512 nodes in each layer performed the best. The performance of the networks beyond 5 hidden layers saturated and hence we limited our DNN to 5 hidden layers.

Table 3.3 shows the correlation results for different modes of mean and variance normalizations. In the global mean (ALLNORM) and variance normalization scheme, all the MFCCs and TVs from the XRMB database were normalized with the global mean and variance estimated from all the utterances. In the speaker-specific normalization approach, the MFCCs and TVs were mean and variance normalized separately for each speaker. The correlation results comparing both these normalization approaches are shown in 3.3.

From Table 3.3, it can be inferred that the speaker-specific mean and variance normalization performs better. This makes sense because each speaker's acoustic

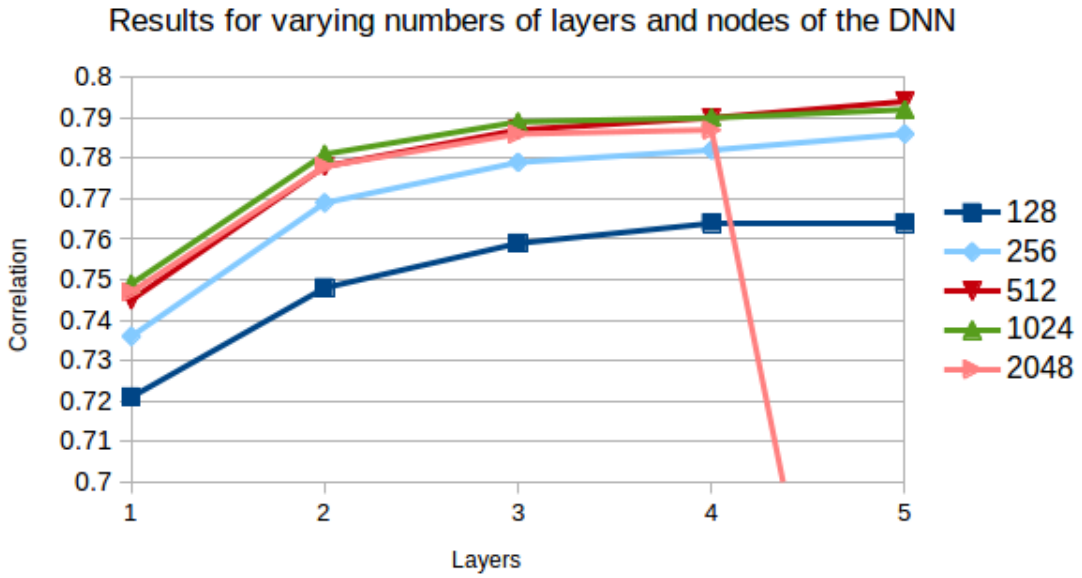


Figure 3.7: Results of varying DNN parameter (layers and number of nodes) on XRMB cross-validation set

Table 3.3: Comparison of mean-variance normalization techniques

	ALLNORM	**SPKNORM**
LA	0.639	0.799
LP	0.818	0.670
TBCL	0.731	0.874
TBCD	0.605	0.749
TTCL	0.529	0.765
TTCD	0.863	0.864
Average	**0.698**	**0.787**

cluster might be centered at a different mean value with a different variance. In the TV domain, there might still be a significant amount of speaker variation due to differences in mean articulatory positions, length of the tongue, range of LA, LP etc.

Thus, normalizing separately for each speaker suppresses the anatomical variations in the TVs.

After performing the fine tuning of the speech inversion system the final best performing neural network architecture was a 5 hidden layer DNN with 512 nodes in each layer. The feature and target normalization chosen was SPKNORM. We will call this speech inversion system as XRMB TV estimator (alternatively, as XRMB speech inversion system) and will be used for various other experiments in the upcoming chapters. The Pearson correlation results of the XRMB speech inversion system are shown in Table 3.4. An example plot of the estimated and actual TVs for an utterance from the XRMB test set has been shown in Figure 3.8

Table 3.4: Correlation results for the final XRMB speech inversion system

	LA	**LP**	**TBCL**	**TBCD**	**TTCL**	**TTCD**	**Average**
Crossval set	0.809	0.678	0.873	0.761	0.769	0.877	**0.794**
Test set	0.856	0.613	0.866	0.745	0.707	0.907	**0.782**

3.6 Discussion

As seen in Figure 3.8, the estimated TVs agree with the actual TVs on most places except a few regions where they are incorrect. We observed that the estimated TVs were more accurate for the phonemes where the concerned articulator is critical for the phoneme's production. This is due to the critical articulator phenomenon due to which movements of articulators that are critical for the production of a phoneme are more precise (less variable) compared to those that are not critical for

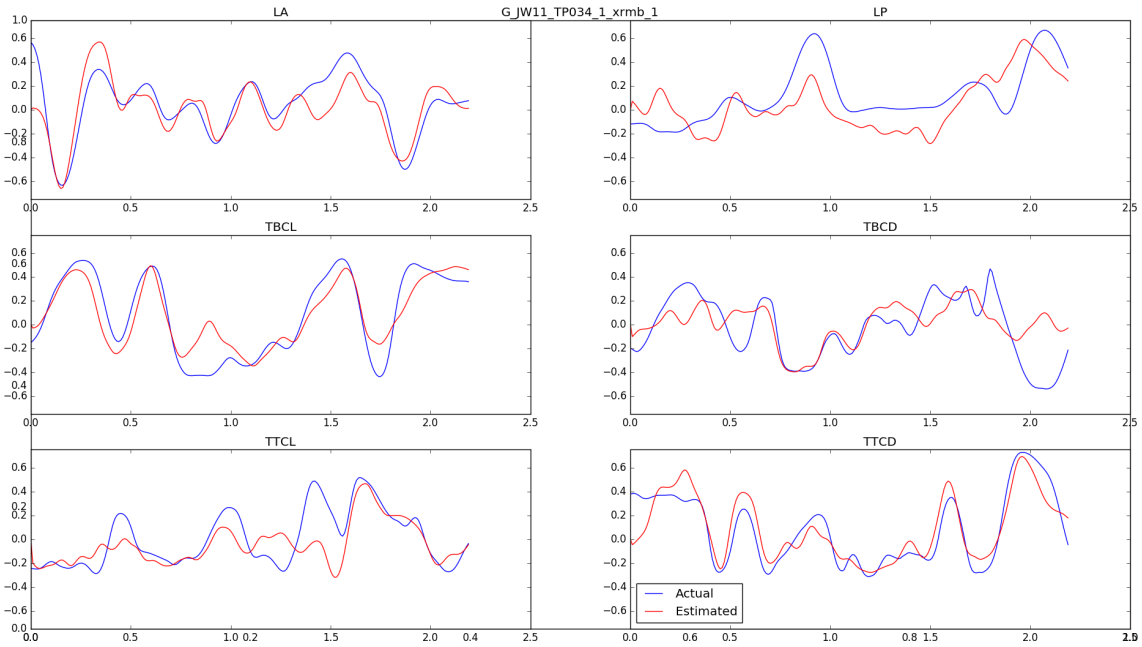


Figure 3.8: Example plot of estimated (red) and actual (blue) TVs for a test set utterance - "Combine all the ingredients in a large bowl"

the production of the phoneme. This is one of the reasons that limit the performance of the speech inversion systems. The variability in the non-critical articulators is compounded when the articulatory data contains multiple speakers.

3.7 Summary

This section discusses the details of the development of a speaker independent speech inversion system on the XRMB dataset. We explored several acoustic features for the speech inversion problem and found that MFCCs are the best acosutic features for learning the acoustic-to-articulatory mapping. We also performed fine tuning of the DNN parameters for the speech inversion using MFCCs as front ends. This speech inversion syhstem, trained on 36 speakers from the XRMB dataset is the first known multi-speaker speaker independent speech inversion

system. The network thus trained can be used to estimate the articulatory features for any speech utterance. A user friendly tool has been developed in Python to easily extract the TVs for any speech waveform of interest using this trained DNN. It can be downloaded from https://github.com/ganesa90/speech-inversion-dnn. A real-time version of the speech inversion system was also developed for a demo and has been made available for download - https://github.com/ganesa90/speech_inversion_rt

Chapter 4

Speech inversion performance across speech variability

4.1 Overview

In this chapter we consider different variabilities of speech and evaluate the performance of speech inversion under those variabilities

4.2 Speaker Variability

4.2.1 Cross speaker performance of speaker dependent systems

It was observed that the acoustic and articulatory variability across speakers was affecting the performance of the speaker independent speech inversion. In order to explore the speaker variability, speaker dependent systems were trained on 10 speakers (5 males and 5 females). The correlation results for the speaker dependent systems for the 10 chosen speakers are shown in Table 4.1 Comparing the numbers from Table 3.3 and Table 4.1, we observe that a speaker dependent speech inversion system is more accurate compared to a speaker independent system. However, the performance of the speaker dependent systems across speakers is mediocre. We tested each speaker dependent system using the test sets of the remaining 9 speakers. Table 4.2 shows the average correlation across the 6 TVs for the cross speaker tests performed on the speaker dependent systems. The cross speaker

test correlations in Table 4.2 highlight the inter-speaker variability of the acoustic and articulatory spaces. The cross-speaker performance of the speaker-dependent systems also showed a clear trend of gender dependence where the female models performed better on female test sets as opposed to the male models.

Thus, one approach to improve the performance of the speaker independent inversion system is to take advantage of data from multiple speakers and domains and perform speaker normalization. The upcoming sections will discuss a speaker normalization approach for speech inversion.

Table 4.1: Correlation results for speaker dependent speech inversion systems

	Spk ID	**Gender**	**LA**	**LP**	**TBCL**	**TBCD**	**TTCL**	**TTCD**	**Average**
Spkr 1	**JW12**	M	0.837	0.821	0.908	0.828	0.792	0.905	**0.848**
Spkr 2	**JW14**	F	0.826	0.698	0.927	0.840	0.864	0.902	**0.843**
Spkr 3	**JW24**	M	0.824	0.769	0.907	0.773	0.764	0.827	**0.811**
Spkr 4	**JW26**	F	0.814	0.825	0.908	0.785	0.804	0.900	**0.839**
Spkr 5	**JW27**	F	0.795	0.796	0.878	0.774	0.733	0.893	**0.811**
Spkr 6	**JW31**	F	0.851	0.782	0.922	0.850	0.809	0.906	**0.853**
Spkr 7	**JW40**	M	0.779	0.551	0.906	0.749	0.833	0.869	**0.781**
Spkr 8	**JW45**	M	0.834	0.785	0.896	0.804	0.845	0.866	**0.838**
Spkr 9	**JW54**	F	0.758	0.529	0.879	0.760	0.884	0.848	**0.776**
Spkr 10	**JW59**	M	0.806	0.769	0.909	0.806	0.815	0.882	**0.831**

4.2.2 Speaker Normalization to combat acoustic variability

The mapping from acoustics to articulations is known to be highly non-linear and non-unique (Qin and Carreira-Perpiñán, 2007). Adding speaker variability

Table 4.2: Cross speaker correlation results for speaker dependent speech inversion systems

Train \ Test	Spkr1 (M)	Spkr 2 (F)	Spkr 3 (M)	Spkr 4 (F)	Spkr 5 (F)	Spkr 6 (F)	Spkr 7 (M)	Spkr 8 (M)	Spkr 9 (F)	Spkr10 (M)
Spkr 1 (M)	**0.847**	0.539	0.617	0.512	0.384	0.288	0.540	0.596	0.535	0.602
Spkr 2 (F)	0.548	**0.844**	0.494	0.637	0.577	0.486	0.335	0.437	0.620	0.487
Spkr 3 (M)	0.655	0.558	**0.808**	0.480	0.319	0.257	0.548	0.605	0.519	0.654
Spkr 4 (F)	0.539	0.593	0.457	**0.837**	0.585	0.576	0.314	0.409	0.602	0.412
Spkr 5 (F)	0.429	0.589	0.304	0.596	**0.813**	0.658	0.142	0.169	0.526	0.320
Spkr 6 (F)	0.310	0.490	0.183	0.553	0.606	**0.852**	0.116	0.114	0.386	0.132
Spkr 7 (M)	0.518	0.323	0.473	0.306	0.195	0.142	**0.784**	0.559	0.375	0.459
Spkr 8 (M)	0.600	0.487	0.626	0.378	0.197	0.050	0.586	**0.838**	0.522	0.633
Spkr 9 (F)	0.591	0.612	0.484	0.626	0.550	0.444	0.425	0.484	**0.777**	0.519
Spkr 10 (M)	0.616	0.527	0.613	0.441	0.313	0.208	0.505	0.612	0.503	**0.829**

to the already challenging problem makes it even more difficult. Most research in speech inversion has been focused on developing accurate speaker dependent systems. Approaches like codebook search (Atal et al., 1978), feedforward neural networks (Mitra et al., 2010a), and Mixture Density Networks have been found to work well for speaker dependent speech inversion. There have been a few attempts to perform speaker independent speech inversion (Afshan and Ghosh, 2015) (Ji, 2014) which have been limited to two speakers from the MOCHA TIMIT dataset (Wrench and Richmond, 2000). Hueber et al. (Hueber et al., 2015) presents a Gaussian mixture regression based speaker adaptation scheme for a Gaussian Mixture Model (GMM) based speech inversion system. However, there has not to date been any effort in performing speaker adaptation for artificial neural network based speech inversion systems. This section presents a Vocal Tract Length Normalization

(VTLN) based approach to speaker adaptation for speech inversion. VTLN is a popular speaker adaptation technique in ASR which has so far not been applied to speech inversion.

Vocal Tract Length Normalization (VTLN) (Eide and Gish, 1996) using a piecewise linear warping function is a commonly adopted approach for speaker adaptation in speech recognition. We applied VTLN in a maximum likelihood framework to adapt the acoustic features of the mismatched speakers to the target speaker. In order to perform VTLN, a speaker dependent acoustic space using Gaussian Mixture Models (GMM) was trained for all the 10 speakers.

The experiments in this paper are performed on a set of 10 speakers from the U. Wisconsin X-ray Microbeam (XRMB) database (Westbury, 1994). The articulatory features are represented by six tract-variable (TV) trajectories (described below). Using a leave-one-out methodology, separate experiments were performed for each speaker in which the acoustic features from the other 9 speakers were transformed using the VTLN approach. The transformed acoustic features were then used to train a speech inversion system. The performance of the system trained on VTLN adapted acoustic features was compared to the performance of speaker dependent systems. The performances of the individual systems were compared using the correlation between the estimated and the actual TVs on the target speaker's test set. More details of the speech inversion system training and the experiments are provided in the upcoming sections

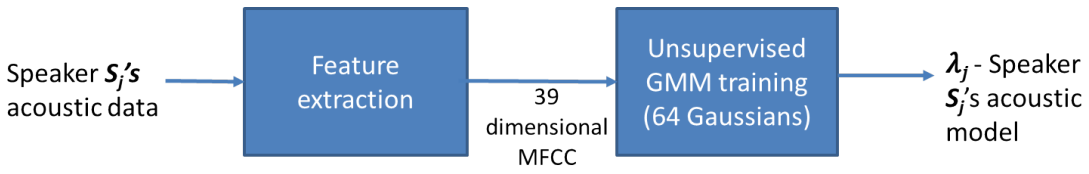


Figure 4.1: Training of GMM speaker acoustic spaces

4.2.2.1 Speaker acoustic spaces

Thirteen dimensional MFCCs with slope and acceleration were used as acoustic features for modeling the speaker acoustic spaces. Gaussian Mixture Models (GMMs) with 64 Gaussian components were trained on the 39 dimensional MFCC features. The diagonal covariance GMMs were trained iteratively by increasing the number of Gaussians from 2 to 64 by doubling the number of components in each stage. The GMM training routines were obtained from the MSR Identity Toolbox v1.0 (Sadjadi et al., 2013). Thus, such GMMs were trained for each of the 10 speakers chosen for the cross-speaker evaluation. Figure 4.1 shows the block diagram of the system used to train unsupervised speaker acoustic spaces. The training is unsupervised becausse we don't use any kind of phone alignments for training phone-wise GMM like in HMM based ASR. Instead we let the GMMs fit the distribution of the acoustic features for each speaker. A visualization of the speaker acoustic spaces is shown in Figure 4.2. Each model λ_i is a 64 component GMM modeling the distribution of MFCCs for speaker S_i.

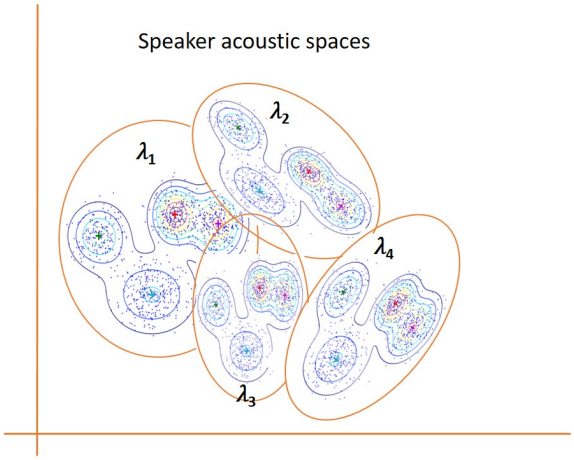


Figure 4.2: A schematic representation of the speaker acoustic spaces

4.2.2.2 Maximum Likelihood based VTLN

Vocal Tract Length Normalization (VTLN) aims to compensate the effects of different vocal tract lengths by warping the frequency spectrum in the filterbank analysis before the computation of the cepstral coefficients. This warping can be implemented by a simple piecewise linear warping function as shown in Figure 4.3. The warping factor determines the nature of the warping function. The warping is implemented between the lower boundary of frequency analysis (LOFREQ) and the upper boundary of frequency analysis (HIFREQ). In order to adapt the acoustic features of speaker S_i to speaker S_j, a single warping factor α_{ij} is used for all utterances from speaker S_i. The warping factor ij is determined by a maximum likelihood approach as outlined below.

Let the GMM acoustic model for speaker S_j be λ_j, and the warped acoustic features for the t^{th} time frame of an utterance of speaker S_i to the target speaker

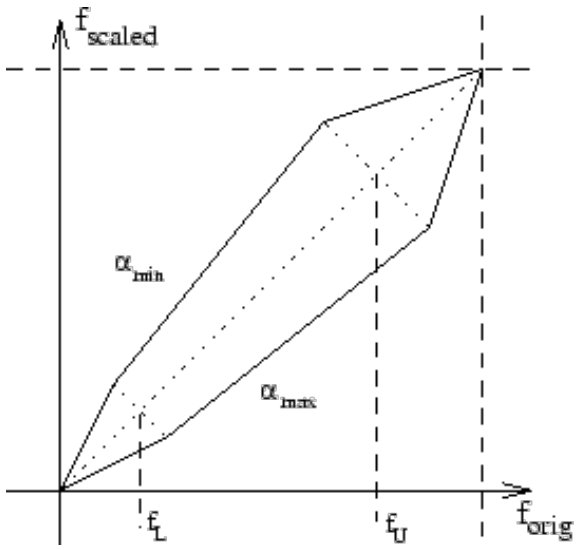


Figure 4.3: Frequency warping function implemented in HTK toolkit (Young et al., 2009)

S_j be $x_{ij}(t)$. Then, the most likely warping factor α_{ij} is given by-

$$\alpha_{ij} = \arg\max_{\alpha} \sum_{t=1}^{N} \log(P(x_{ij}(t)|\lambda_j, \alpha)) \tag{4.1}$$

In the equation 4.1 $\sum_{t=1}^{N} \log(P(x_{ij}(t)|\lambda_j, \alpha))$ is the log likelihood of the transformed features of speaker S_i with respect to speaker S_j's acoustic model. The optimal α_{ij} is obtained by sweeping the value of α_{ij} from 0.8 to 1.2 in steps of 0.025. Using the optimal α_{ij}, we compute the speaker adapted acoustic features for speaker S_i adapted to speaker S_j.

4.2.2.3 Speech inversion system

We trained speech inversion systems using a single hidden layer feed-forward neural network. Since only small amounts of data were available for each speaker, single hidden layer networks were chosen as the architecture. The inputs to the

neural network were the 13 dimensional MFCCs contextualized with MFCC features from 8 frames on either side. Thus, the input dimension was 13x17 = 221. The outputs of the network were six dimensional TVs. We trained networks with 100, 200, 300, 400 and 500 nodes in the hidden layer and selected the best performing network based on performance on the test set. The outputs of the trained neural network were found to be noisy. The outputs were smoothed using a Kalman smoothing technique to obtain smooth TV estimates. 3.6 shows the block diagram of our speech inversion system.

4.2.2.4 Experiments

Speaker transformed datasets: Using the VTLN method described in Section 4.2.2.2, each speaker's data was transformed to each of the other 9 speakers' data. Thus, for each speaker, we have 10 sets of data – 1 from the speaker and other 9 transformed to the target speaker from the other 9 speakers using VTLN. The following figure shows the schematic of the transformation procedure for transforming data from speakers $S_b ... S_j$ to speaker S_a's acoustic space to create the transformed datasets $S_{ba} ... S_{ja}$. In this way, we created 90 transformed datasets tailored to each of the 10 speakers' acoustic spaces. 4.4 shows the schematic of the procedure adopted to create the speaker transformed datasets.

Speech inversion systems trained on speaker transformed datasets: We trained four types of speech inversion systems for each speaker as described in Section 4.2.2.3. The following are the descriptions of the different inversion systems trained.

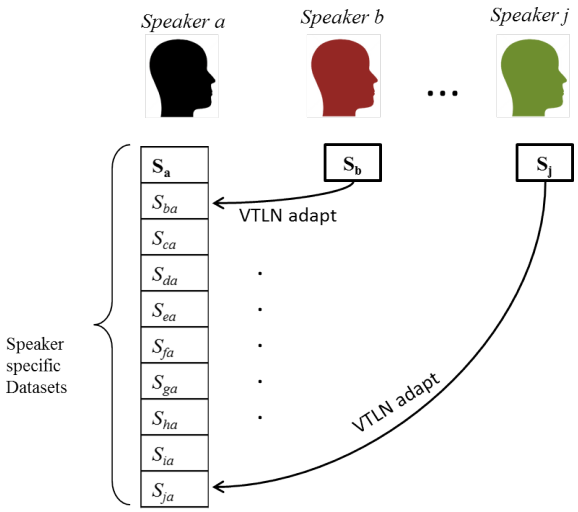


Figure 4.4: Schematic of speaker transformed datasets creation

• SD: 10 Speaker Dependent (SD) speech inversion systems.

• *Sys1*: For each speaker, data from other 9 speakers were randomly chosen to match the amount of data from the target speaker and an inversion system was trained. In total, 10 such systems were trained. For example, for speaker 'a', data from $S_b ... S_j$ was randomly sampled to match the amount of data in S_a

• *Sys2*: For each speaker, VTLN transformed data from other 9 speakers were randomly chosen to match the amount of data from the target speaker and an inversion system was trained. In total, 10 such systems were trained. For example, for speaker 'a', data from $S_{ba} ... S_{ja}$ was randomly sampled to match the amount of data in S_a

• *Sys3*: For each speaker, data from the target speaker and the VTLN transformed data from other 9 speakers were randomly chosen to match the amount of data from the target speaker and an inversion system was trained. In total, 10 such systems were trained. For example, for speaker 'a', data from S_a, $S_{ba} ... S_{ja}$

was randomly sampled to match the amount of data in S_a. The difference between System3 and System2 is that System3 has some of the target speaker's data in the training set.

In total, 40 speech inversion systems were trained. In the above described systems, the amount of training data for each system was kept the same in order to have a fair comparison with the SD system. However the transformed data available for each target speaker was about 10 times more because of the other 9 speakers' data put together. We created versions of Systems 1, 2, and 3 using all the transformed data. We call these systems *Sys1_alldata*, *Sys2_alldata*, and *Sys3_alldata*.

4.2.3 Results of Speaker Normalization experiments

For each speaker, a test set containing 10% of the speaker's data was created which was kept separate from all the speech inversion training and VTLN procedure. Each of the systems SD, System1, 2, and 3 were evaluated on each speaker's test set. The Pearson product Moment Correlation (PPMC) was computed between the actual and estimated TVs. Table 4.3 shows the correlation results of all the speech inversion systems across all speakers. The numbers show correlation values averaged across all 6 TVs. The correlation for LP tract variable is the least and that for TBCL is the highest. The performance of *Sys1* is very poor compared to SD because the training dataset for this system consists of a small number of utterances from multiple speakers. Transforming the data from the other 9 speakers to the target speaker's acoustic space using the proposed VTLN approach provides an average

of 7% absolute improvement in correlation over *Sys1*. The amount of improvement in correlation varies across all speakers. Some speakers like JW14 and JW24 show marginal or no improvement in the performance, whereas for JW31 we see a large 13% improvement. In order to see the influence of speaker specific training data on the performance, we created *Sys3* which contained a part of the target speaker's training set data. The overall amount of training data for *Sys3* was kept same as the amount of training data available for each target speaker. This provided an average of 3% improvement in correlation compared to *Sys2*. However, the correlations of *Sys3* were still 13% below the average correlation of the SD systems. Figure 4 shows the plots of the estimated and actual TVs for a randomly selected test utterance from speaker JW26's test set. Table 4.4 shows the correlation results for the speech inversion systems trained with all the available data from the other 9 speakers. These are the systems *Sys1_alldata*, *Sys2_alldata* and *Sys3_alldata* as described in section 4.2.2.4. We observe that the results are much better than those in Table 4.3. The performance gain obtained by performing the VTLN adaptation is around 4% on an average above the correlation results of *Sys1_alldata*. It is interesting to observe that adding all the training data of the target speaker, as done in the training of *Sys3_alldata* provides a system that performs as well as the speaker dependent SD systems. This demonstrates that adding VTLN adapted data from multiple speakers does not degrade the performance of the speaker dependent systems.

In addition to training different speaker adapted speech inversion systems, we also evaluated the VTLN based speaker adaptation approach by applying the adaptation to cross-speaker tests. We evaluated each speaker dependent (SD) speech

Table 4.3: Correlation results of SD, Sys1, Sys2, and Sys3 for all speakers

Speech inversion system	Average amount of training data (mins)	Spk 1	Spk 2	Spk 3	Spk 4	Spk 5	Spk 6	Spk 7	Spk 8	Spk 9	Spk 10	Average
		JW12	JW14	JW24	JW26	JW27	JW31	JW40	JW45	JW54	JW59	
		M	F	M	F	F	F	M	M	F	M	
SD	5.68	0.848	0.843	0.811	0.839	0.811	0.853	0.781	0.838	0.776	0.831	**0.823**
Sys1	5.68	0.669	0.659	0.608	0.631	0.556	0.507	0.560	0.615	0.635	0.642	**0.608**
Sys2	**5.68**	**0.714**	**0.656**	**0.630**	**0.708**	**0.627**	**0.635**	**0.648**	**0.668**	**0.656**	**0.697**	**0.664**
Sys3	5.68	0.738	0.699	0.715	0.738	0.660	0.708	0.583	0.685	0.687	0.717	**0.693**

Table 4.4: Correlation results of SD, Sys1_alldata, Sys2_alldata, and Sys3_alldata for all speakers

Speech inversion system	Average amount of training data (mins)	Spk 1	Spk 2	Spk 3	Spk 4	Spk 5	Spk 6	Spk 7	Spk 8	Spk 9	Spk 10	Average
		JW12	JW14	JW24	JW26	JW27	JW31	JW40	JW45	JW54	JW59	
		M	F	M	F	F	F	M	M	F	M	
SD	5.68	0.848	0.843	0.811	0.839	0.811	0.853	0.781	0.838	0.776	0.831	**0.823**
Sys1_alldata	5.68	0.712	0.731	0.703	0.716	0.676	0.611	0.652	0.706	0.691	0.718	**0.692**
Sys2_alldata	**5.68**	**0.755**	**0.748**	**0.736**	**0.773**	**0.710**	**0.698**	**0.709**	**0.730**	**0.714**	**0.753**	**0.733**
Sys3_alldata	5.68	0.819	0.803	0.793	0.830	0.776	0.809	0.790	0.806	0.782	0.817	**0.802**

inversion system on the test sets of the other 9 speakers. This experiment was similar to the one described in section 4.2.1 but in this case we used the speaker adapted MFCC features of the test speakers instead of their original features. We computed the correlation between the actual and estimated TVs using the Pearson correlation. Table 4.5 shows the average correlations for cross speaker tests after VTLN. We refer the reader to compare and contrast Table 4.5 with Table 4.2 which shows the performance without the VTLN based speaker adaptation.

In order to easily visualize the performance of the speech inversion systems

Table 4.5: Average correlations for cross speaker tests after VTLN. Each row represents correlations of one speaker dependent system on the test sets of other speakers

Train \ Test	Spkr1 (M)	Spkr 2 (F)	Spkr 3 (M)	Spkr 4 (F)	Spkr 5 (F)	Spkr 6 (F)	Spkr 7 (M)	Spkr 8 (M)	Spkr 9 (F)	Spkr10 (M)
Spkr 1 (M)	**0.847**	0.559	0.617	0.659	0.566	0.562	0.598	0.596	0.553	0.588
Spkr 2 (F)	0.605	**0.844**	0.586	0.637	0.593	0.581	0.472	0.557	0.620	0.625
Spkr 3 (M)	0.655	0.604	**0.808**	0.647	0.509	0.590	0.567	0.611	0.592	0.654
Spkr 4 (F)	0.657	0.612	0.625	**0.837**	0.589	0.610	0.563	0.630	0.629	0.668
Spkr 5 (F)	0.587	0.611	0.568	0.596	**0.813**	0.672	0.483	0.453	0.529	0.594
Spkr 6 (F)	0.601	0.586	0.580	0.604	0.639	**0.852**	0.423	0.452	0.508	0.548
Spkr 7 (M)	0.553	0.323	0.468	0.569	0.422	0.424	**0.784**	0.560	0.454	0.483
Spkr 8 (M)	0.600	0.568	0.626	0.677	0.475	0.450	0.584	**0.838**	0.622	0.639
Spkr 9 (F)	0.581	0.608	0.593	0.640	0.559	0.567	0.557	0.624	**0.777**	0.644
Spkr 10 (M)	0.600	0.611	0.607	0.665	0.564	0.575	0.549	0.621	0.619	**0.829**

with and without the speaker adaptation, we have plotted the correlation values as gray-scale colors in Figure 4.5

4.2.4 Summary

Based on the results shown in Tables 4.3 and 4.4, we can conclude that the amount of training data plays a great role in the accuracy of the speech inversion system. Even if the data is from multiple speakers, more data is always good. The VTLN speaker adaptation normalizes multiple speakers' acoustic data to match a target speaker. VTLN provides an average of 7% absolute improvement of correlation (Sys1 to Sys2) on the speech inversion system trained on the 9 speakers' dataset. Adding a small amount of the target speaker's data in the training set

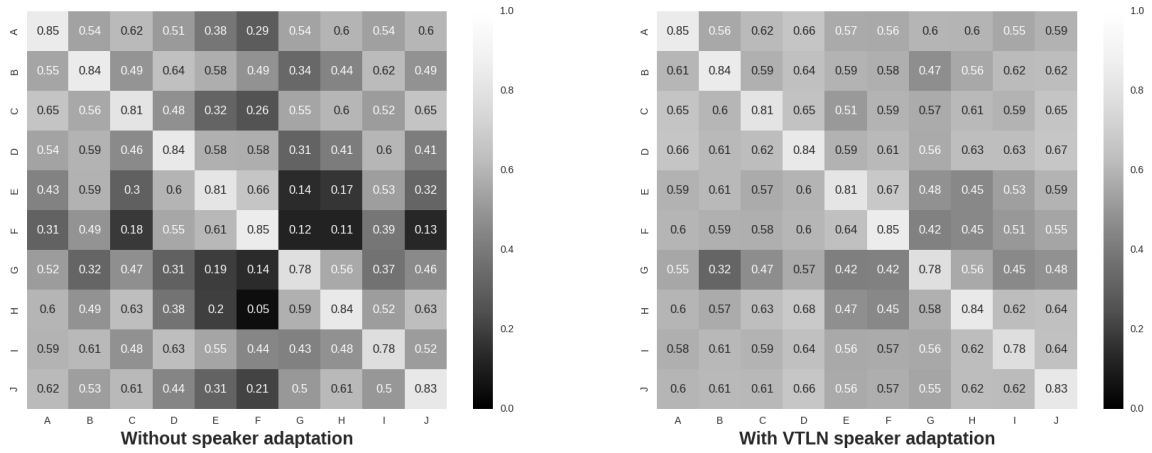


Figure 4.5: Visualization of the cross speaker test correlations. Correlation of 1 corresponds to white and 0 corresponds to black

improves the correlation further by 3% over Sys2. In spite of performing VTLN, the correlation performance of Sys2 trained on the transformed data is 16% poorer than the Speaker dependent system. The systems trained with all data shows that having more training data from multiple speakers can make the speech inversion system better. The accuracy of Sys1_alldata is 10% poorer than SD due to the mismatch between the acoustic spaces of the training speakers and the test speakers. With the VTLN based transformation of the training data, the accuracy improves by 4%. This means our proposed adaptation technique helps reduce the mismatch between the acoustic spaces. Adding all of the target speakers' training data along with the transformed data of the other 9 speakers' does not degrade the speaker dependent performance. This approach of transforming the training data from multiple speakers to create multiple speaker adapted versions can be used to create a model selection based approach to speech inversion. In such a system we will have multiple speaker tuned models and then select the best matching model for a test utterance based on a maximum likelihood speaker matching approach. In

this chapter so far, we have examined the variability of speech and articulations across speakers and developed a speaker adaptation approach to normalize the speaker differences. The experiments in this chapter show that data from multiple speakers can be normalized and combined to create better speaker independent speech inversion systems. This approach can be extended to combine data from different articulatory datasets to create a single improved speech inversion system.

4.3 Variability due to speaking rate

Speaking rate is a very common cause of variability in speech. Conversational speech often involves varying speaking rates which significantly affect the accuracy of ASR systems. The variability induced in the speech due to speaking rate is predominantly due to casual and incomplete articulatory gestures. The effects of speaking rate on speech are not uniformly manifested across all phonemes. Hence it is not possible to study speaking rate effects by performing uniform time-scale modification of speech. This section focuses on the study of speaking rate variability from the perspective of articulatory representations. We collected concurrent acoustic and articulatory data from eight speakers of American English speaking at normal and fast rates. We then trained acoustic -to-articulatory inversion systems to investigate the effects of speaking rate on the performance of the speech inversion systems. Careful cross-speaker and cross speaking rate experiments show the effects of speaking rate on the performance of speech inversion systems and also highlight the acoustic and articulatory variability of speech due to speaking rates.

4.3.1 The EMA-IEEE Articulatory dataset

A 5-D electromagnetic articulometry (EMA) system (WAVE; Northern Digital) was used to record the 720 phonetically balanced Harvard sentences (Rothauser et al., 1969) from eight speakers (4 males, and 4 females) at normal and fast production rates. Participants produced each sentence twice, first at their preferred 'normal' speaking rate followed by a 'fast' production (for a subset of the sentences two normal rate productions were elicited). They were instructed to produce the 'fast' repetition as quickly as possible without making errors. EMA trajectories were obtained at 100 Hz from sensors placed on the tongue (tip (TT), body (TB), root (TR)), lips (upper (UL) and lower (LL)) and mandible, together with reference sensors on the left and right mastoids, and upper and lower incisors (UI, LI). The data were low-pass filtered at 5 Hz for references and 20 Hz for articulator sensors, corrected for head movement and aligned to the occlusal plane. Synchronized audio was recorded at 22050 Hz, using a directional shotgun microphone placed 50 cm from the speaker's mouth. Phone and word labels were placed using the U. Pennsylvania forced aligner (Yuan and Liberman, 2008).

4.3.2 Conversion of EMA sensor positions to TVs

The EMA sensor trajectory data was converted to TVs using geometric transformations. The TVs that we defined are described in Table 4.6. The

transformations were computed using the following equations:

$$LA[n] = \sqrt{(LL_x[n] - UL_x[n])^2 + (LL_z[n] - UL_[n])^2} \tag{4.2}$$

$$LP[n] = LL_x[n] - \underset{m \in allutterances}{\text{median}} \{LL_x[m]\} \tag{4.3}$$

$$JA[n] = \sqrt{(LI_x[n] - UL_x[n])^2 + (LI_z[n] - UL_z[n])^2} \tag{4.4}$$

$$TTCD[n] = \underset{x \in (-50,0)}{\text{Min}} \{Dist(TT, pal(x))\} \tag{4.5}$$

$$TTCL[n] = \underset{m \in allutterances}{\text{median}} \{TT_x[m]\} - TT_x[n] \tag{4.6}$$

The nine TVs were: Lip Aperture (LA), Lip Protrusion (LP), Jaw Angle (JA), Tongue Tip Constriction Location (TTCL), Tongue Tip Constriction Degree (TTCD), Tongue Middle Constriction Location (TMCL) and Tongue Middle Constriction Degree (TMCD), Tongue Root Constriction Location (TMCL) and Tongue Root Constriction Degree (TMCD). LA was defined as the Euclidean distance between the UL and the LL sensors as shown in equation 4.2. LP was defined as the displacement along the x-axis of the LL sensor from its median position as shown in equation 4.3. JA was defined as the Euclidean distance between the UL sensor and the LI sensor as shown in equation 4.4. Two TVs were computed for each tongue sensor - constriction degree and location. Constriction degree for a tongue sensor was defined as the minimum distance between the sensor and the

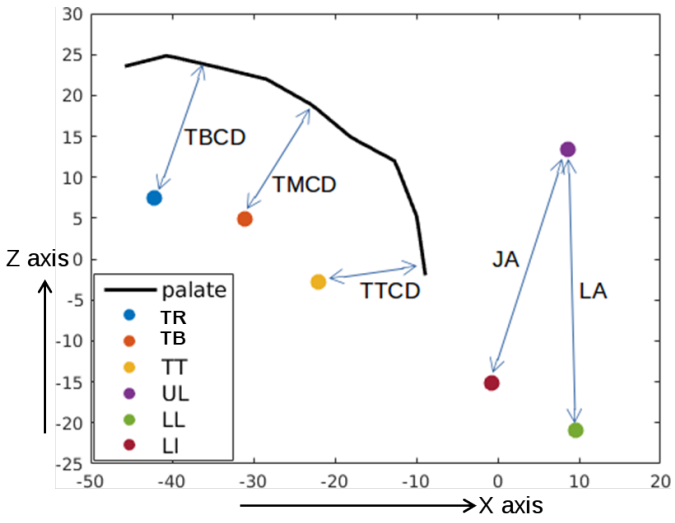


Figure 4.6: Transformation of EMA sensor positions to TVs

palate trace as shown in equation 4.5. This way the TTCD, TMCD, and TBCD TVs were computed from the TT, TM and TB sensor positions and the palate trace. The constriction location for a tongue sensor was defined as the displacement of the sensor along the x-direction from its median position as shown in equation 4.6. Thus, TTCL, TMCL, and TBCL were computed from the TT, TM, and TB sensor positions.

4.3.3 Speech inversion experiments

The EMA articulatory data collected from the 8 subjects; each consisted of 720 utterances produced in fast and normal speaking rates. The total dataset contains 7.05 hours of speech and concurrent articulatory trajectories. The 720 sentences from the IEEE dataset was randomly divided into 3 subsets for training, cross-validation and testing. The training subset contained 576 sentences while the test and cross-validation sets contained 72 sentences each. With this split of

Table 4.6: Definition of TVs from EMA sensors for the EMA-IEEE dataset

Tract Variables	**Meaning**	**EMA Sensors involved**
LA	Lip Aperture	LL, UL
LP	Lip Protrusion	LL
JA	Jaw Angle	LI, UL
TTCL	Tongue Tip Constriction Location	TT, palate
TTCD	Tongue Tip Constriction Degree	TT, palate
TBCL	Tongue Body Constriction Location	TB, Palate
TBCD	Tongue Body Constriction Degree	TB, Palate
TRCL	Tongue Root Constriction Location	TR, Palate
TRCD	Tongue Root Constriction Degree	TR, Palate

sentences, we created train, cross-validation and test sets for each of the 8 speakers and the 2 speaking rates. Thus, we created 8 pairs (Normal and Fast rates) of subsets from the EMA-IEEE dataset. Note that the same set of sentences were used across all speakers for their respective train, cross-validation and test sets.

We trained several DNN based acoustic-to-articulatory speech inversion systems using different combinations of the data subsets. We used 3-hidden layer neural networks as the architecture for all our systems. Based on the amount of training data, we varied the number of nodes in each hidden layer from 256 to 1024. The speech inversion system remained same as the one shown in Figure 3.6. The different speech inversion systems trained using the dataset are described in Table 4.7.

Table 4.8 shows the correlation results of the various speech inversion systems

Table 4.7: Description of various speech inversion systems trained on EMA-IEEE dataset

System Name	**Description**
SD_N	Speaker dependent Normal rate
SD_F	Speaker dependent Fast rate
SD_all	Speaker dependent all utterances
LOO_N	Leave one speaker out Normal rate
LOO_F	Leave one speaker out Fast rate
LOO_all	Leave one speaker out all utterances

trained on the EMA-IEEE dataset. We observe that the accuracy of estimating TVs for fast rate speech is at least 4% lower compared to the normal rate speech. This difference in performance is due to the higher acoustic and articulatory variability of the fast rate speech. When subjects speak fast they tend to be quick in their articulations in order to increase the speaking rate thereby resulting in unmet articulatory targets leading to coarticulations.

We finally trained a speech inversion system using all the training data (both speaking rates) from all the speakers. This system was trained in order to estimate TVs in the future for ASR experiments. We will refer to this system as the EMA-IEEE speech inversion system in the future. Table 4.9 shows the correlation between actual and estimated TVs on the complete test set (test sets of all speakers at both speaking rates) for the EMA-IEEE speech inversion system.

Table 4.8: Average correlation results of various speech inversion systems trained on the EMA-IEEE dataset

	LA	**LP**	**JA**	**TTCL**	**TTCD**	**TBCL**	**TBCD**	**TRCL**	**TRCD**	**Average**
SD_N	0.819	0.771	0.852	0.841	0.862	0.806	0.886	0.770	0.802	**0.823**
SD_F	0.745	0.704	0.809	0.777	0.793	0.742	0.815	0.716	0.714	**0.757**
SD_all	0.788	0.739	0.829	0.810	0.827	0.772	0.861	0.749	0.760	**0.793**
LOO_N	0.713	0.627	0.778	0.708	0.771	0.675	0.786	0.625	0.616	**0.700**
LOO_F	0.656	0.562	0.730	0.681	0.720	0.648	0.743	0.601	0.609	**0.661**
LOO_all	0.697	0.601	0.766	0.698	0.751	0.668	0.776	0.620	0.611	**0.688**

Table 4.9: Correlations between actual and estimated TVs for the EMA-IEEE speech inversion system

LA	**LP**	**JA**	**TTCL**	**TTCD**	**TBCL**	**TBCD**	**TRCL**	**TRCD**	**Average**
0.869	0.771	0.891	0.815	0.856	0.790	0.884	0.722	0.826	**0.825**

4.3.4 Evaluation across speaking rates

To study the effect of speaking rate on speech inversion performance, we evaluated the systems trained on the EMA-IEEE datasets across speaking rates. We performed cross speaking rate evaluations on the speaker dependent (SD) and speaker independent (LOO) speech inversion systems. For example, we evaluated the SD_N system for each speaker on the fast utterances from the corresponding speaker. Similarly, we evaluated the LOO_F system for each speaker with the normal utterances of the corresponding left out speaker. This way we performed cross speaking rate experiments with SD_N, SD_F, LOO_N, and LOO_F systems for all 8 speakers in the EMA-IEEE dataset. We computed the average correlation

between the actual and estimated TVs for each evaluation. The average correlations for the matched and mismatched speaking rate experiments are shown in table —. Note that the matched speaking rate numbers are the same average correlation numbers shown in Table 4.8.

Table 4.10: Correlations between actual and estimated TVs for matched and mismatched speaking rate evaluations for various speech inversion systems trained on the EMA-IEEE dataset. Numbers in brackets show the standard deviations of the correlations.

	Normal rate	**Fast rate**
SD_N	0.823 (0.037)	0.749 (0.035)
SD_F	0.756 (0.080)	0.757 (0.073)
LOO_N	0.716 (0.063)	0.621 (0.048)
LOO_F	0.672 (0.035)	0.645 (0.053)

The results shown in Table 4.10 show that the systems trained on normal rate speech perform similar to the fast rate trained systems on Fast rate speech. However, the performance of fast speech trained systems is much poor on normal rate speech compared to the normal trained systems. The reason for this disparity in performance across speaking rates is due to the higher variability in the acoustic and articulatory spaces in the fast rate speech data.

4.4 Variability due to accent and language

The goal of this study is to assess how appropriate normalization and deep and shallow neural network techniques may help in creating an adequate speaker-independent acoustic-to-articulatory speech inversion system. To reliably assess the performance of our system, we use articulatory data of more than 40 speakers collected in a research project investigating native and non-native pronunciation of English (Wieling et al., 2015). Specifically, we focus on two subsets of data collected in this project. The first subset consists of English and Dutch utterances from 21 L1 Dutch speakers (NL data), whereas the second subset consists of English utterances from 22 British English speakers (UK data). Both sets of data contain simultaneously recorded acoustic and electromagnetic articulograpy (EMA) data. Besides using the actual EMA sensor trajectories, we converted the sensor trajectories to Tract Variables (TVs) (Saltzman and Munhall, 1989) using geometric transformations (explained in Section 4.4.2).

We trained separate speech inversion systems on both the NL data as well as the UK data to estimate the EMA sensor positions as well as the TVs. In order to compute the accuracy of the speaker-independent speech inversion systems, we trained and tested them using leave-one-speaker-out cross validation. For the NL data, we trained separate speech inversion systems on exclusively Dutch utterances, English utterances, and both Dutch and English utterances. In the following, we compare the performance of these speaker-independent speech inversion systems across the two datasets.

4.4.1 Dataset description

4.4.1.1 EMA data

The data used in this study was collected to compare the pronunciation and articulation of English by Dutch speakers to the English pronunciation of native Southern Standard British English speakers (see also (Wieling et al., 2015)). The articulatory data was collected on site (in Groningen, the Netherlands for the Dutch speakers, and in London, UK for the native English speakers) using an NDI Wave 100 Hz 16-channel articulography device. For the articulatory data collection, three sensors were attached to the midline tongue: one at about half a cm. behind the tongue tip (TT), one about three cm. behind the TT sensor (TB), and the other midway between TT and TB (TM). We further attached three sensors to the lips and two to the teeth: one at the center of the upper lip (at the vermilion border; UL), one at the center of the lower lip (at the vermilion border; LL), and the third in the right corner of the lips (SL). The teeth sensors were attached to the lower incisor (LI) and to the upper incisor (UI). To correct for head movement, we attached four sensors to the head (left and right mastoid process and two at the front of the head), and we used a biteplate with three sensors to rotate all other sensors to a common coordinate system relative to the occlusal plane. The articulatory data was synchronized with the acoustic data, which was collected using a sampling rate of 22.05 kHz (using an Audio Technica AT875R microphone).

In London, we collected data for 22 speakers, whereas we collected data for 21 speakers in Groningen, the Netherlands. For the Dutch speakers, the experiment

consisted of two parts. In the first (native Dutch) phase of the experiment, we collected articulatory and acoustic data when the speakers pronounced one paragraph of text (the Dutch version of the North Wind and the Sun), which was followed by the collection of pronunciation data for about 125 words and non-words (in random order, all repeated twice). Each word was preceded and succeeded by a schwa to ensure a neutral articulatory context at the beginning and end of the word pronunciation. In the second (English) phase of the experiment, the participants first pronounced two paragraphs of text (i.e. the North Wind and the Sun, and a paragraph of text used in the Speech Accent Archive (Weinberger, 2010)), which was followed by about 175 English words and non-words (in random order, each repeated twice, and preceded and followed by the schwa). Finally, if there was still time left, participants were asked to pronounce sentences from the Mocha-TIMIT corpus (Wrench, 2000). For the native English speakers, there was no Dutch phase of the experiment, but the individual words were pronounced both without the schwa context and with the schwa context. In total, this resulted in about 185 minutes of speech for the 21 Dutch speakers (NL data) and 235 minutes of speech for the 22 native English speakers (UK data).

The raw EMA data was corrected for head movement and aligned to the occlusal plane. Missing sensor data (due to sensors which malfunctioned, or came off during the experiment) was estimated using the algorithm outlined in (Qin and Carreira-Perpiñán, 2010). In short, a probability density of the sensor positions was estimated, and the missing sensor coordinates were approximated using conditional distributions derived from the modeled density (Qin and Carreira-Perpiñán, 2010).

4.4.2 Conversion of EMA sensors to Tract Variables

The specific EMA data greatly depends on the anatomy of the speaker and the points where the sensors are placed. Vocal tract constriction variables, or tract variables (TVs), are measures of constriction position and location along the vocal tract. Instead of actual coordinates (x: anterior-posterior axis, z: inferior-superior axis) of the sensors, the TVs represent relative positions of the articulators. We converted the EMA sensor trajectories to ten TVs using geometric transformations as shown in Figure 4.6. The ten TVs were: Lip Aperture (LA), Lip Protrusion (LP), Lip Width (LW), Jaw Aperture (JA), Tongue Tip Constriction Location (TTCL), Tongue Tip Constriction Degree (TTCD), Tongue Middle Constriction Location (TMCL) and Tongue Middle Constriction Degree (TMCD), Tongue Root Constriction Location (TMCL) and Tongue Root Constriction Degree (TMCD). LA was defined as the Euclidean distance between the UL and the LL sensors. LP was defined as the displacement along the x-axis of the LL sensor from its median position. Lip Width (LW) was defined as the Euclidean distance between the SL sensor and the centroid of the UL and LL sensors. JA was defined as the Euclidean distance between the UL sensor and the LI sensor. Two TVs were computed for each tongue sensor - constriction degree and location. Constriction degree for a tongue sensor was defined as the minimum distance between the sensor and the (automatically determined, data-driven) palate trace. This way the TTCD, TMCD, and TBCD TVs were computed from the TT, TM and TB sensor positions and the palate trace. The constriction location for a tongue sensor was defined

as the displacement of the sensor along the x-direction from its median position. Thus, TTCL, TMCL, and TBCL were computed from the TT, TM, and TB sensor positions.

4.4.3 Results

4.4.3.1 Leave one speaker out tests

Given the large number of speakers in the UK and NL data, we used leave-one-speaker-out cross-validation (LOCV) to evaluate the speaker-independent speech inversion performance within each dataset. These experiments were performed for both subsets of data separately. The NL data, which consisted of both English and Dutch utterances, was divided into three sets: Dutch utterances (NL_dutch), English utterances (NL_english), and all utterances (NL_all). The UK data only consisted of English utterances (UK_english). The LOCV tests were performed for each of these four sets. Table 4.11 provides an overview of these systems and the corresponding subsets of data. For the UK dataset, 18 speakers were randomly selected for neural network training, 3 speakers were randomly selected for the validation step (to determine the stopping criterion for the neural network training), and finally the system was tested on the single remaining speaker (i.e. in the LOCV approach, each speaker was included in the test set exactly once). For the Dutch data, a similar approach was used.

The neural networks for the UK_english system had three hidden layers with 300 nodes in each layer. Due to the limited amount of Dutch utterances available

in the NL data (see Table 4.11), the NL_dutch systems were trained with a single hidden layer (with 300 nodes). Similarly, we restricted the number of hidden layers to two (with 300 nodes in each layer) for the NL_english systems. The NL_all systems, which were trained with both the English and Dutch utterances, were given three hidden layers with 300 nodes in each layer. The LOCV experiments were performed separately for estimating EMA sensor positions as well as TVs. For the EMA sensor positions, we estimated the x and z coordinates for all the sensors except for the SL sensor, for which we estimated the x and y (i.e. left-right) positions. The average correlations (on the basis of the LOCV test set results) for the EMA sensor positions are shown in Figure 4.7, whereas Figure 4.8 shows the same for the TVs.

Table 4.11: Speech inversion systems and their training data

System name	**Data**	**Amount of data**
UK_english	English utterances from 22 UK English speakers	235 min.
NL_dutch	Dutch utterances from 21 L1 Dutch subjects	60 min.
NL_english	English utterances from 21 L1 Dutch subjects	126 min.
NL_all	English and Dutch utterances from 21 L1 Dutch subjects	186 min.

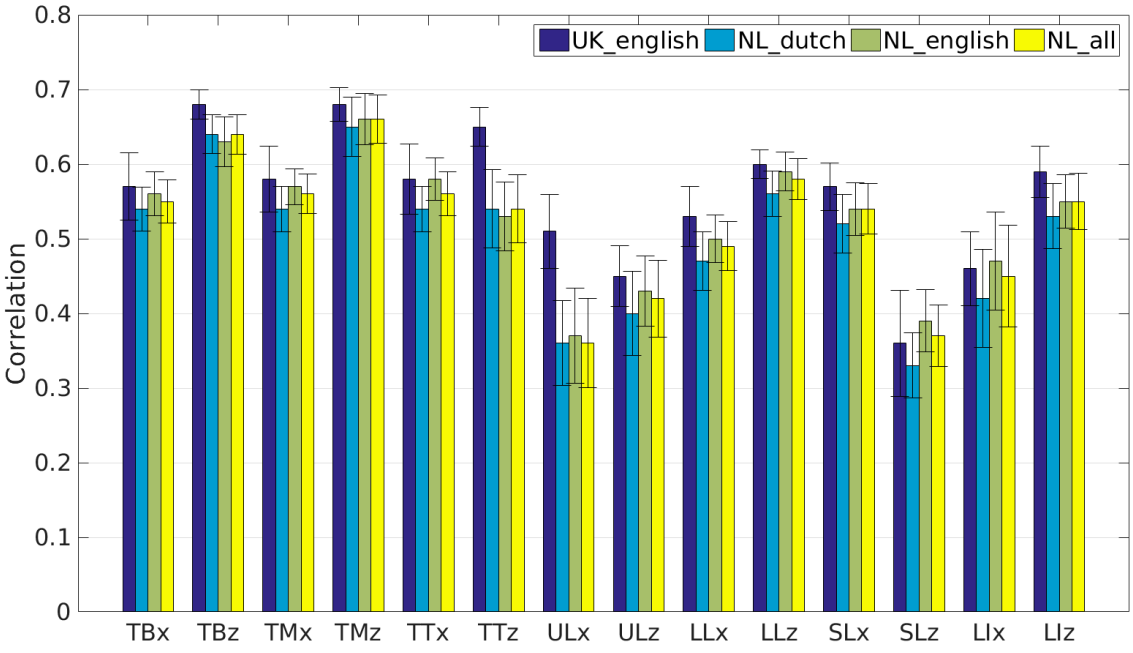


Figure 4.7: Average (across all speakers) correlations between actual and estimated EMA sensor positions. Error bars denote two standard errors.

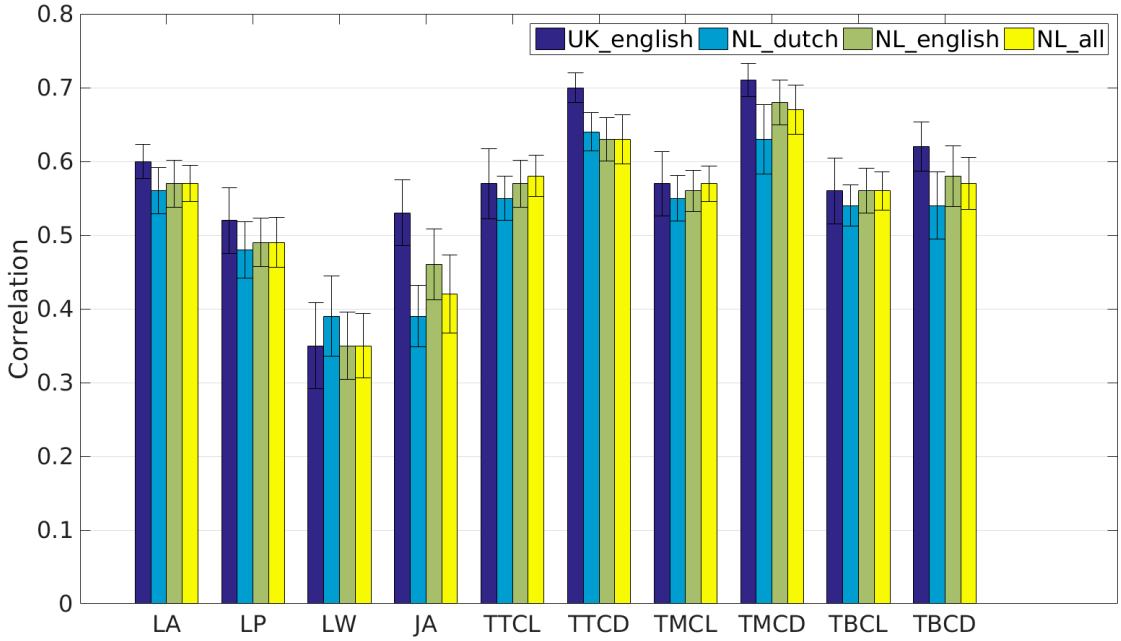


Figure 4.8: Average (across all speakers) correlations between actual and estimated tract variables. Error bars denote two standard errors.

4.4.3.2 Cross-domain experiments

The performance of the speech inversion systems illustrated above shows how well the system has learned to estimate the articulatory patterns of that language. In this section, we report how well our speech inversion system is able to perform in a cross-language setting. For this purpose, we evaluated how well a system trained on the data from the UK English speakers was able to predict the articulatory trajectories of the Dutch speakers (in accented English, Dutch, or both) and vice versa. Instead of training on all data to obtain a new model, we used the best-performing model from the LOCV approach.[1] We evaluated the performance of each of the four systems on all speakers on the basis of the other three subsets of data. For example, the UK_english system was tested on every speaker from the NL dataset (separately for the three subsets of data: Dutch only, non-native English only, or both). The results of these experiments are presented in Tables 4.12 and 4.13 (for the EMA sensor positions: correlations and RMSE) and in Table 4.14 (for the TVs). Note that the values on the diagonals reflect the average LOCV performance for each subset of the data shown in Figures 4.7 and 4.8.

4.4.4 Discussion

In this study we have shown that our system is able to model speaker-independent articulatory positions, with a correlation of about $r = 0.53$.

This is substantially lower than the correlation of about $r = 0.62$ reported in

[1] While it is likely that a model on the basis of all data would have been slightly better performing, it is unlikely that this would have impacted the results substantially.

Table 4.12: Average correlations (including standard error) for actual and estimated EMA sensor positions based on different datasets. The left-most column indicates the speech inversion system. The top row indicates the test set. The numbers in the brackets indicate standard errors

Train \ Test	UK english	NL dutch	NL english	NL all
UK_english	0.56 (0.012)	0.42 (0.015)	0.48 (0.014)	0.45 (0.015)
NL_dutch	0.42 (0.010)	0.51 (0.012)	0.46 (0.012)	0.47 (0.011)
NL_english	0.48 (0.011)	0.48 (0.013)	0.53 (0.011)	0.51 (0.011)
NL_all	0.49 (0.011)	0.51 (0.013)	0.53 (0.011)	0.52 (0.012)

Table 4.13: Average RMSE between actual and estimated EMA sensor positions based on different datasets. The left-most column indicates the speech inversion system. The top row indicates the test set. The RMSE values are in mean-variance normalized sensor coordinates

Train \ Test	UK english	NL dutch	NL english	NL all
UK_english	0.83	0.86	0.91	0.90
NL_dutch	0.91	0.81	0.91	0.88
NL_english	0.88	0.83	0.87	0.86
NL_all	0.87	0.81	0.87	0.85

(Ji et al., 2016), but our result does not depend on a specific reference group of speakers. Furthermore, if we exclude the performance with respect to UL and SL

Table 4.14: Average correlations (including standard error) for actual and estimated tract variables based on different datasets. The left-most column indicates the speech inversion system. The top row indicates the test set. The numbers in the brackets indicate standard errors

Train \ **Test**	**UK english**	**NL dutch**	**NL english**	**NL all**
UK_english	0.57 (0.012)	0.44 (0.013)	0.51 (0.012)	0.48 (0.014)
NL_dutch	0.43 (0.010)	0.52 (0.012)	0.48 (0.011)	0.49 (0.011)
NL_english	0.50 (0.011)	0.49 (0.013)	0.54 (0.010)	0.52 (0.011)
NL_all	0.51 (0.011)	0.54 (0.011)	0.56 (0.010)	0.54 (0.010)

for the EMA sensor positions, and the LW tract variable (not included by (Ji et al., 2016)), these correlations increase to $r = 0.58$. The UL and SL sensors (and tract variables) are difficult to predict as their influence on the speech signal is relatively limited.

The objective of speaker-independent speech inversion is to accurately capture the trend of the articulatory movements, even though there might be offsets in actual sensor positions due to the anatomical mismatch between training speakers and the speaker used to evaluate the model performance. The performance on the basis of tract variables was only marginally better than the performance based on the EMA sensor positions. As the EMA sensor positions were normalized with respect to their mean and variance, this also (just as tract variables) abstracts away from most anatomical variation.

While cross-language modeling of the trajectories resulted in a lower correlation than the within-language results, the drop in performance was only limited, especially when more data was available (i.e. comparing NL_all to UK_english). Tables 4.12 and 4.14 show several evaluations of the speaker-independent speech inversion systems across different test sets. Table 4.13 shows the root mean squared error between actual and estimated (normalized) EMA sensor coordinates across the different test sets. The results in the table highlight the performance of the systems in different mismatch conditions. The native language mismatch condition is highlighted comparing UK_english to NL_all. The NL_all system performs better on the UK english set than vice versa. This might be due to the fact that the UK data is cleaner (due to being recorded in a soundproof booth) than the NL data. Consequently, the system trained on the clean UK data performs poorly on the NL data. The accent mismatch is highlighted by comparing UK_english to NL_english. We observe that the performance of the UK_english system on the NL english set is close to the within dataset (NL_english-NL english) performance. By contrast, the NL_english system performs much worse than the UK_english system on the UK english dataset. On the one hand, this can be attributed to the higher amount of training data in the UK dataset. On the other hand, the amount of variability in the acoustics and articulatory movements is likely higher for the L2 English speakers (leading to poorer NL_english speech inversion models). Finally, the performance when the language is completely mismatched is shown by the UK_english vs. NL_dutch comparison. Unsurprisingly, we see lower correlations in these comparisons, which can be attributed to both language

mismatch as well as a data mismatch. By contrast, the NL_english vs. NL_dutch comparison avoids the problem of mismatched data (i.e. collected at different sites), and their comparison highlights the effect of language mismatch in speech inversion performance (i.e. about 0.05 reduction in the correlation coefficient).

4.4.5 Summary

The experiments performed in this study shed light on the effects of the amount of training data, the different types of data (i.e. collected in different environments), and different accents and languages on the performance of speech inversion systems. Our results highlight that with appropriate normalizations of the acoustic features and articulatory trajectories, speaker independent systems can estimate the sensor positions and TVs reasonably well with a correlation of about 0.53 with matched training and testing conditions. For mismatched data, the performance drops to about 0.43. Speaker normalization techniques (Sivaraman et al., 2016; Girin et al., 2017) may further improve the performance of these systems. This section also highlights that data collected using the same protocol may be combined in order to generate improved speech inversion systems, even if the languages are different. In future work, we plan to develop methods for combining data collected with different protocols and potentially even different modalities for the creation of speech inversion systems.

Chapter 5

Uncovering acoustically weak articulatory maneuvers

Speech patterns vary significantly due to speaking rate, speaking style, context and emotions. Among all the variability observed in speech, in this chapter we are particularly interested in the changes occurring in speech acoustics due to speaking rate and speaking style. We split this chapter into two major sections - 5.1 Coarticulation and lenition, and 5.2 Distinguishing acoustically similar articulatory maneuvers: The case of American English /r/

Coarticulation and lenition are common phenomena that occur in fast rate speech, especially affecting the acoustic properties of the speech signal that relate to manner and place of articulation. The resulting acoustic variability continues to offer serious challenges for the development of automatic speech recognition (ASR) systems that can perform well with minimal constraints. The phenomena of coarticulation and lenition often manifest as deletion or substitution of phone units when looked at from the acoustic perspective. However, articulatory phonology explains coarticulation through spatio-temporal changes in the patterns of underlying gestures. This chapter studies the coarticulation occurring in certain fast spoken utterances using articulatory constriction tract-variables (TVs) estimated from acoustic features. The objectives of this study are to study the effects of coarticulation in fast rate speech from the articulatory perspective and

to test whether an acoustic-to-articulatory speech inversion system can predict the hidden articulatory gestures underlying the coarticulated syllables. The implications of this study are significant since, if our speech inversion system is able to "uncover" seemingly hidden gestures, then the robustness and accuracy of ASR systems will be vastly improved. Furthermore, such results will also provide the means for improving a variety of speech applications and leading, for example, to the strengthening of speech pronunciation tools in the classroom and clinic, and to the development of more natural sounding synthetic speech that will better reflect idiosyncratic individual differences between speakers.

It has been shown using Magnetic Resonance Imaging (MRI) techniques that the American English /r/ sound has a range of possible tongue configurations that range from a retroflex tongue shape on one end to a bunched tongue shape on the other. (Zhou et al., 2008). It is also known that the acoustic distinction between the bunched and retroflex tongue shapes differ in the frequency spacing between the formants F4 and F5 which are known to be very weak compared to F1, F2, and F3. In this chapter, we try to answer the question whether the acoustically weak distinction between the bunched and the retroflex configurations can be uncovered using the articulatory trajectories estimated by a speech inversion system. We use a speaker independent speech inversion system trained on the XRMB dataset (Chapter 3) to estimate the TVs for two example utterances containing bunched and retroflexed /r/s. We show that the speech inversion system is able to distinguish between the two forms of /r/ even though the speakers from the test utterances were never part of the training set of the speech inversion system.

5.1 Coarticulation and lenition

Coarticulation is the overlap of articulatory movements of adjacent sounds in speech leading to what many refer to as substitution or deletion of sounds. A phoneme is influenced by, and becomes more like, a preceding or following phoneme based on the context. For example, in the fast production of the phrase "perfect memory", the /t/ sound appears to be deleted due to the lip closure gesture for /m/ overlapping with the tongue tip constriction for /t/. Lenition occurs when a gesture for a particular phoneme does not reach its intended target leading to a weakening of the sound. It commonly occurs in consonants where a stop consonant becomes a fricative.

Articulatory Phonology (AP) provides a unified framework for understanding how spatiotemporal changes in the pattern of underlying speech gestures can lead to corresponding changes in the extent of intergestural temporal overlap and in the degree of gestural spatial reduction; in turn, these changes in overlap and reduction create acoustic consequences that are typically reported as assimilations, insertions, deletions and substitutions.

In this chapter, we focus on the capabilities of the speech inversion system for modeling changes in temporal overlap and spatial magnitude of gestures. Specifically, we address the following questions: (1) With proper contextualization, can our speech inversion system uncover gestures 'hidden' acoustically by increases in overlap (coarticulation) and/or decreases in magnitude (lenition)?; (2) Is undershoot of articulatory targets accurately reflected in the output of the speech

inversion system?; (3) Will the use of naturally-spoken data (e.g., concurrently recorded speech acoustics and kinematics) for training the speech inversion system result in AP gestural trajectories that accurately reflect articulatory movements during and between gestures?; and finally (4) What is the best methodology for training the speech inversion system with naturally-spoken speech and articulatory data? The implications of successfully answering these questions are significant since, if our speech inversion system is able to "uncover" seemingly hidden gestures, then the robustness and accuracy of ASR systems will be vastly improved. Furthermore, such results will also provide the means for improving a variety of speech applications and leading, for example, to the strengthening of speech pronunciation tools in the classroom and clinic, and to the development of more natural sounding synthetic speech that will better reflect idiosyncratic individual differences between speakers.

5.1.1 Articulatory datasets and speech inversion systems

Fast rate speech leads to significant coarticulation and reduction phenomena. For example, in "perfect memory", the '/t/' often appears to be deleted acoustically due to the overlap of the lip closure for '/m/' with the tongue tip constriction for '/t/'; examination of the TV trajectories, however, shows that the underlying gestures persist. To obtain data to investigate such contexts we recorded speech at normal and fast rates concurrently with Electromagnetic Articulograph (EMA) data, using the IEEE sentences (Rothauser et al., 1969) as the corpus for this task.

We will refer to this dataset as the EMA-IEEE database. A complete description of the recordings of the EMA-IEEE sentences is given in Section 4.3.1. We trained artificial neural networks (ANNs) for acoustic-to-articulatory speech inversion using speech and articulatory data obtained from the U.W. X-ray Microbeam (XRMB) database (Westbury, 1994). A description of the speech inversion system is given in chapter 3 . The trained speech inversion systems were used to estimate TVs for specific fast and normal rate utterances from the EMA-IEEE database. TVs were estimated from the sensor positions recorded using EMA (using the method outlined in Section 4.3.2), and were compared to the actual TVs obtained from the Electromagnetic Articulograph (EMA) recordings of the same IEEE sentences. A speech inversion system was also trained on the articulatory (EMA) data recorded for this experiment. This section compares the ability of various speech inversion systems to detect an utterance's underlying gestures given the significant coarticulation effects of fast speech.

Four different TV estimators were trained using the two datasets - EMA-IEEE (Section 4.3.1) and XRMB (Section 3.4). A TV estimator was trained on the complete XRMB database. This estimator is referred to as X_NORM. To normalize gender specific acoustic variations, the XRMB database was divided into male and female speaker utterances and a TV estimator was trained on each of these subsets. The systems trained on these gender specific subsets are referred to as XF_NORM and XM_NORM. Another TV estimator was trained using the EMA-IEEE dataset. This system estimates only 3 TVs (LA, TBCD, and TTCD) as the other TVs were not computed from EMA trajectories. We refer to this estimator as E_IEEE. Table

5.1 summarizes these TV estimators. The trained TV estimators were tested on 10% of their respective datasets held out for testing where the sentences were chosen randomly. The performance of the TV estimator was measured by the Pearson Product Moment Correlation (PPMC) between the estimated and ground-truth TVs using the test set. The results for the different TV estimators are given in Table 5.2.

Table 5.1: Description of different speech inversion systems trained

TV estimator name	**Training dataset**
X_NORM	XRMB utterances converted to TVs
XF_NORM	Female speakers' utterances from XRMB database converted to TVs
XM_NORM	Male speakers' utterances from XRMB database converted to TVs
E_IEEE	Single female speaker EMA data converted to TVs

Table 5.2: Correlation results of trained TV estimators on their respective test data sets. (NA: TVs were not estimated)

TV estimator name	**LA**	**TBCD**	**TTCD**	**LP**	**TBCL**	**TTCL**
X_NORM	0.66	0.59	0.76	0.56	0.78	0.65
XF_NORM	0.72	0.66	0.79	0.62	0.82	0.66
XM_NORM	0.68	0.64	0.78	0.57	0.83	0.72
E_IEEE	0.64	0.80	0.72	NA	NA	NA

5.1.2 Analysis of specific examples of coarticulation

Effects of coarticulation and reduction can be expressed in many forms in fast rate speech including deletion, assimilation and substitution. In this paper, we selected two utterances from the EMA-IEEE dataset and one utterance from an earlier study (Tiede et al., 2001) illustrating coarticulation effects. Both fast rate and normal rate utterances of these selected sentences were analyzed. Articulatory data was converted to TV representation using the same method described in Section 4.3.2. The following are the three sentences chosen for analysis.

1. The empty **flask stood** on the tin tray.

2. The beam dropped down on the **workman's head**.

3. She had a **perfect memory** for details. (from (Tiede et al., 2001))

The words in bold contain the clusters of interest. None of the above utterances were included in any of the TV estimators trained. Each of these utterances was analyzed using the TV estimators described in section 5.1.1. We analyzed only the LA, TBCD, and TTCD TVs.

The average correlations of the estimated TVs with actual TVs for the three selected utterances are shown in Table

From Table 5.3, we can see that the E_IEEE system has the highest correlations for sentences 1 and 2 since those utterances were produced by the same speaker (note that these utterances were not included in the training of this system). Hence, we plotted the estimated TVs from E_IEEE system for analysis of sentences 1 and 2.

Table 5.3: Correlations (PPMC) of estimated TVs from different TV estimators for the selected sentences (n = normal rate, f = fast rate)

TV estimator name	**flask stood**		**workman's head**		**perfect memory**	
	n	f	n	f	n	f
X_NORM	0.56	0.59	0.61	0.75	0.40	0.51
XF_NORM	0.56	0.59	0.55	0.72	0.28	0.55
XM_NORM	0.56	0.59	0.59	0.63	0.44	0.58
E_IEEE	0.86	0.82	0.75	0.79	0.18	0.44

5.1.2.1 Analysis of "flask stood"

Figure 5.1 shows spectrograms and the TVs for the normal-rate and fast-rate productions of sentence 1. In the case of the normal-rate production, the consonant cluster /sk/ at the end of "flask" and the /st/ at the beginning of "stood" are clearly seen in the acoustics and both the actual and estimated TVs show constrictions in the right regions.

However, in the fast-rate production of this utterance, the acoustics suggest that the /k/ in "flask" was not produced. Instead, it appears as if the /s/ in "flask" and the /s/ in stood are combined (the duration of this /s/ is about 30ms longer than the ones in the normal-rate production) and this /s/ is then followed by a the /t/ in "stood". This appears to be a case where the fast-rate production resulted in no gesture being made for the /k/. Although there is lowering of the TBCD gesture during the /t/, this lowering appears to be due to the /t/ closure and can be seen in situations where a /s/ or /t/ is produced without an adjacent velar consonant.

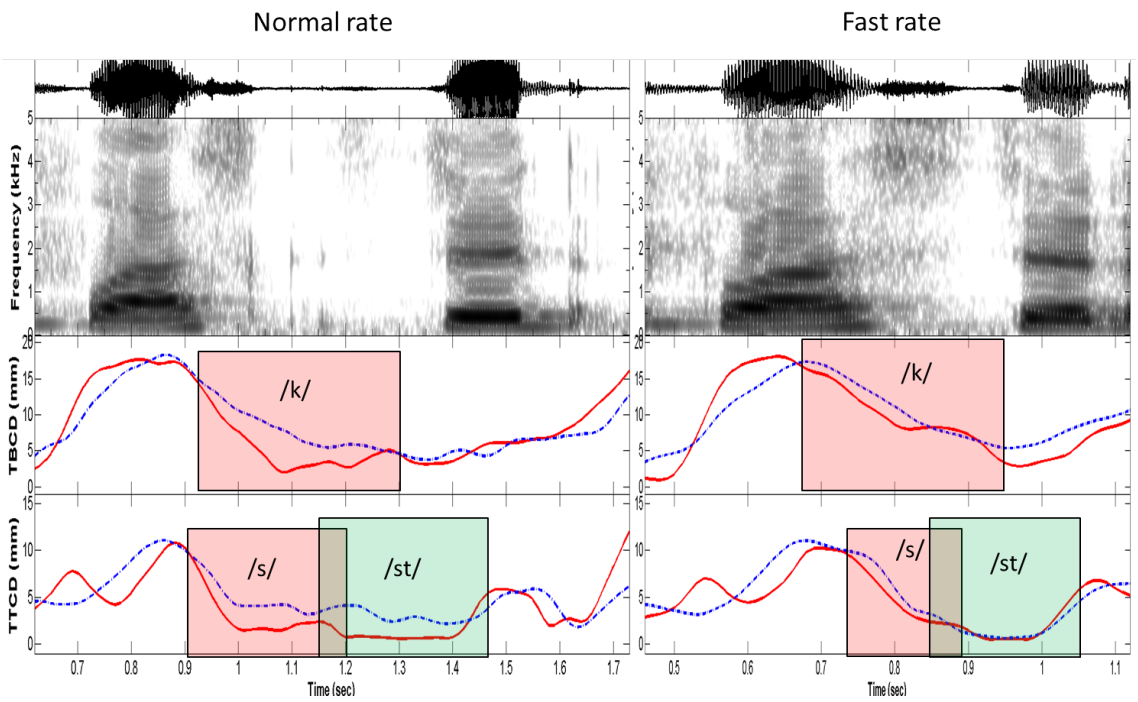


Figure 5.1: Actual (red) and estimated (blue) TVs for “flask stood”

It is possible that this apparent deletion of the /k/ gesture is due to the complexity of these cluster sequences, which include four consecutive consonants.

5.1.2.2 Analysis of ”workman”

Figure 5.2 shows spectrograms and the TVs for the normal-rate and fast-rate productions of sentence 2. The actual and estimated TVs are strongly correlated across the utterance. In particular, both show the /k/ constriction when it is produced as a stop in the normal-rate production and as a fricative in the fast-rate production. Note that the /k/ gesture in the fast-rate production of the utterance is weaker than it is in the normal-rate production of the same. This not surprising given the estimated TVs are derived from the acoustics. Finally, note that both sets of TVs show the closure of the lips for the /m/.

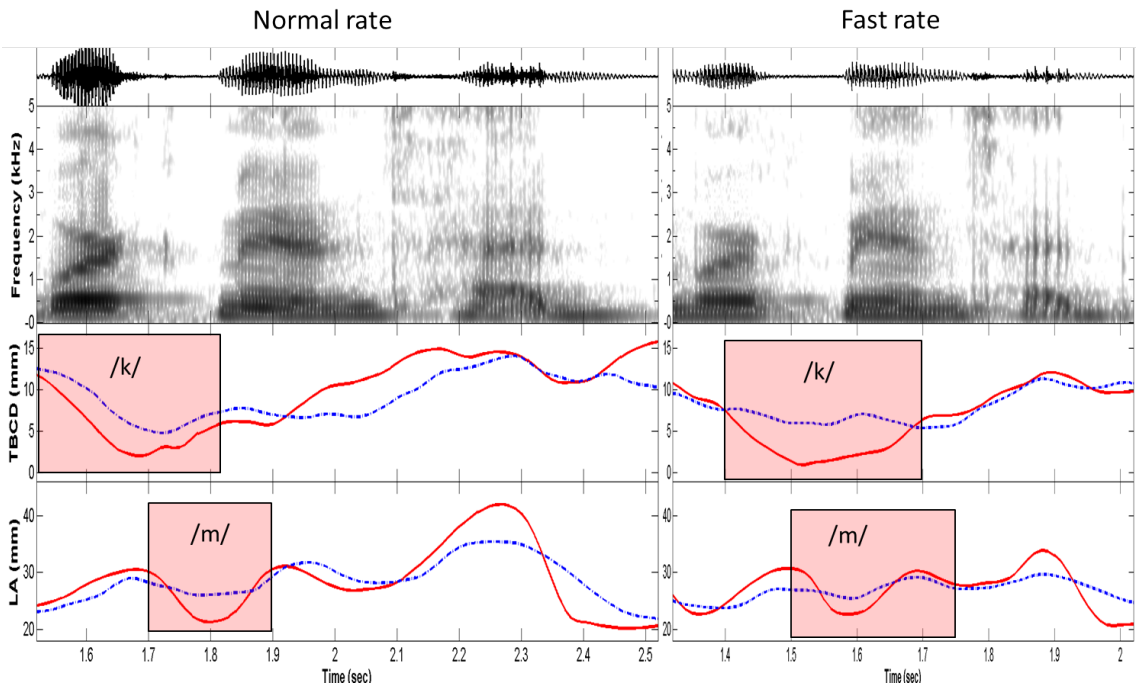


Figure 5.2: Actual (red) and estimated (blue) TVs for “workman’s head”

5.1.2.3 Analysis of ”perfect memory”

From Table 5.3, it can be seen that none of the TV estimators provide a reliable estimate of the 3 TVs that we are interested in analyzing. As a result, we trained speaker dependent TV estimators on all 46 speakers of the XRMB database. We then estimated TVs using each speaker’s TV estimator and then selected the system that best correlated with the actual TVs. We observed that the TV estimator trained on speakers JW29 provided best correlations for normal rate and that trained on JW28 provided best correlations for fast rate. We used the estimated TVs from these models to analyze the “perfect memory” utterance.

Figure 5.3 shows spectrograms and the actual and estimated TVs for sentence 3. As can be seen in the normal-rate production, the acoustics show a release burst for /k/ but not /t/, followed by a period of silence and then the /m/ murmur at the

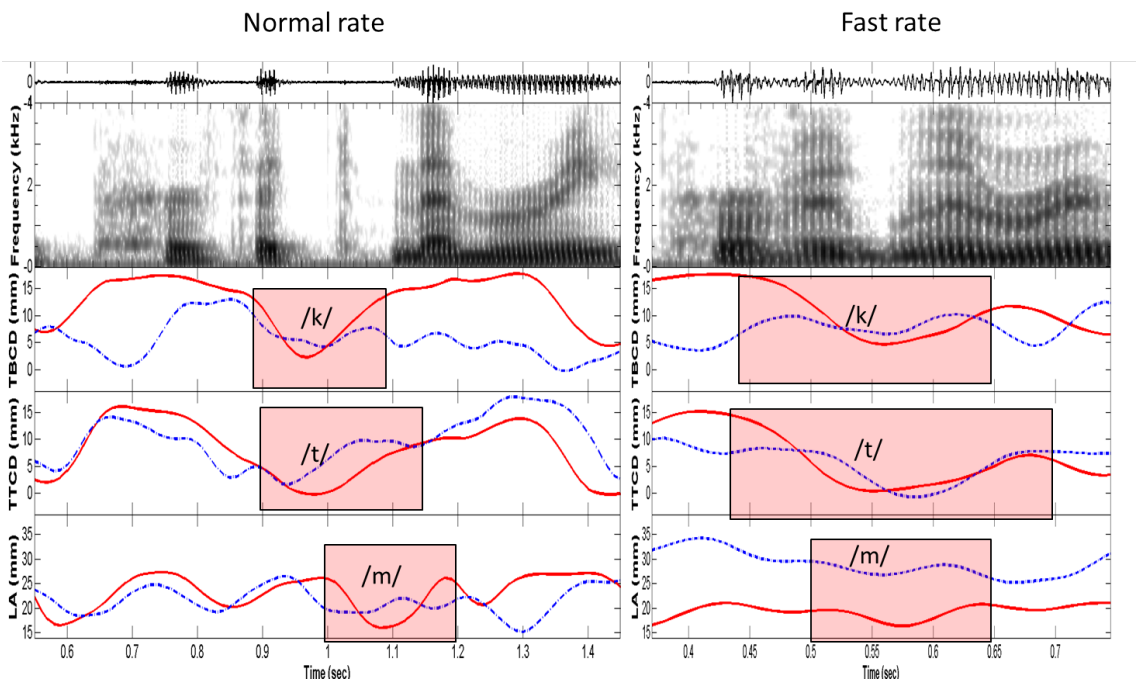


Figure 5.3: Actual (red) and estimated (blue) TVs for "perfect memory"

beginning of "memory". Both sets of TVs show a tongue-body gesture for the /k/ that overlaps with the tongue-tip gesture for the /t/ and the lip gesture for the /m/. In contrast, there is no silence between the last vowel in "perfect" and the first vowel in "memory". Instead, this region appears acoustically as one sonorant consonant, i.e., the /m/. However, the articulatory data tell a different story. As in the normal rate speech, we see gestures for the /k/ and /t/, but with considerably more overlap between the gestures. In particular, the /m/ gesture is fully overlapped with that of the other consonants. Thus, this fast-rate production of "perfect memory" contains what we refer to as "hidden gestures" for the /k/ and the /t/. Note that both of these gestures are apparent in the estimated TVs, although the closure for the lip gesture is weaker than the actual gesture.

5.1.3 Summary

The results show that the speech inversion systems perform reasonably well on unseen data containing challenging coarticulatory phenomena. Working with naturally-spoken data can result in speech inversion systems that produce TVs that closely match TVs computed directly from articulatory data. However, the variability in the training data needs to be properly normalized or restricted. Thus, a future goal of this work is to develop methodologies for coping with variability and choosing which of several different speech inversion systems will work for any given speaker, especially if that speaker's data has not been used as part of the training data.

5.2 Distinguishing acoustically similar articulatory maneuvers: The case of American English /r/

It is well known that different tongue configurations are used by different speakers to produce the rhotic /r/ sound in American English (Delattre and Freeman, 1968), (Hagiwara, 1995), (Espy-Wilson and Boyce, 1999), (Espy-Wilson et al., 2000), (Tiede et al., 2004). Among the myriad tongue shapes for /r/ production, the two particular shapes that exhibit the greatest amount of contrast are the "bunched" /r/ (produced with a lowered tongue tip and a raised tongue dorsum) and the "retroflex" /r/ (produced with a raised tongue tip and a lowered tongue dorsum). Figure 5.4 shows the MRI image of the two tongue configurations of two different subjects producing their natural sustained /r/ sound (as in "pour").

In spite of the two tongue configurations for the /r/ production being significantly different, they do not show clear acoustic (Delattre and Freeman, 1968) (Westbury et al., 1998) or perceptual differences (Twist et al., 2007).

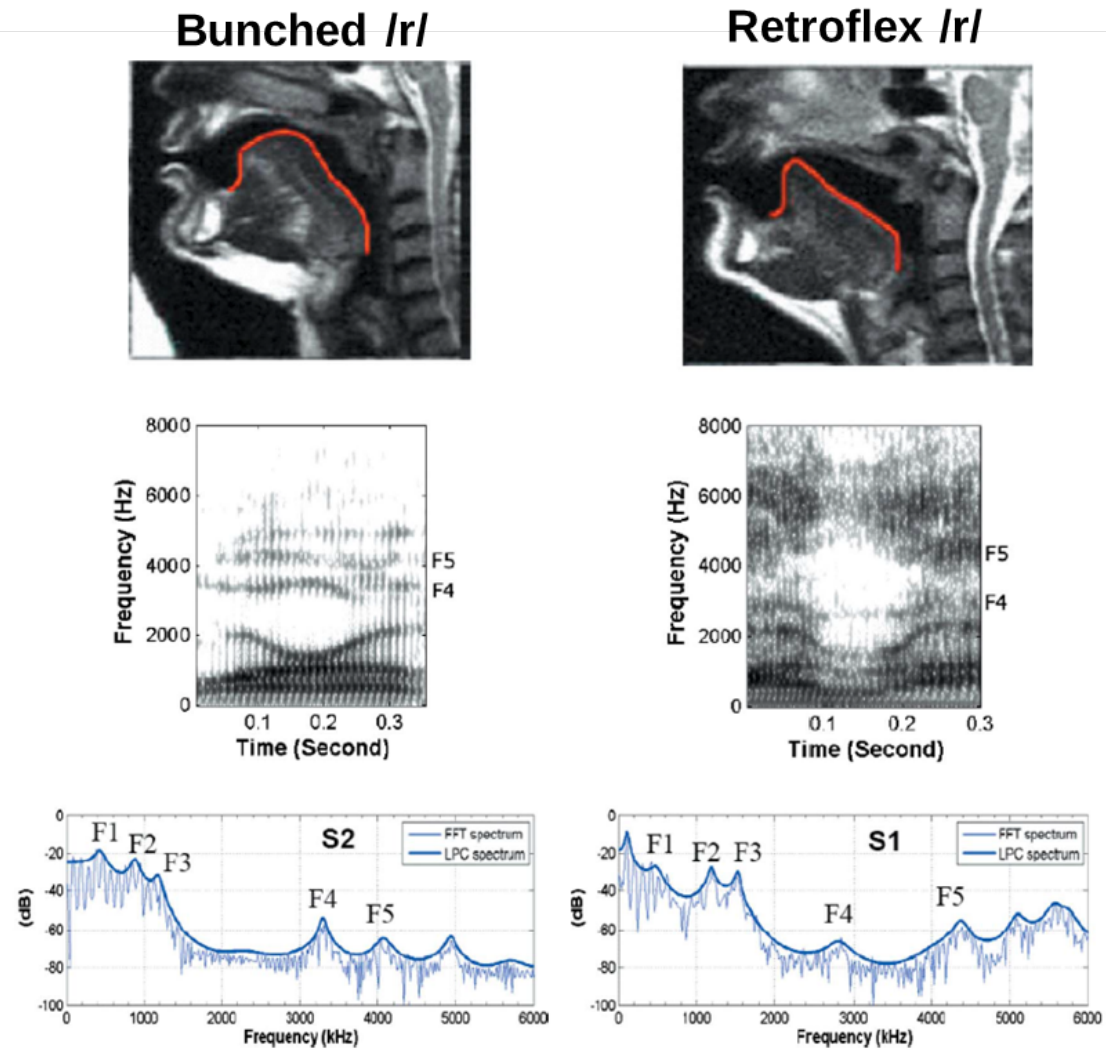


Figure 5.4: Bunched vs Retroflex /r/ production - Top panel: Midsagittal MR images of two tongue configurations for American English /r/. Middle panel: Spectrograms for nonsense word "warav." Lower panel: Spectra of sustained /r/ utterance. The left side is for S1 and the right side is for S2. Figure taken from (Zhou et al., 2008)

The American English /r/ is characterized by the lowered third formant frequency (F3) and often approaching F2 (Hagiwara, 1995), (Espy-Wilson, 1987).

This characteristic formant trajectory can be seen in the spectrogram shown Figure 5.4. Studies focused on F1-F3 formants failed to find a relationship between the tongue shape and the formant frequencies. It has been shown that higher formants specifically F4 and F5 may contain cues to differentiating the tongue configuration used to produce the /r/ sound (Espy-Wilson and Boyce, 1999), (Espy-Wilson et al., 2000). These higher formants are typically lower in amplitude in the spectrum making them difficult to detect. In addition, human perception appears to rely largely on the first three formants.

The /r/ sound is a difficult to produce for many children and some non-native English speaking adults. Studies have estimated that about 2-3% of college age people have trouble producing the /r/ sound (Reddy, 2014). Speech therapy is often sought to help children correctly produce the American English /r/. New methods of treatment providing visual articulatory feedback using ultrasound have been found to improve the pronunciation of /r/ sound (Cavin, 2015)(Preston et al., 2014). Ultrasound measurement techniques need expensive equipment in a speech therapist clinic. If a visual articulatory feedback can be provided based on the estimation of articulatory movements from speech acoustics, such visual feedback based speech therapy will become cheaper and easily accessible to people from all economic backgrounds. If an acoustic to articulatory speech inversion system can accurately estimate the tongue and lip movements from speech, it can potentially be used in speech therapy.

In this chapter we evaluate whether a speaker independent acoustic-to-articulatory speech inversion system can estimate the tongue

configurations for the two productions of /r/. We used the speaker independent speech inversion system trained on the XRMB database developed in Chapter 3 to estimate the TVs for the utterances by subject S1 and S2 producing the bunched and the retroflex /r/ respectively. Both the subjects were asked to produce a sentence containing the word "warav" in which the /r/ sound in "warav" is of interest to us. The subjects S1 and S2 produce the /r/ naturally and are not consciously biased towards either of the tongue configurations. Also note that the two subjects S1 and S2 are not part of the XRMB database which is the training set for our speech inversion system.

Figure 5.5 shows the estimated TVs for the bunched and the retroflex /r/ productions. These utterances are the same ones shown in Figure 5.4. The panels below the spectrograms show the Tongue Body Constriction Degree (TBCD), Tongue Tip Constriction Degree (TTCD), and the Lip Aperture (LA) estimated using the XRMB speech inversion system. We have not shown the plots of TBCL, TTCL and LP because of they are not critical in distinguishing the two configurations of /r/. For TBCD and TTCD, a peak in the plot indicates increased distance between the palate and the tongue (tongue body for TBCD and tongue tip for TTCD). A dip in the plot indicates constriction made by the tongue (tongue body for TBCD and tongue tip for TTCD) with the palate. Similarly, for the LA plot, a dip (low value) indicates constriction at the lips and a peak (high value) indicates wodening of the lips. For the bunched /r/ on the left panel, we see raised tongue dorsum as a dip in TBCD and a lowered tongue tip as a high value for TTCD. We see the constriction of the lip as a dip in the LA plot. For the retroflex /r/ on the

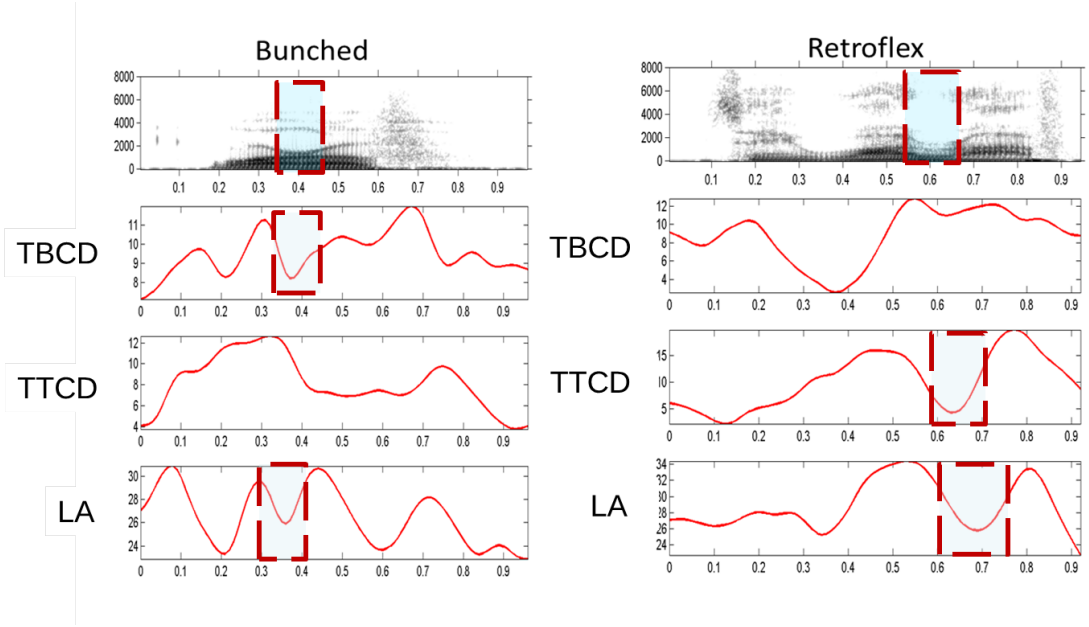


Figure 5.5: Estimated TVs for bunched and retroflex /r/ from the subjects S1 and S2. The red boxes on the spectrograms indicate the position of the /r/ sound in the utterance. The panels below the spectrograms show the Tongue Body Constriction Degree (TBCD), Tongue Tip Constriction Degree (TTCD), and the Lip Aperture (LA) estimated using the XRMB speech inversion system

right panel, we see the raised tongue tip as a dip in the TTCD plot and a lowered tongue dorsum as a high value for TBCD. We again see the constriction of the lips as a dip in LA. Thus, we see that the speech inversion system successfully estimates the tongue configurations for the bunched and the retroflex /r/ productions.

In order to test the hypothesis that higher formants may contain cues to differentiating between the two tongue configurations, we lowpass filtered the speech waveforms at 3000Hz to get rid of all formants higher than F3 that can possibly distingush bunched vs retroflex /r/. We then passed the filtered speech waveforms through the XRMB speech inversion system. Figure 5.6 shows the estimated TVs for the filtered speech waveforms from the same bunched and the retroflex examples. In this case we can clearly see that the speech inversion system fails to estimate the

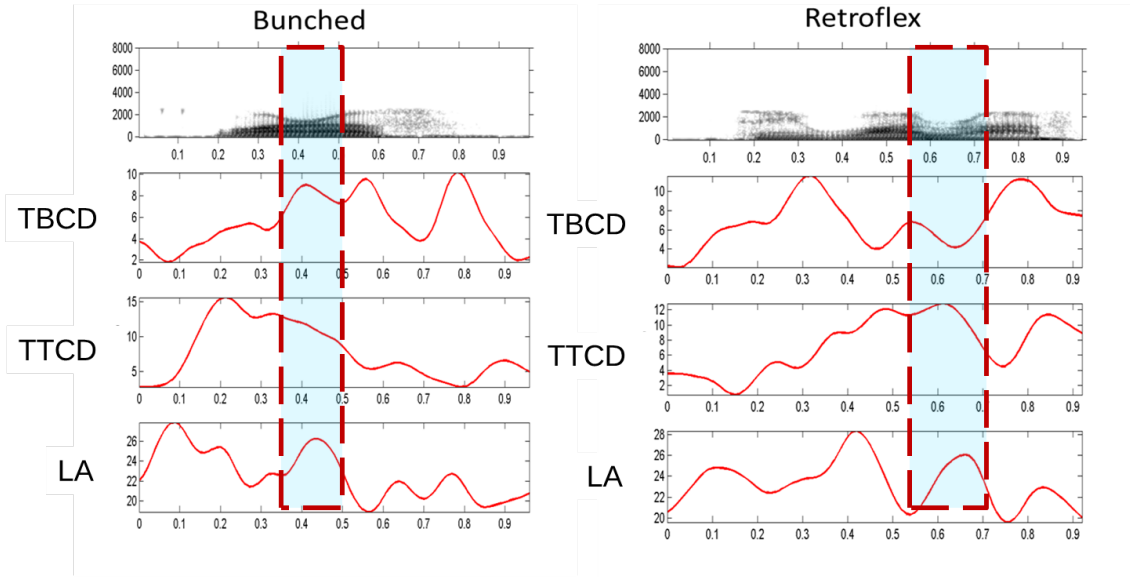


Figure 5.6: Estimated TVs for filtered utterances of bunched and retroflex /r/ from the subjects S1 and S2

correct tongue configurations for the bunched and the retroflex /r/, thus reinforcing the hypothesis that the higher formants contain cues to distinguish the bunched and the retroflex /r/.

5.2.1 Summary

The results of this experiment show that the speaker independent speech inversion system can successfully estimate the articulatory configurations for the bunched and retroflex productions of /r/. Thus the speech inversion system correctly distinguishes the two productions of the /r/ sound which are acoustically and perceptually similar. This work was presented at the Fall 2014 meeting of the Acoustical Society of America (Sivaraman et al., 2014)

Chapter 6

Phone place of articulation classification using articulatory features

In articulatory phonetics, a phoneme's identity is specified by its articulator-free (manner) and articulator-bound (place) features. Previous studies have shown that acoustic-phonetic features (APs) can be used to segment speech into broad classes determined by the manner of articulation of speech sounds. The effort in this chapter is to extend previous efforts (Juneja and Espy-Wilson, 2008) to develop a landmark system by adding in components to recognize place of articulation. Previous studies (Stevens and Blumstein, 1981) (Juneja and Espy-Wilson, 2008) (Chen and Alwan, 2000) have shown that finding the acoustic correlates of place of articulation is a very challenging task compared to manner of articulation. The objective of this chapter is to test the performance of estimated articulatory trajectories for place of articulation classification. In the first stage, the speech signal was segmented into broad classes using ideal phonetic transcriptions into 5 broad classes (Vowels – V, Fricatives – Fr, Sonorant Consonants – SC, Stops – ST and Silence – SIL). A single feature vector composed of Mel Frequency Cepstral Coefficients (MFCCs) and estimated articulatory trajectories (estTV) were extracted from the broad class segments. Fixed length feature vectors were obtained from variable length segments using a statistical parameterization of the MFCCs and estTVs. The combination of MFCCs with estTVs provided an average of 2% relative

improvement in recognition of the place features compared to MFCCs alone. This work was presented at the Spring 2015 meeting of the Acoustic Society of America (Sivaraman et al., 2015a)

6.1 Datasets and systems

The phonetically rich TIMIT speech dataset was used for performing the place of articulation classification experiments. The training set consisted of 462 speakers while the test set contained 168 speakers not present in the training set. All the sentences (si, sa, and sx) were used in the training and testing. The articulatory features were extracted using the speaker-independent speech inversion system trained on the XRMB dataset. This was the same speech inversion system developed in Chapter 3. Note that there was no speaker overlap between the XRMB dataset and the TIMIT dataset.

6.2 Phonetic feature hierarchy and phone broad classes

The phonemes of American English can be characterized by three general descriptors - source characteristics, manner of articulation, and place of articulation (Juneja, 2004).

1. *Source characteristics:* The source or excitation of speech can be periodic when air is pushed from the lungs at a high pressure that causes the vocal folds to vibrate, or aperiodic when either the vocal folds are spread apart or the source is produced at a constriction in the vocal tract. The source characteristics are

encoded by a binary voiced feature. The voiced feature is assigned a "+" for phonemes which involve periodic excitation by the vocal folds and a "-" for phonemes that do not contain a periodic excitation.

2. *Manner of Articulation:* Manner of articulation refers to how open or close the vocal tract is, how strong the constriction is and whether the there is air flow through the nasal cavity or not. Manner phonetic features are also called articulator-free features (Stevens, 2002) which means that these features are independent of the main articulator and are related to the manner in which the articulators are used. The manner features are shown above the horizontal dashed line (in capital letters) in the Figure 6.1.

3. *Place of Articulation:* The place of articulation phonetic feature determines the place in the vocal tract where the constriction (or shaping) happens in order to produce the sound. For example the /p/, /b/, and /m/ sounds have a labial place of articulation indicating that the sounds are produced by the labial constriction. These phonetic features are called articulator-bound features (Stevens, 2002). The acoustic correlates of the place features are more subtle than that of the manner features (Juneja, 2004). Articulatory features are best suited to estimate the place of articulation features. The place of articulation features are shown below the horizontal dashed line (in capital letters) in the Figure 6.1.

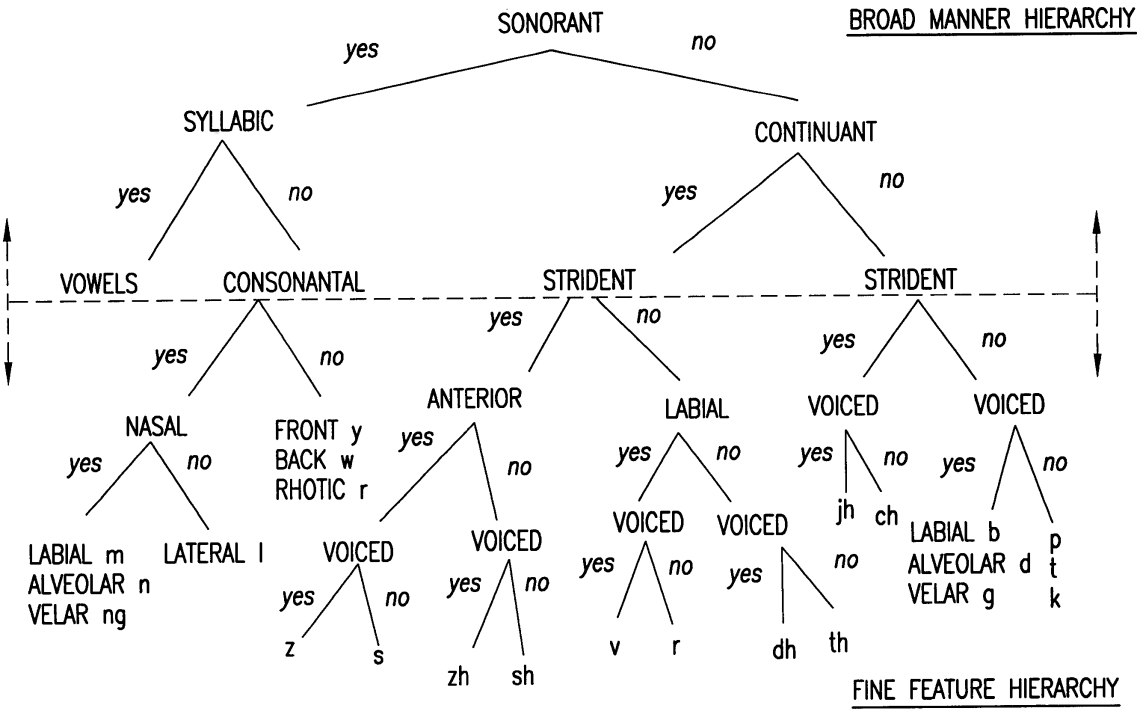


Figure 6.1: Phonetic feature hierarchy for American English phonemes. Figure taken from (Espy-Wilson and Juneja, 2010)

6.3 Place of articulation classification system using acoustic and articulatory features

We developed a broad class segment based place of articulation classification system taking advantage of the articulatory features estimated from speaker independent speech inversion systems. This task was designed to only highlight the efficacy of articulatory features for place of articulation classification. We assume perfect phone segmentation using the TIMIT phone transcriptions. The task here is only to classify the phone segment into different place of articulation features. The phonemes grouped according to manner of articulation are called broad classes. As seen in 6.1, places of articulation are different for different broad classes. In this experiment, we group the TIMIT phonemes into 4 broad classes - Vowels, Sonorant

consonants, Stop consonants, and Fricatives. We use the phone labels to obtain these broad class identities for each phone segment. In future this assumption can be replaced with a phone broad classification system as developed in (Juneja and Espy-Wilson, 2008). The task at hand is to further classify phone segments under each of these broad classes into their place of articulation features. The Table 6.1 shows the list of place of articulation features for each broad class.

Table 6.1: Places of articulation for various broad classes

Broad class	**Places of articulation**
Vowels (V)	Back vs Central vs Front High vs Mid vs Low
Stops (ST)	Bilabial vs Alveolar vs Velar
Fricatives (Fr)	Labial vs Dental vs Alveolar vs Palatal
Sonorant Consonants (SC)	Bilabial vs Alveolar vs Frontal vs Retroflex vs Back vs Velar vs Glottal

The vowels are classified according two sets of place features - front vs mid vs back and high vs mid vs low. The Sonorant consonants were classified into 7 place features - Bilabial vs Alveolar vs Frontal vs Retroflex vs Back vs Velar vs Glottal.

We developed a hierarchical classification system for this classification task. The system used both MFCC acoustic features as well as estimated TVs (estTV) from the XRMB speech inversion system. Figure 6.2 shows the block diagram of the palace of articulation classification system.

As already explained earlier, the system assumes idea broad class segmentation from phone transcriptions of the TIMIT dataset. The broad class segments of

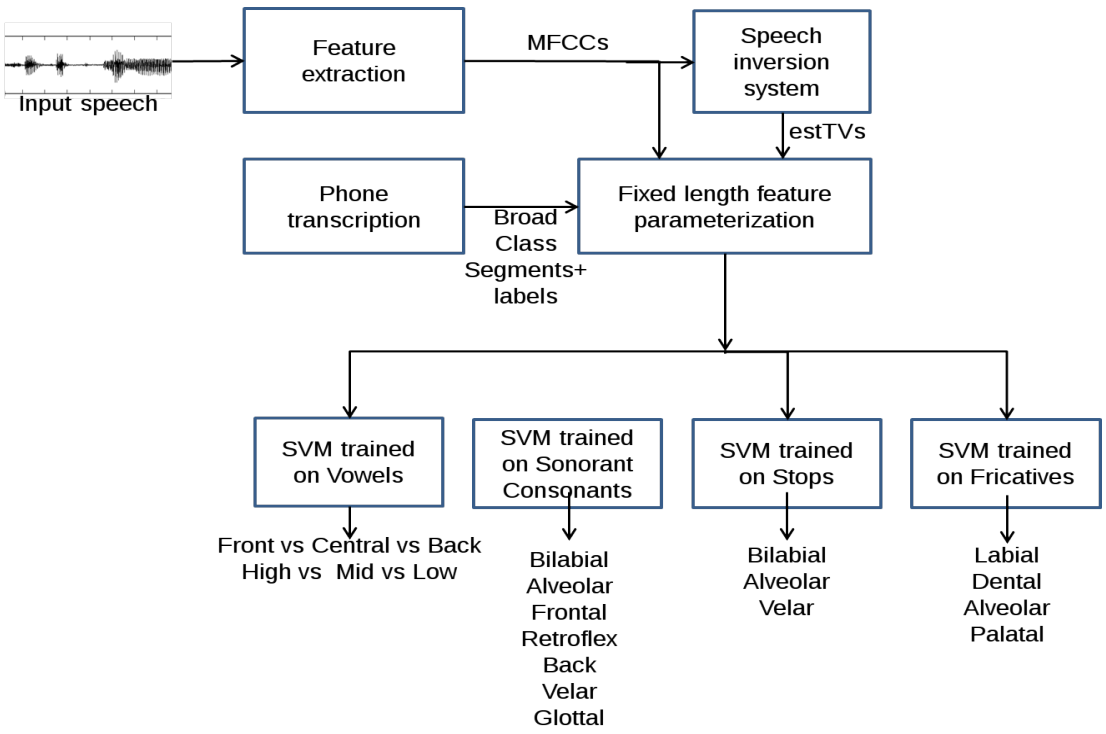


Figure 6.2: Block diagram of place of articulation classification

the TIMIT utterances obtained from the transcription were of different lengths. Classification of place using SVMs required a fixed length feature vector that accurately summarized the MFCCs and estTV features in the broad class segment. Table 6.2 shows the functional features obtained from each component of MFCCs and estTVs for each broad class segment. This functional feature extraction from the frame level MFCC and estTV features provides a fixed length feature vector for each broad class segment. The Support Vector Machine (SVM) classifiers for place classification were trained on the corresponding broad class segments from the TIMIT training set.

Table 6.2: List of fixed length features extracted from segmental frame level MFCCs and TVs

	Description	**Expression**
1	Min	$Min\{Feat_i(t)\}$ across time t
2	Max	$Max\{Feat_i(t)\}$ across t
3	Mean	$Mean\{Feat_i(t)\}$
4	Max slope	$Max\{\frac{dFeat_i}{dt}\}$
5	Min slope	$Min\{\frac{dFeat_i}{dt}\}$
6	Min absolute slope	$Min\{\frac{dFeat_i}{dt}\}$

6.4 Results of phone broad classification

Support Vector Machines (SVM) with Radial basis function kernels were trained to perform the sub-classification of broad class segments. We experimented with different kernel functions and parameters to tune the classifiers. The tuning was performed on a held out 10% subset from the training set broad class segments. A separate SVM was trained for each broad class. We performed classification experiments with different combinations of features namely - MFCCs, estTVs and MFCC+estTV. The bar charts in Figure 6.3 show the classification accuracies for each broad class.

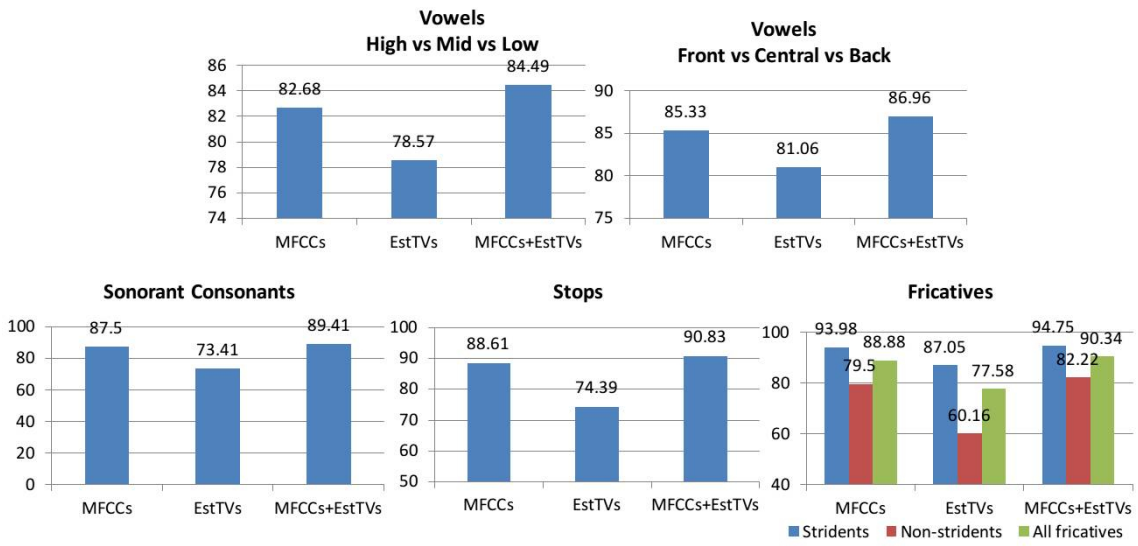


Figure 6.3: Classification accuracies of place of articulation classification

6.5 Summary

Augmenting acoustic features with estimated articulatory features provides an average of 2% relative improvement in accuracy for Vowels, Stops and Sonorant consonants. EstTVs alone do not perform as well as MFCCs. For fricatives, place classification of strident fricatives is more accurate than the non-strident fricatives. Adding contextual information can help in improving the classification accuracy. The articulatory features alone do not provide superior performance compared to acoustic features. Overall, articulatory features combined with acoustic features do help improve the accuracy significantly for classifying the place of articulation of phonemes.

In the future we plan to use an automatic landmark detection system (Juneja and Espy-Wilson, 2008) to segment utterances into broad classes. Instead of using a speech inversion system trained on complete utterances, we plan to train specific

speech inversion systems for each broad class. Features from such tuned speech inversion systems might be more accurate than a generic system. This kind of a place classification system can be combined with the broad class probability estimation system to decode phone sequences. This kind of system will be a completely acoustic phonetic approach to phone recognition.

Chapter 7

Speech recognition experiments incorporating articulatory representations

This chapter deals with the evaluation of articulatory features for continuous speech recognition tasks. Studies have shown that articulatory information can help model speech variability and, consequently, improves speech recognition performance. Studies have explored using DNNs (Mitra et al., 2010b)(Mitra, 2010)(Uria et al., 2011)(Canevari et al., 2013) for learning the nonlinear inverse transform of acoustic waveforms to articulatory trajectories (a.k.a. speech-inversion or acoustic-to-articulatory inversion of speech). Results have demonstrated that using articulatory representations in addition to acoustic features improves phone recognition (Badino et al., 2016)(Badino et al., 2016)(Mitra et al., 2011)(Deng et al., 1997) and speech recognition performance (Mitra et al., 2010b)(Mitra et al., 2014a). However most studies have focused on using articulatory features derived from speech inversion systems trained on one or two speaker datasets, or synthetic speech. In this work, we have trained speech inversion systems with datasets containing multiple speakers and also converted the raw Cartesian coordinates of articulator positions into tract variables (TVs) which are relative constriction measures as explained in 3. In this chapter ASR experiments have been performed on the Wall Street Journal (WSJ) corpus using state-of-the-art ASR architectures. The ASR

systems were evaluated across corpus and domains using the TIMIT and EMA-IEEE datasets. Articulatory features estimated from the XRMB speech inversion system and the EMA-IEEE speech inversion system were fused together with acoustic features represented as Gammatone filterbank energies. Apart from the estimated TVs a voicing probability feature from the Kaldi pitch estimator (Ghahremani et al., 2014) was appended to the TVs to account for the missing glottal articulatory feature. In order to highlight critical regions of the TVs, a kinematics based binary features representing articulatory gesture activations were extracted from the estimated TVs and appended to the articulatory feature stream. Experiments and results show the impact of these features on the ASR performance. In order to optimally combine acoustic and articulatory feature streams, this work proposes a hybrid convolutional neural network (HCNN), where two parallel layers are used to jointly model the acoustic and articulatory spaces, and the decisions from the parallel layers are fused at the output context-dependent (CD) state level. The acoustic model performs time-frequency convolution on filterbank- energy-level features, whereas the articulatory model performs time convolution on the articulatory features. Previous results using synthetic TVs have demonstrated that HCNN-based model achieves lower word error rates compared to the CNN/DNN based systems. Results show that combining articulatory features with acoustic features through the hybrid CNN improves the performance of the ASR system in matched and mismatched evaluation conditions.

7.1 Datasets for ASR experiments

7.1.1 Wall Street Journal

The DARPA WSJ1 CSR dataset was used in the experiments presented in this chapter. For training, a set of 35,990 speech utterances (77.8 hours) from the WSJ1 collection, having 284 speakers was used. For testing, the WSJ-eval94 dataset composed of 424 waveforms (0.8 hours) from 20 speakers was used. Note that for all the experiments reported here, speaker-level vocal tract length normalization (VTLN) was not performed. We denote this dataset as WSJ1 in our experiments described in this chapter.

7.1.2 TIMIT

TIMIT is a widely used corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of 8 major dialects of American English, each reading 10 phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance. Out of the 630 speakers in the datset, 462 speakers' data is usually assigned for training and the remaining 168 speakers are used for testing. For our experiments in this chapter, we have used the TIMIT test set for cross corpus evaluation of the ASR systems trained on the WSJ1 train set.

7.1.3 EMA-IEEE dataset

The EMA-IEEE dataset is the same dataset described in 4.3.1. It consists of 720 phonetically balanced Harvard sentences (Rothauser et al., 1969) recorded from eight speakers (4 males, and 4 females) at normal and fast production rates. The dataset also consists of parallel EMA recordings. As described in 4.3.3, the 720 Harvard sentences were randomly divided into 3 subsets for training, cross-validation and testing. The training subset contained 576 sentences while the test and cross-validation sets contained 72 sentences each. With this split of sentences, we created train, cross-validation and test sets for each of the 8 speakers and the 2 speaking rates. Thus, we created 8 pairs (Normal and Fast rates) of subsets from the EMA-IEEE dataset. For our experiments in this chapter we used only the test utterances from the EMA-IEEE dataset to perform cross-corpus evaluation of the acoustic models trained on WSJ1. We split the EMA-IEEE test set into two subsets based on speaking rates - EMA-IEEE-F consisting of fast rate utterances and EMA-IEEE-N consisting of normal rate utterances. The evaluation of the acoustic models were carried out separately for each of these subsets thus highlighting the effect of mismatched speaking rate on ASR accuracy.

7.2 Acoustic and articulatory features

7.2.1 Acoustic features - Gammatone Filterbanks energies

The Gammatone filters are a linear approximation of the auditory filterbank of the human ear. In GFB processing, speech is analyzed by using a bank of 40 gammatone filters equally spaced on the equivalent rectangular bandwidth (ERB) scale. For this work, the power of the bandlimited time signals within an analysis window of 26ms was computed at a frame rate of 10ms. Subband powers were then root compressed by using the 15th root, and the resulting 40-dimensional feature vector was used as the GFB. It was shown (Mitra et al., 2014b) that CNNs give lower WERs compared to DNNs when using filterbank features for the Aurora-4 ASR task, and GFBs offered performance gain over mel-filterbank energies (MFBs). This observation was sufficient to safely assume that GFB is a strong baseline acoustic feature for our experiments in this chapter.

7.2.2 Articulatory features - Estimated TVs

Articulatory features in the form of TVs were estimated from acoustics using speaker independent speech inversion systems trained in Chapter 3 and Section 4.3.3. We used TVs estimated by two different speech inversion systems - (1) the XRMB speech inversion system (Chapter 3) and (2) the EMA-IEEE speech inversion system (Section 4.3.3). The XRMB system was trained on 36 speakers from the XRMB dataset while the EMA-IEEE system was trained on both fast and normal utterances from the EMA-IEEE training set. Note that none of the utterances in the WSJ1

dataset or the TIMIT and EMA-IEEE test sets were included in the training of the speech inversion systems. For the XRMB system, there was no overlap between the speakers of the XRMB train and test sets. For the EMA-IEEE dataset all the test speakers were in the training set but the utterances were mutually exclusive. We will refer to the TVs estimated by the XRMB system as XTV and those estiamted by the EMA-IEEE system as ETV

7.2.3 Voicing probability

The Kaldi Pitch tracker (Ghahremani et al., 2014) comes with the Kaldi speech recognition toolkit (Povey et al., 2011) and provided two-dimensional output consisting of pitch tracks and a normalized cross-correlation function that gave an indication about voicing information. We converted the cross-correlation function into voicing probability estimates that ranged from 0 to 1. This one dimensional voicing probability feature was extracted for the WSJ1 dataset, XRMB test set, and the EMA-IEEE test set. We will refer to this feature as *vad*.

7.2.4 Articulatory gestural activations

As motivated by Articulatory phonology, articulatory gestures are action units of the articulators wheich actuate the movements of the articulators to produce speech. Gestures are like movement primitives (Ramanarayanan et al., 2013) and are precursors to TVs. Studies have shown that articulatory gestures derived from EMA data preserve discriminatory information about phone categories (Ramanarayanan

et al., 2015). In this work we extracted gesture linke binary activation features from the estimated TVs to highlight the critical regions of the articulations. Since articulatory features are less variable (and more accurately predictable) in the critical regions, we developed binary activation gesture features that highlight the regions where the constriction actions happen. For example, for the Tongue tip, the critical articulation is the constriction with the palate. Hence, a Tongue tip gesture will highlight regions where the tongue tip is performing an action of constriction by moving towards the palate before the instant of constriction and away from it after the constriction. The movement towards the constriction instant is called a gestural onset whereas the movement following the constriction is called the gestural offset. We define gestural activation as the time duration spanning the gestural onset and offset. The lengths of the onset and offset windows are determined by defining a threshold on the velocity of the associated TV approaching a gesture and departing from the gesture. Table 7.1 shows the list of articualtory gestures defined from the TVs.

Table 7.1: Articulatory Gestures and their associated articulators

Gesture name	**Associated TVs**	**Articulators**
Lip Aperture (LA)	LA	Upper & Lopwer lips
Tongue Tip (TT)	TTCD, TTCL	Tongue tip
Tongue Middle (TM)	TMCD, TMCL	Tongue middle
Tongue Back (TB)	TBCD, TBCL	Tongue back
Jaw Angle (JA)	JA	Jaw, Lower incisors, Upper lip

Figure 7.1: Example plot of gestural activations and TVs for a TIMIT utterance - *"the reasons for this dive seemed foolish now"*

Figure 7.2 shows a plot of the estimated TVs and their annotated gestures for an example utterance from the TIMIT dataset.

We extracted 3 gestures - LA, TM, and TT from the TVs estimated by the XRMB TV estimator and 5 gestures - LA, JA, TT, TM, and TB from the TVs estimated by the EMA-IEEE TV estimator. We will refer to the gestures estimated from XTV as XG and those from the ETV as EG. These articulatory gestural activations were appended to the estimated TVs to form the complete set of articulatory features used for ASR. In total, the articulatory features all put together (XTV+vad+XG+ETV+EG) were 24 dimensional.

7.3 ASR system architecture

We trained different acoustic models for the WSJ1 dataset with different deep (DNN) and convolutional neural network (CNN) architectures.

Time Frequency Convolutional Neural Networks (TFCNN): Apart from the basic feedforward DNN architecture, we explored Time-Frequency Convolutional Neural Netowkrs (TFCNN) for the acoustic features used to train the acoustic models. Figure shows a schematic of the TFCNN architecture. The TFCNN architecture was based upon (Mitra and Franco, 2015), where two parallel convolutional layers were used at the input, one performing convolution across time, and the other across the frequency scale of the input filterbank features. The results of that work showed that the TFCNNs performed better compared to their CNN counterparts. Here, we used 75 filters to perform time convolution, and 200 filters to perform frequency convolution. We used the optimal configuration learned from (Mitra and Franco, 2015) for the experiments reported in this chapter. For time and frequency convolution, eight bands were used. A max-pooling over three samples was used for frequency convolution, while a max- pooling over five samples was used for time convolution. The feature maps after both the convolution operations were concatenated and then fed to a fully connected neural net, which had 1024 nodes and four hidden layers. This architecture was used to train an acoustic feature only ASR system using GFB as the features.

Hybrid Convolutional Neural Networks (HCNN): In order to optimally combine the acoustic and articulatory features in an ASR architecture, we developed a modified deep neural network architecture to jointly model the acoustic and the articulatory space. The following description of the HCNN architecture is taken from the paper (Mitra et al., 2017) coauthored by the author of this dissertation.

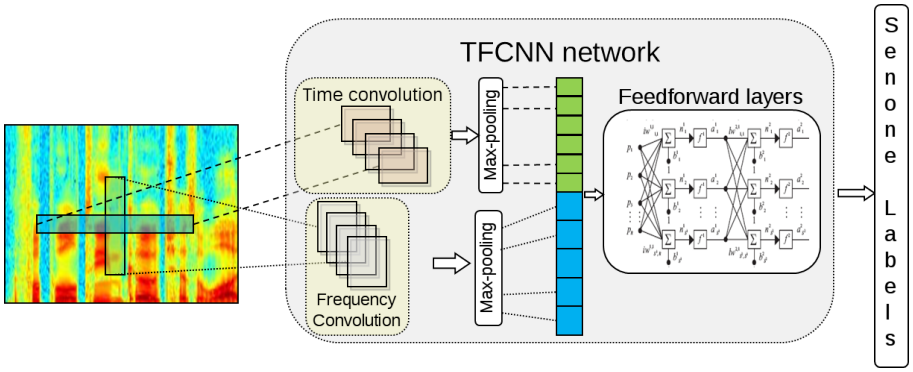


Figure 7.2: Block diagram showing time-frequency convolution neural nets (TFCNN). The top block shows convolution filters working across time, and the bottom dotted block shows convolution filters working across frequency. The max-pooled outputs of these convolution filters are fed to a fully connected four-layered deep neural net. (Mitra and Franco, 2015)

The diagram of the network is shown in Figure 7.3, illustrating two parallel neural networks trained simultaneously. These two parallel neural networks modeled two things: (1) learning the acoustic space from the GFB features and (2) learning the articulatory space from the TV trajectories. The acoustic space was learned by using a time-frequency convolution layer, where two separate convolution filters operate on the input GFB features. These two convolution layers had the same parameter specification as that used in the TFCNNs. The articulatory space was learned by using a time-convolution layer that contained 75 filters, followed by a max-pooling over five samples. Note that the cross-TV convolution operation may not produce any meaningful information, whereas time convolution on the TVs can help in extracting TV modulation-level information, which was the motivation behind selecting a time-convolution layer for learning the articulatory space. The fully connected DNN layers were different in size; we observed that 800 neurons was nearly optimal for learning the acoustic space, and that 256 neurons was nearly

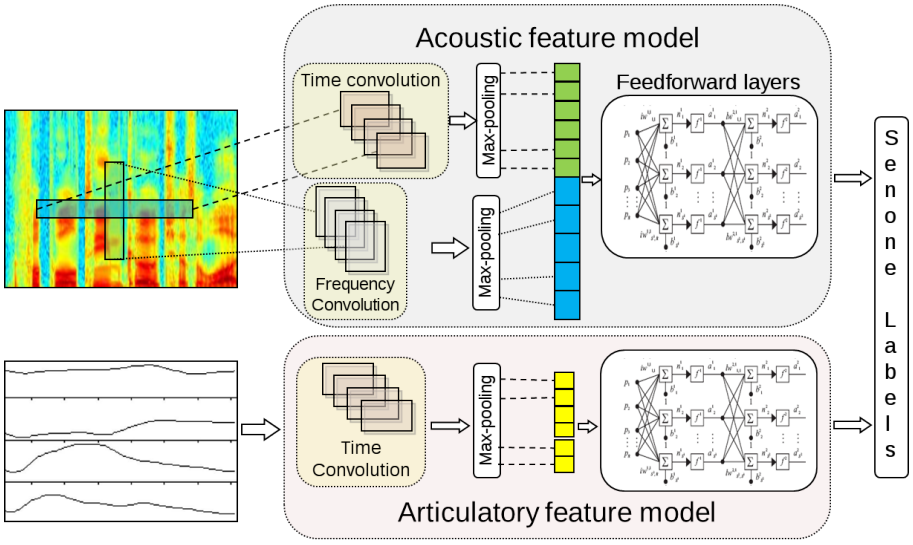


Figure 7.3: Schematic of the hybrid convolutional neural network (HCNN). The top layer represents the acoustic model, whose input is filterbank features, and the bottom layer represents the articulatory model, whose input is TV trajectories. (Mitra et al., 2017)

optimal for learning the articulatory space. Note that the parallel networks were jointly trained. We used this HCNN architecure for our ASR experiments with articulatory features.

7.4 Experiments and results on the WSJ dataset

In order to generate the alignments necessary for training the CNN system, a Gaussian mixture model (GMM)–hidden Markov model (HMM) model was used to produce the senones' labels. Altogether, the GMM-HMM system produced 1659 context-dependent (CD) states for WSJ1. The input features to the acoustic models were formed by using a context window of 15-95 frames (half of the frames on either side of the current frame). We performed experiments by varying the context window from 15 to 95 frames in steps of 20. The acoustic models were trained by using

cross-entropy on the alignments from the GMM-HMM system. The output layer of the networks included as many nodes as the number of CD states for the given dataset. The networks were trained by using an initial four iterations with a constant learning rate of 0.008, followed by learning-rate halving based on cross-validation error decrease. Training stopped when either no further significant reduction in cross-validation error was noted or when cross-validation error started to increase. Backpropagation was performed using stochastic gradient descent with a mini-batch of 256 training examples. For the DNN systems, we used five layers with 1024 neurons in each layer, with similar learning criteria as the CNNs.

Table 7.2 shows the Word Error Rates (WER) on the WSJ1 evaluation set for different feature combinations. We observe that the WER remains the same for all the feature combinations. The fact that the articulatory features did not help improve the WER indicates that the concatenation of GFB with the articulatory features though a DNN model is not an optimal combination of the features.

Table 7.2: Word Error Rates on WSJ eval set for DNN acoustic models

Features	**WER**
GFB	6.0
GFB+XTV+vad	6.0
GFB+XTV+vad+XG+ETV+EG	6.0

We experimented with different splicing widths of the input features to figure out the best splicing widths for the features. The plot in figure 7.4 shows the Word Error Rates (WER) on the development set of WSJ1 for different splicing widths and

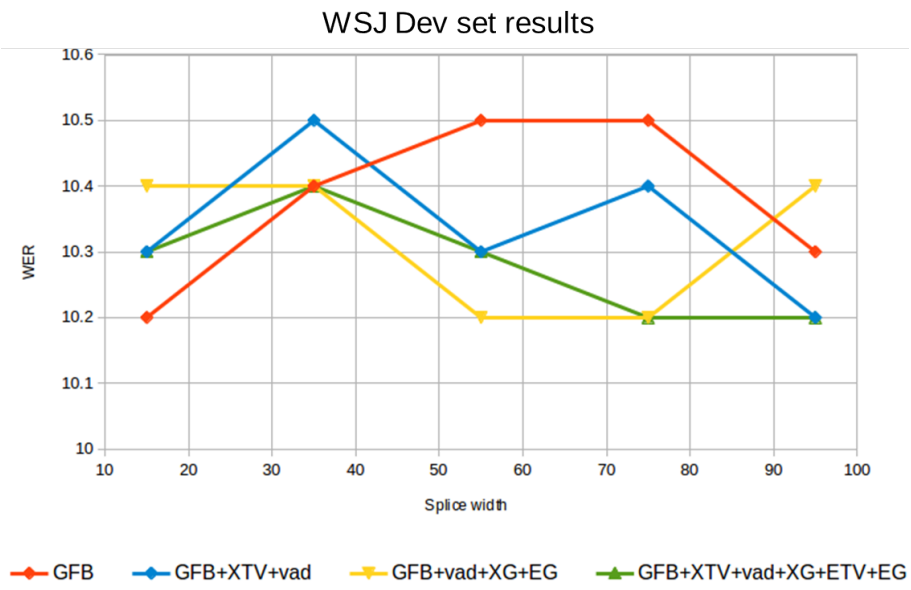


Figure 7.4: WER on the WSJ1 dev set for the HCNN model at different splicing widths for various feature combinations

feature combinations. Figure 7.5 shows the WER results on the WSJ1 evaluation set for the same set of splicings. We observed that a splicing width of 95 frames was optimal for the GFB+XTV+vad+XG+ETV+EG feature combination. It reduced the baseline WER of 10.3% to 10.2% on the WSJ dev set. The WER improvement on the dev set translated to a 0.2% reduction in the evaluation set WER. The GFB TFCNN system is already very strong with an impressive 5.6% WER on the WSJ1 eval set. Even small improvements over this highly competitive performance with the GFB features is quite challenging. The feature combination containing GFB and articulatory gestures alone (GFB+vad+XG+EG) did not give superior performance compared to GFB.

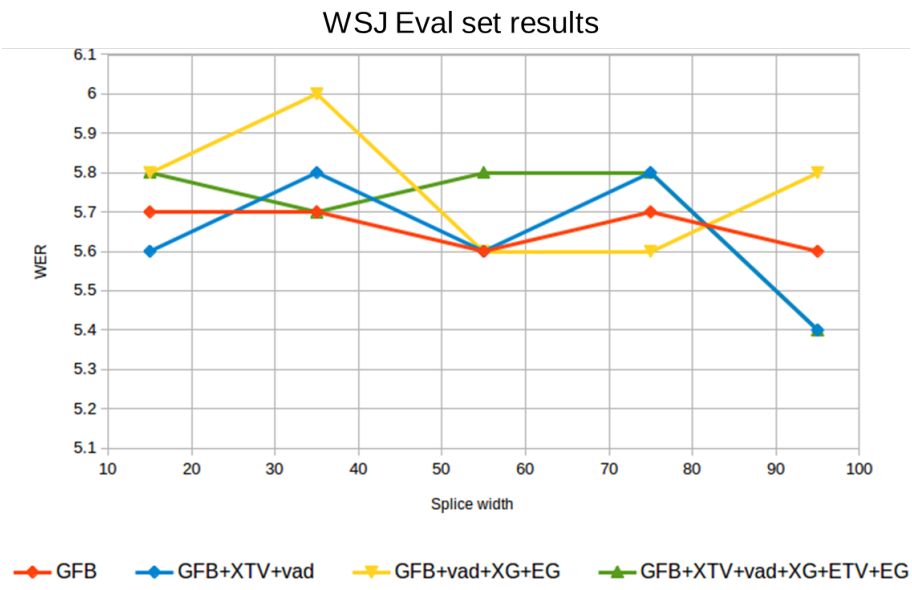


Figure 7.5: WER on the WSJ1 eval set for different splicing widths for various feature combinations

7.5 Results of cross-corpus testing

In order to test the accuracy of the acoustic models across corpora, we evaluated the WSJ1 trained acoustic models on the TIMIT (timit_test), EMA-IEEE-Normal (ema_n_test) and EMA-IEEE-Fast (ema_f_test) datasets. Since these datasets contained some words not included in the WSJ1 corpus, we chose to evaluate the phone recognition accuracy instead of complete sentences (word sequences). We used a trigram phone language model trained on the WSJ1 train set to decode the hypotheses. Although the language models are biased towards the WSJ1 corpus we are interested in the improvement of the acoustic model hypothesis for same language models. Table 7.3 shows the Phone Error Rates (PER) for the cross-corpus evaluations. The best PERs for each of the evaluation sets are highlighted in bold.

Table 7.3: Phone Error Rates on different test sets for the HCNN model for two best performing splicing widths (75 and 95)

	Splice = 75			**Splice = 95**		
	GFB	GFB+XTV+vad+ETV	GFB+XTV+vad+XG+ETV+EG	GFB	GFB+XTV+vad+ETV	GFB+XTV+vad+XG+ETV+EG
WSJ_dev	**13.6**	13.7	14.0	14.0	**13.6**	**14.0**
WSJ_eval	**14.6**	14.7	14.9	14.7	**14.5**	14.9
timit_test	**30.6**	**30.4**	30.5	30.9	30.7	31.0
ema_n_test	**36.8**	**36.3**	36.8	37.2	36.9	37.0
ema_f_test	**53.5**	**53.5**	53.6	54.5	53.9	**54.0**
Average	**29.8**	**29.7**	30.0	**30.3**	**29.9**	30.2

7.6 Summary

In this chapter, we presented DNN and CNN based acoustic models for WSJ1 dataset using acoustic and articulatory features. We developed and explored a hybrid convolutional neural network (HCNN) architecture for optimally combining the acoustic and articulatory feature streams. Experiments on WSJ1 dataset showed that the HCNN system combining all the articulatory features and the GFB acoustic feature reduced the WER by 0.2% over the acoustic only (GFB) TFCNN acoustic model, and by 0.6% over the DNN acoustic model trained on GFB acoustic feature. Cross-corpus phone recognition with the WSJ1 acoustic models highlighted the reduction in performance in mismatched corpus condition. Adding articulatory features to acoustic features reduced the PER by 0.1% on an average with a splicing width of 75 frames and by 0.4% an average for a feature splicing width of 95 frames.Thus articulatory features (TVs and gestures) can be used in state-of-the-art ASR systems to improve the performance.

The articulatory gestures that we estimated in this experiment was using the

TVs estimated from the acoustics. The errors in the TV estimation gets propagated to the estimated gestures as well. This might be a reason due to which the gestures did not help much in improving the ASR performance. Estimating the binary gesture features directly from speech would probably help improve the ASR performance. Further explorations into effective parametrization of the articulatory features could further improve the performance.

The results of the cross-corpus evaluations show the sensitivity of the acoustic models and is an important area of future ASR research.

Chapter 8

Summary and future work

8.1 Summary

This dissertation analyzes the inherent variability of speech due to speaker voice, speaking rates, and accents from the perspective of articulatory phonology. It then proposes a set of articulatory representations and articulatory feature based speech recognition architectures that can help to deal with the speech variability in an Automatic Speech Recognition (ASR) system. Articulatory phonology provides a unified framework to represent speech as a constellation of coordinated gestures whose temporal overlap and spatial reduction explain coarticulation and lenition. In order to represent speech in terms of articulatory gestures, the gestures needed to be estimated from the speech signal. Chapter 3 of this dissertation dealt with the development of speaker independent acoustic-to-articulatory speech models that transformed acoustic representations to their articulatory counterparts and hence their gestural representations, using real speech and articulatory datasets. In contrast with the previously published work on speech inversion using synthetic or limited real speaker datasets, this thesis deals with a much more challenging problem of training a single speaker independent speech inversion system using articulatory data from multiple speakers (46 speakers) from the XRMB articulatory dataset. In order to suppress the anatomical differences between the speakers we converted

the flesh point X-Y positions of the articulators in the XRMB dataset to relative measures of articulatory constrictions known as Tract Variables (TV). Feed-forward neural network architectures were explored for the acoustic to articulatory mapping. A thorough analysis by tuning different neural network parameters was carried out to figure out the best architecture for the speech inversion system. These experiments led to a 5 hidden layer feedforward neural network with 512 nodes in each layer. In order to smooth the estimates of the neural network, a low pass filtering Kalman smoother was used. This speech inversion system which we refer to as XRMB speech inversion system was further used for various experiments in different parts of the thesis. The XRMB speech inversion system estimates the TVs with correlation of 0.782 between the actual and estimated TVs on a held out set of 5 speakers from the XRMB dataset. The performance of the speech inversion system varied significantly across speakers. In chapter 4 we studied and quantified this variability by analyzing cross speaker performance of speaker dependent speech inversion systems. To address the impact of speech inversion performance due to speaker variability, we developed a Vocal Tract Length Normalization (VTLN) based speaker adaptation system. The speaker adaptation approach provided a 7% absolute improvement in correlation performance of the speech inversion systems. Speaking rate is another kind of variability that we analyzed and addressed in this dissertation. In order to study the effects of speaking rate on acoustics and articulations we analyzed a variable speaking rate dataset containing Electromagentic Articulometry (EMA) data. The dataset consisted of 8 speakers producing the IEEE sentences at normal and fast rates. We performed

extensive speech inversion experiments like cross speaking rate and cross speaker experiments to assess the performance across speaking rates on this dataset. The cross speaking rate correlations of the TV estimates provided a quantitative analysis of variability in acoustic and articulatory spaces due to speaking rate. Next we analyzed the variability due to accents by performing cross accent and even cross language speech inversion experiments on a Dutch, Dutch accented English and British English articulatory dataset. The cross accent results showed that speech inversion systems trained on a particular accented speech can be used to reliably estimate the articulaotry trajectories of a different accent with some reduction in performance compared to matched accent conditions. Coarticulation and lenition are commonly observed in conversational and fast rate speech. They often manifest as deletion or substitution of phone units when looked at from the acoustic perspective. Articulatory phonology explains coarticulation through spatio-temporal changes and overlap in the patterns of underlying gestures. However these spatio-temporal overlap in the articulatory gestures are weakly manifested in the acoustics thus leading to apparent deletion or substitution. In chapter 6 we studied the coarticulation occurring in certain fast spoken utterances using articulatory constriction tract-variables (TVs) estimated from acoustic features. The objective of this study was to test the efficacy of the speech inversion system (XRMB system) in estimating acoustically weak articulatory maneuvers like coarticulations. Specifically, we studied three example utterances of fast and normal rate speech to estimate the articulatory gestures. We showed that the speech inversion systems which were never trained on the target utterances reliably estimated the TVs that

matched the actual TVs from the corresponding articulatory measurements. We also considered the special case of American English /r/ sound, the production of which has two possible tongue configurations - the bunched and the retroflexed (Zhou et al., 2008). It is known that the acoustic correlates of the differences in these tongue configurations lie in the 4th and 5th formant frequencies which are acoustically weak signatures. We showed that the XRMB speech inversion shystem could accurately distinguish between the bunched and the retroflex configurations even though the examples considered were not part of the XRMB dataset. The analyses carried out in chapters 4 and 5 indicate that articulatory features are a more invariant representation of speech compared to the acoustic features and hence hold potential to improving the performance of ASR systems Chapter 6 considers a proof of concept experiment of classifying place of articulations of phonemes using estimated articulatory features. Phone place of articulation classification experiments were carried out using the TIMIT dataset. Results showed that a combination of acoustic features (MFCCs) and estimated TVs improve the place of articulation classification accuracy by 2% relative to the accuracy of MFCCs alone. Encouraged by this observation we carried out full fledged continuous speech recognition experiments in chapter 7. In chapter 7 we carry out medium vocabulary speech continuous recognition experiments on the Wall Street Journal 1 (WSJ1) dataset. We experimented with various combinations of acoustic and articulatory features. The acoustic features were characterized by Gammatone Filterbank energies (GFB) and the articulatory features were estimated using the XRMB and the EMA-IEEE speech inversion systems. We also developed binary articulatory

gesture like activation features to highlight the critical regions of articulatory constrictions from the TVs. In order to best combine the acoustic and articulatory features we developed a hybrid convolutional neural network (HCNN) architecture for the ASR system. The HCNN architecture performed time and frequency convolutions on the acoustic features and time convolutions on the articulatory features. We compared the performance of the HCNN acrhitecture with the Time Frequency Convolutional neural network (TFCNN) and DNN architectures operatin on acoustic only GFB features. Results showed that the HCNN system combining all the articulatory features and the GFB acoustic feature reduced the WER by 5.2% relative to the acoustic only (GFB) TFCNN acoustic model, and by 0.6% over the DNN acoustic model trained on GFB acoustic feature. Cross-corpus phone recognition with the WSJ1 acoustic models highlighted the reduction in performance in mismatched corpus condition. Adding articulatory features to acoustic features reduced the PER by 0.1% on an average with a splicing width of 75 frames and by 0.4% an average for a feature splicing width of 95 frames.

The findings based on this dissertation indicate that articulatory representations extracted from acoustics can be used to address acoustic variability in speech observed due to speakers, accents, and speaking rates and further be used to improve the performance of Automatic Speech Recognition systems.

8.2 Future directions

There are several directions of research that could be pursued based on the findings of this dissertation.

8.2.1 Consolidating multi-modal articulatory data for speech inversion

Articulatory data is being collected by different groups using different modalities like EMA, rt-MRI, ultrasound, and EMG. All these datasets measure/image different regions of the vocal tract with different temporal and spatial resolutions. Since each of these techniques are expensive and time consuming, in a particular dataset only a limited amount of data is recorded from a few speakers. Methods developed in this thesis to convert EMA trajectories to TVs and speaker normalization offer promise for combination of articulatory data from multiple modalities for speech inversion training. This would effectively augment the size of the articulatory dataset as well as the diversity.

8.2.2 Assistive devices for pronunciation training

The acoustic to articulatory speech inversion systems developed in this thesis show promise that that the articulatory movements can be accurately estimated from speech. A real-time implementation of the XRMB speech inversion system during the course of the thesis makes such a system suitable for practical use in pronunciation training. The estimated TVs from a speech inversion system could be

used in creating a 3-D visualization of the articulatory movements by actuating a 3-D vocal track model using the estimated TVs. Such a visualization system could potentially help second language learners to improve their pronunciations. Often certain sounds (e.g., the liquids /r/ and /l/ in english) are difficult to produce in a given language. Subjects speaking a non-native language may fail to reach the target articulation or may use an incorrect articulation that results in substitution of sounds. Given that the TVs can be estimated form the subjects' speech signal, providing visual comparison of what they are doing with their articulators and what they should be doing to produce a canonical articulation of the sound or phrase. This corrective feedback mechanism holds promise in pronunciation training.

8.2.3 Speech Synthesis

Articulatory features have been explored for synthesis of accent normalized speech. With the advent of deep recurrent neural networks for speech synthesis articulatory features can effectively be incorporated in a speech synthesis system. Modeling articulatory trajectories and gestures as latent variables in a speech synthesis system can potentially enable a speech synthesis system to produce accented speech. A recurrent neural network based articulatory synthesizer can be implemented to map phoneme orthographic transcriptions to TVs and gestures just like the dynamical system implementation in the Task Dynamics and Articulations (TADA) system. Such a system would be a robust data driven articulatory synthesizer and further the understanding of articulatory speech synthesis

8.2.4 Accent normalization for ASR

Variability due to speech accents pose a great challenge to ASR systems. Current ASR systems deal with accents by creating accent specific acoustic models or performing accent adaptation. Articulatory features are an invariant representation of speech that offer simple and intuitive understanding of accents. Accents are manifested due to the inaccuracy in reaching articulatory targets and changes in timings between contending articulatory targets. With the knowledge of articulatory variations in accented speech a system can be developed to normalize accents to improve the recognition accuracy of ASR systems for accented speech. Future experiments need to be performed to evaluate the cross accent ASR performance of articulatory features and develop methods for accent normalization.

Bibliography

Afshan, A. and Ghosh, P. K. (2015). Improved subject-independent acoustic-to-articulatory inversion. *Speech Communication*, 66:1–16.

Arora, R. and Livescu, K. (2013). Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 7135–7139.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, 63(5):1535–1555.

Badino, L., Canevari, C., Fadiga, L., and Metta, G. (2016). Integrating articulatory data in deep neural network-based acoustic modeling. *Computer Speech and Language*, 36:173–195.

Bengio, Y. and Lecun, Y. (2007). Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, J. W., editor, *Large-scale kernel machines*, pages 321–360. MIT Press.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11):763–786.

Browman, C. P. and Goldstein, L. (1992). Articulatory Phonology : An Overview *. *Phonetica*, 49:155–180.

Canevari, C., Badino, L., Fadiga, L., and Metta, G. (2013). Relevance-weighted-reconstruction of articulatory features in deep-neural-network-based acoustic-to-articulatory mapping. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.

Cavin, M. (2015). The use of ultrasound biofeedback for improving English/r. *Working Papers of the Linguistics Circle*, 25(1):32–41.

Çetin, Ö., Kantor, A., King, S., Bartels, C., Magimai-Doss, M., Frankel, J., and Livescu, K. (2007). An articulatory feature-based tandem approach and factored observation modeling. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 4, pages IV–645–IV–648.

Chen, W. S. and Alwan, A. (2000). Place of articulation cues for voiced and voiceless plosives and fricatives in syllable-initial position. *Proceedings of the 6th International Conference of Spoken Language Processing*, 4:113–116.

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper & Row.

Delattre, P. and Freeman, D. C. (1968). A dialect study of American English r's by x-ray motion picture. *Linguistics*, 6(44):28–69.

Denes, P. B. and Pinson, E. N. (2015). *The Speech Chain: The Physics and Biology of Spoken Language, Second Edition*. Bell Telephone Laboratories.

Deng, L., Ramsay, G., and Sun, D. (1997). Production models as a structural basis for automatic speech recognition. *Speech Communication*, 22(2-3):93–111.

Diehl, R. L., Lotto, A. J., and Holt, L. L. (2004). Speech Perception. *Annual Review of Psychology*, 55(1):149–179.

Dusan, S. V. and Deng, L. (2000). Acoustic-to-Articulatory Inversion using Dynamical and Phonological Constraints. In *5th Seminar on Speech Production*, pages 237–240.

Espy-Wilson, C. and Juneja, A. (2010). System and method for automatic speech recognition from phonetic features and acoustic landmarks.

Espy-Wilson, C. Y. (1987). *An Acoustic-Phonetic Approach to Speech Recognition : Application to the Semivowels*. PhD thesis, Massachusetts Institute of Technology.

Espy-Wilson, C. Y. (1994). A feature-based semivowel recognition system. *The Journal of the Acoustical Society of America*, 96(1):65–72.

Espy-Wilson, C. Y. and Boyce, S. (1999). The relevance of F 4 in distinguishing between different articulatory configurations of American English /r/. *The Journal of the Acoustical Society of America*, 105(2):1400–1400.

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., and Alwan, A. (2000). Acoustic modeling of American English /r/. *The Journal of the Acoustical Society of America*, 108(1):343–56.

Frankel, J., Wester, M., and King, S. (2007). Articulatory feature recognition using dynamic Bayesian networks. *Computer Speech and Language*, 21(4):620–640.

Ghahremani, P., Babaali, B., Povey, D., Riedhammer, K., Trmal, J., and Khudanpur, S. (2014). A pitch extraction algorithm tuned for automatic speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 2494–2498. Institute of Electrical and Electronics Engineers Inc.

Ghosh, P. K. and Narayanan, S. (2010). A generalized smoothness criterion for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 128(4):2162–72.

Ghosh, P. K. and Narayanan, S. S. (2011). A subject-independent acoustic-to-articulatory inversion. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 4624–4627. IEEE.

Girin, L., Hueber, T., and Alameda-Pineda, X. (2017). Extending the Cascaded Gaussian Mixture Regression Framework for Cross-Speaker Acoustic-Articulatory Mapping. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 25(3):662–673.

Guenther, F. H., Espy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M., and Perkell, J. S. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *The Journal of the Acoustical Society of America*, 105(5):2854–2865.

Hagiwara, R. E. (1995). *Acoustic realizations of American /r/ as produced by women and men*. PhD thesis, University of California Los Angeles.

Halle, M. and Stevens, K. M. (1962). Speech recognition: a model and a program for research. *IRE Transactions on Information Theory*, 8(2):155–159.

Hanson, H. M. and Stevens, K. N. (2002). A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn. *The Journal of the Acoustical Society of America*, 112(3 Pt 1):1158–82.

Hardcastle, W. J. (1972). The use of electropalatography in phonetic research. *Phonetica*, 25(4):197–215.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hiroya, S. and Honda, M. (2004). Estimation of Articulatory Movements From Speech Acoustics Using an HMM-Based Speech Production Model. *IEEE Transactions on Speech and Audio Processing*, 12(2):175–185.

Hueber, T., Girin, L., Alameda-Pineda, X., and Bailly, G. (2015). Speaker-adaptive acoustic-articulatory inversion using cascaded Gaussian mixture regression. *IEEE Transactions on Audio, Speech and Language Processing*, 23(12):2246–2259.

Ji, A. (2014). *Speaker Independent Acoustic-To-Articulatory*. PhD thesis, Marquette University.

Ji, A., Johnson, M. T., and Berry, J. J. (2016). Parallel Reference Speaker Weighting for Kinematic-Independent Acoustic-to-Articulatory Inversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1865–1875.

Juneja, A. (2004). *Speech recognition based on phonetic features and acoustic landmarks*. PhD thesis, University of Maryland College Park.

Juneja, A. (2012). A comparison of automatic and human speech recognition in null grammar. *The Journal of the Acoustical Society of America*, 131(3):EL256–EL261.

Juneja, A. and Espy-Wilson, C. (2008). A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition. *The Journal of the Acoustical Society of America*, 123(2):1154–68.

Kirchhoff, K. (1999). Robust speech recognition using articulatory information.

Kobayashi, T., Yagyu, M., and Shirai, K. (1991). Application of neural networks to articulatory motion estimation. In *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 489–492 vol.1. IEEE.

Lieberman, P. (1970). Towards a Unified Phonetic Theory. *Linguistic Inquiry*, 1(3):307–322.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15.

Liu, P., Yu, Q., Wu, Z., Kang, S., Meng, H., and Cai, L. (2015). A deep recurrent approach for acoustic-to-articulatory inversion. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2015-Augus, pages 4450–4454. IEEE.

Livescu, K., Bezman, A., Borges, N., Yung, L., ??etin, z., Frankel, J., King, S., Magimai-Doss, M., Chi, X., and Lavoie, L. (2007). Manual transcription of conversational speech at the articulatory feature level. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 4, pages IV–953–IV–956. IEEE.

Maeda, S. (1990). Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech Production and Speech Modelling*, pages 131–149. Springer Netherlands, Dordrecht.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, 53(4):1070–1082.

Mitra, V. (2010). Articulatory Information For Robust Speech Recognition.

Mitra, V. and Franco, H. (2015). Time-frequency convolutional networks for robust speech recognition. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 317–323. IEEE.

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (2012). Recognizing articulatory gestures from speech for robust speech recognition. *The Journal of the Acoustical Society of America*, 131(3):2270–2287.

Mitra, V., Nam, H., Espy-Wilson, C. Y., Saltzman, E., and Goldstein, L. (2010a). Retrieving tract variables from acoustics: A comparison of different machine learning strategies. *IEEE Journal on Selected Topics in Signal Processing*, 4(6):1027–1045.

Mitra, V., Nam, H., Espy-Wilson, C. Y., Saltzman, E., and Goldstein, L. (2010b). Retrieving tract variables from acoustics: A comparison of different machine learning strategies. *IEEE Journal on Selected Topics in Signal Processing*, 4(6):1027–1045.

Mitra, V., Nam, H., Espy-Wilson, C. Y., Saltzman, E., and Goldstein, L. (2011). Speech inversion: Benefits of tract variables over pellet trajectories. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5188–5191. IEEE.

Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C., and Saltzman, E. (2014a). Articulatory features from deep neural networks and their role in speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.

Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C., Saltzman, E., and Tiede, M. (2017). Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition. *Speech Communication*, 89:103–112.

Mitra, V., Wang, W., Franco, H., Lei, Y., Bartels, C., and Graciarena, M. (2014b). Evaluating robust features on Deep Neural Networks for speech recognition in noisy and channel mismatched conditions. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 895–899.

Nagamine, T., Seltzer, M. L., and Mesgarani, N. (2015). Exploring how deep neural networks form phonemic categories. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1912–1916.

Nam, H., Goldstein, L., Saltzman, E., and Byrd, D. (2004). TADA: An enhanced, portable Task Dynamics model in MATLAB. *The Journal of the Acoustical Society of America*, 115(5):2430.

Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (2012). A procedure for estimating gestural scores from speech acoustics. *The Journal of the Acoustical Society of America*, 132(6):3980–3989.

Narayanan, S., Bresch, E., Ghosh, P., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A., Proctor, M., Ramanarayanan, V., and Zhu, Y. (2011). A multimodal real-time MRI articulatory corpus for speech research. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 837–840. International Speech Communication Association.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. (2004). An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America*, 115(4):1771–1776.

Neiberg, D., Ananthakrishnan, G., and Engwall, O. (2008). The acoustic to articulation mapping: Non-linear or non-unique? In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1485–1488.

Oppenheim, A. V. (1969). Speech Analysis-Synthesis System Based on Homomorphic Filtering. *The Journal of the Acoustical Society of America*, 45(2):458–465.

Ouni, S. and Laprie, Y. (2005). Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 118(1):444–460.

Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zacks, J., and Levy, S. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *The Journal of the Acoustical Society of America*, 92(2 Pt 1):688–700.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 1–4.

Preston, J. L., McCabe, P., Rivera-Campos, A., Whittle, J. L., Landry, E., and Maas, E. (2014). Ultrasound visual feedback treatment and practice variability for residual speech sound errors. *Journal of Speech, Language, and Hearing Research*, 57(6):2102–2115.

Qin, C. and Carreira-Perpiñán, M. Á. (2007). An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. *Interspeech*, pages 74–77.

Qin, C. and Carreira-Perpiñán, M. A. (2010). Estimating missing data sequences in X-ray microbeam recordings. In *INTERSPEECH*, pages 1592–1595.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rabiner, L. R., Juang, B. ., Levinson, S. E., and Sondhi, M. M. (1985). Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities. *AT&T Technical Journal*, 64(6):1211–1234.

Rahim, M., Keijn, W., Schroeter, J., and Goodyear, C. (1991). Acoustic to articulatory parameter mapping using an assembly of neural networks. In *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 485–488 vol.1. IEEE.

Ramanarayanan, V., Goldstein, L., and Narayanan, S. S. (2013). Spatio-temporal articulatory movement primitives during speech production: Extraction, interpretation, and validation. *The Journal of the Acoustical Society of America*, 134(2):1378–1394.

Ramanarayanan, V., Van Segbroeck, M., and Narayanan, S. S. (2015). Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. *Computer Speech & Language*.

Reddy, S. (2014). New Tool for Children With Speech Errors; Making The 'R' Sound Is One of the Most Common Problems.

Richmond, K., King, S., and Taylor, P. (2003). Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech and Language*, 17(2-3):153–172.

Rothauser, E., Chapman, W., and Guttman, N. (1969). IEEE Recommnded Pratice for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246.

Sadjadi, S. O., Slaney, M., and Heck, L. (2013). MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research. *IEEE Speech and Language Processing Technical Committee Newsletter*, pages 1–4.

Saltzman, E. L. and Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology*, 1(4):333–382.

Sangwan, A. and Hansen, J. H. (2012). Automatic analysis of Mandarin accented English using phonological features. *Speech Communication*, 54(1):40–54.

Saon, G., Kuo, H. K. J., Rennie, S., and Picheny, M. (2015). The IBM 2015 English conversational telephone speech recognition system. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015-Janua:3141–3144.

Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., and Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31(1):26–35.

Schroeter, J. and Sondhi, M. M. (1994). Techniques For Estimating Vocal-Tract Shapes from the Speech Signal. *IEEE Transactions on Speech and Audio Processing*, 2(1):133–150.

Shinozaki, T. and Furui, S. (2003). An Assessment of Automatic Recognition Techniques for Spontaneous Speech in Comparison with Human Performance. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 95–98.

Shirai, K. and Kobayashi, T. (1986). Estimating articulatory motion from speech wave. *Speech Communication*, 5(2):159–170.

Sim, K. C. (2016). On constructing and analysing an interpretable brain model for the DNN based on hidden activity patterns. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*, pages 22–29. IEEE.

Sivaraman, G., Espy-Wilson, C., Mitra, V., Nam, H., and Saltzman, E. (2014). Analysis of acoustic to articulatory speech inversion for natural speech. In *The Journal of the Acoustical Society of America*, volume 136, pages 2082–2082. Acoustical Society of America.

Sivaraman, G., Espy-Wilson, C., and Wieling, M. (2017). Analysis of Acoustic-to-Articulatory Speech Inversion Across Different Accents and Languages. In *Interspeech 2017*, pages 974–978, Stockholm. ISCA.

Sivaraman, G., Mitra, V., Nam, H., Saltzman, E., and Espy-Wilson, C. (2015a). Augmenting acoustic phonetics with articulatory features for phone recognition. *Journal of the Acoustical Society of America*, 137(4):2302–2302.

Sivaraman, G., Mitra, V., Nam, H., Tiede, M., and Espy-Wilson, C. (2016). Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 08-12-Sept, pages 455–459.

Sivaraman, G., Mitra, V., Tiede, M., Saltzman, E., Goldstein, L., and Espy-Wilson, C. (2015b). Analysis of coarticulated speech using estimated articulatory trajectories. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015-Janua:369–373.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17(1):3–45.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4):1872.

Stevens, K. N. and Blumstein, S. E. (1981). The Search for Invariant Acoustic Correlates of Phonetic Features. In *Perspectives on the Study of Speech*, pages 1–38. Lawrence Erlbaum Associates Inc.

Tiede, M. K., Boyce, S. E., Holland, C. K., and Choe, K. A. (2004). A new taxonomy of American English /r/ using MRI and ultrasound. *The Journal of the Acoustical Society of America*, 115(5):2633–2634.

Tiede, M. K., Perkell, J., Zandipour, M., and Matthies, M. (2001). Gestural timing effects in the ‘‘perfect memory" sequence observed under three rates by electromagnetometry. *The Journal of the Acoustical Society of America*, 110(5):2657.

Toda, T., Black, A., and Tokuda, K. (2004). Acoustic-to-articulatory inversion mapping with gaussian mixture model. In *ICSLP*, pages 1129–1132, Jeju Island, Korea.

Twist, A., Baker, A., Mielke, J., and Archangeli, D. (2007). Are "Covert" /r/ Allophones Really Indistinguishable? *University of Pennsylvania Working Papers in Linguistics*, 13(2):207–216.

Uria, B., Renals, S., and Richmond, K. (2011). A Deep Neural Network for Acoustic-Articulatory Speech Inversion. *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, pages 1–9.

Weinberger, H. S. (2010). The Speech Accent Archive.

Westbury, J. R. (1994). Speech Production Database User ' S Handbook. *IEEE Personal Communications - IEEE Pers. Commun.*, 0(June).

Westbury, J. R., Hashi, M., and J. Lindstrom, M. (1998). Differences among speakers in lingual articulation for American English /r/. *Speech Communication*, 26(3):203–226.

Wieling, M., Veenstra, P., Adank, P., Weber, A., and Tiede, M. (2015). Comparing L1 and L2 speakers using articulography. In *Proceedings of ICPhS 2015*.

Wrench, A. and Richmond, K. (2000). Continuous speech recognition using articulatory data. In *Proc. ICSLP*, pages 145–148.

Wrench, A. A. (2000). A Multichannel Articulatory Database and its Application for Automatic Speech Recognition. In *In Proceedings 5 th Seminar of Speech Production*, pages 305–308.

Young, S. J., Evenmann, G., Gales, M. J. F., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2009). *The HTK Book, Version 3.4*. Cambridge university engineering department, 3 edition.

Yuan, J. and Liberman, M. Y. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics '08*, pages 5687–5790.

Zhou, X., Espy-Wilson, C. Y., Boyce, S., Tiede, M., Holland, C., and Choe, A. (2008). A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. *The Journal of the Acoustical Society of America*, 123(6):4466–81.

Zlokarnik, I. (1995). Adding articulatory features to acoustic features for automatic speech recognition. *The Journal of the Acoustical Society of America*, 97(5):3246.

Zue, V. and Lamel, L. (1986). An expert spectrogram reader: A knowledge-based approach to speech recognition. In *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 1197–1200. Institute of Electrical and Electronics Engineers.