ABSTRACT

Title of Dissertation: ARTICULATORY INFORMATION FOR ROBUST SPEECH RECOGNITION.

Vikramjit Mitra, Doctor of Philosophy, 2010

Dissertation directed by: Dr. Carol Y. Espy-Wilson
Department of Electrical and Computer Engineering

Current Automatic Speech Recognition (ASR) systems fail to perform nearly as good as human speech recognition performance due to their lack of robustness against speech variability and noise contamination. The goal of this dissertation is to investigate these critical robustness issues, put forth different ways to address them and finally present an ASR architecture based upon these robustness criteria.

Acoustic variations adversely affect the performance of current phone-based ASR systems, in which speech is modeled as ‘beads-on-a-string’, where the beads are the individual phone units. While phone units are distinctive in cognitive domain, they are varying in the physical domain and their variation occurs due to a combination of factors including speech style, speaking rate etc.; a phenomenon commonly known as ‘coarticulation’. Traditional ASR systems address such coarticulatory variations by using contextualized phone-units such as triphones. Articulatory phonology accounts for coarticulatory variations by modeling speech as

a constellation of constricting actions known as articulatory gestures. In such a framework, speech variations such as coarticulation and lenition are accounted for by gestural overlap in time and gestural reduction in space. To realize a gesture-based ASR system, articulatory gestures have to be inferred from the acoustic signal. At the initial stage of this research an initial study was performed using synthetically generated speech to obtain a proof-of-concept that articulatory gestures can indeed be recognized from the speech signal. It was observed that having vocal tract constriction trajectories (TVs) as intermediate representation facilitated the gesture recognition task from the speech signal.

Presently no natural speech database contains articulatory gesture annotation; hence an automated iterative time-warping architecture is proposed that can annotate any natural speech database with articulatory gestures and TVs. Two natural speech databases: X-ray microbeam and Aurora-2 were annotated, where the former was used to train a TV-estimator and the latter was used to train a Dynamic Bayesian Network (DBN) based ASR architecture. The DBN architecture used two sets of observation: (a) acoustic features in the form of mel-frequency cepstral coefficients (MFCCs) and (b) TVs (estimated from the acoustic speech signal). In this setup the articulatory gestures were modeled as hidden random variables, hence eliminating the necessity for explicit gesture recognition. Word recognition results using the DBN architecture indicate that articulatory representations not only can help to account for coarticulatory variations but can also significantly improve the noise robustness of ASR system.

IMPROVING ROBUSTNESS OF SPEECH RECOGNITION SYSTEMS

By

Vikramjit Mitra

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
2010

Advisory Committee:
Professor Carol Y. Espy-Wilson, Chair/Advisor
Professor Rama Chellappa
Professor Jonathan Z. Simon
Professor Mark Hasegawa-Johnson
Dr. Hosung Nam
Professor William J. Idsardi

Acknowledgements

Though my name appears as the sole author of this dissertation, I could never have performed this research without the support, guidance and efforts of a lot of people. I owe my sincere gratitude to all of them who have made this dissertation possible and have made my graduate experience a cherish-able one.

Words fall short to describe my gratefulness to my mentor, advisor and coach Dr. Carol Y. Espy-Wilson, who has been an unwavering source of intellectual and spiritual support throughout this long journey. Carol believed in me and my ability to perform fundamental research from day one. She introduced me to a very challenging and intriguing research problem and has been an extremely dedicated coach who knew how to motivate and energize me. She never lost confidence in me, even during my worst days.

My dissertation includes concepts borrowed from linguistics, a subject I barely had any prior exposure. This dissertation would never have been possible without the help from Dr. Hosung Nam, who like a big brother has been my constant source of inspiration and motivation. Most of the ideas presented in this dissertation are the outcomes of our long discussions through Skype. Hosung's positive attitude to life and research problems is unique. Whenever I felt down Hosung was there to bring me up on my feet. Whenever I needed him, he was always available for discussions setting aside his own work and patiently listening to my ideas, sharing his own two cents when necessary.

I am deeply indebted to Dr. Elliot Saltzman, Dr. Louis Goldstein and Dr. Mark Hasegawa-Johnson. Elliot and Louis had been instrumental throughout this work, being the main brains behind the Task Dynamic model and Articulatory Phonology; they have helped me to steer through the complex theoretical/analytical problems and helped to strengthen this inter-disciplinary study. Of particular note is the role of Dr. Mark Hasegawa-Johnson, who shared with me many of his ideas and opinions without any hesitation. No matter if it's a conference or a workshop, wherever I asked for his time to listen about my work he gladly accommodated me in his schedule.

In addition to Carol, Hosung and Mark, I would also like to sincerely thank the rest of my thesis committee members: Dr. Rama Chellappa, Dr. Jonathan Simon and Dr. William Idsardi for their patience, helpful comments, suggestions and encouragement. My gratitude also goes to Mark Tiede of Haskins Laboratories for sharing many of his suggestions, Dr. Karen Livescu and Arthur Kantor for their insightful discussions on Dynamic Bayesian Network, Yücel Özbek for his contribution on realizing the Kalman smoother and Xiaodan Zhuang for his several insightful discussions. Thanks to our collaborators Dr. Abeer Alwan, Jonas Börgström, Dr. Jenifer Cole and Dr. Mary Harper.

I am deeply indebted to all of my lab members and fellow graduate students Vladimir Ivanov, Xinhui Zhou, Srikanth Vishnubhotla, Daniel Garcia-Romero, Tarun Pruthi, Vijay Mahadevan, Jing-Ting Zhou and many others for their discussions, insightful comments and suggestions. Their friendship and warmth have given me tons of fun-filled memories that I can treasure for the rest of my life. My sincere thanks also go to all the faculty and staff members and help-desk personnel's of University of Maryland, who with their sincerity, diligence and collaboration have created a nurturing and fertile ground for fundamental research.

Of particular note is the role of my family, who were super caring, motivating and patient throughout the course of my graduate studies. I am profoundly indebted to my dear wife and friend Satarupa for her love, encouragement, dedicated support and patience. She had been a powerhouse of moral support, always cheering me up, sharing her words of encouragement and lending her patient ears to listen about my research and its mundane events. I am thankful to my son Ruhan for bringing unbounded joy in my life and giving me more fuel to work harder toward the end of this dissertation. I am very grateful to my father Mr. Subhajit Mitra and mother Ms. Jyotirmoyee Mitra for their firm belief in me and properly

educating me that made me what I am today. I am indebted to my father-in-law Mr. Asim K. Banerjee and mother-in-law Ms. Uma Banerjee for their love, support, motivation and unbounded confidence.

Finally I would like to thank the National Science Foundation (NSF) for supporting majority of this work through their grant (#IIS0703859), the Electrical and Computer Engineering Department and Graduate school of University of Maryland, College Park for supporting me with their Distinguished Dissertation Fellowship and Wylie Dissertation Fellowship that helped to financially cover the final days of my research.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Automatic Speech Recognition (ASR) is a critical component in applications requiring Human-Computer interaction such as automated telephone banking, hands-free cellular phone operation, voice controlled navigation systems, speech-to-text systems etc. To make such applications suitable for daily use, the ASR system should match human performance in a similar environment. Unfortunately the inherent variability in spontaneous speech as well as degradation of speech due to ambient noise, severely limits the capability of ASR systems as compared to human performance. The study reported in this dissertation aims to improve ASR robustness against speech variability and noise contamination.

One of the earlier studies that compared the performance of human speech recognition (HSR) and automatic speech recognition (ASR) was done by Van Leeuwen *et al.* (1995). They used eighty sentences from the Wall Street Journal database to compare their performances and reported a total word error rate (WER) of 2.6% for HSR as compared to 12.6% for ASRs. They noted that the ASR systems had greater difficulty with sentences having higher perplexity. Later, Lippman (1997) performed a similar study and showed that for word recognition experiments HSR performance was always superior to ASR performance as shown in Figure 1.1(a). Note, the ASR result in Figure 1.1(a) is from a recent study (Dharanipragada *et al.*, 2007) and the Lippman's (1997) actual work showed an even greater performance difference between HSR and ASR. More recently, Shinozaki & Furui (2003) compared HSR performance with that of the state-of-the-art Hidden Markov Model (HMM) based ASR, using a corpus of spontaneous Japanese speech. They have shown that the recognition error rates from HSR are almost half as those from the ASR system. They stated that this difference 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".

Several studies have also been performed to compare HSR and ASR capability in background noise. It was observed (Varga & Steeneken, 1993) that the HSR error rate on a digit recognition task was less than 1% in quiet and also at a signal-to-noise ratio (SNR) of 0dB. Another study (Pols, 1982) showed that HSR error rate was less than 1% at quiet and at an SNR as low as -3dB. For noisy speech, Varga & Steeneken (1993) showed that the least ASR error rate was about 2% in quiet condition and the error rates increased to almost 100% in noisy scenarios. This result was obtained when there was no noise adaptation of the HMM-based back-end. However, with noise adaptation, ASR error rate was reduced to about 40%. Cooke *et al.* (2006) and Barker & Cooke (2007) studied the performance of HSR as opposed to ASR systems, where the speech signals were corrupted with speech-shaped noise at different SNR levels. The obtained results are shown in Figure 1.1(b), where the HSR and ASR performance were very close in clean condition, but the ASR accuracy falls drastically as the SNR level is reduced. This performance difference in noisy conditions clearly shows that ASR systems are still far below human speech perception capabilities.

The above results have inspired a new direction in the field of speech recognition research which deals with incorporating robustness into existing systems and designing a new robust ASR architecture altogether. The comparison of HSR and ASR suggests that a robust ASR system should incorporate linguistic and speech processing factors that govern acoustic variations in speech, and also should consider the physiological speech production as well as speech perception model to distinguish and understand the dynamic variations in speech, both in clean and noisy scenarios.

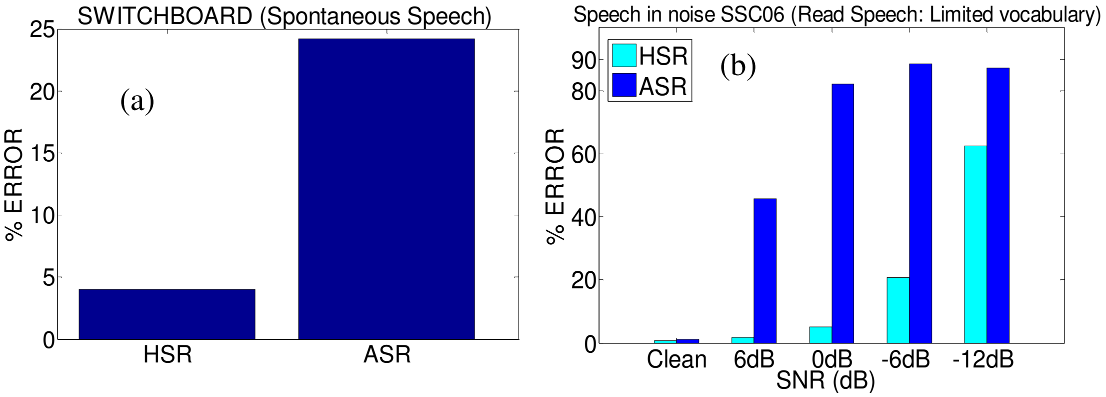


Figure 1.1 Comparison of WER of HSR and ASR for (a) spontaneous speech dataset (HSR result taken from [Lippman, 1997] and ASR result taken from [Dharanipragada *et al.*, 2007]) (b) read-speech at different signal-to-noise ratios (SNR) (Cooke *et al.*, 2006)

1.1What is meant by robustness in speech recognition systems?

Robustness in speech recognition refers to the need to maintain reasonable recognition accuracy despite acoustic and/or articulatory and/or phonetic characteristic mismatch between the training and testing speech samples. Human speech, even for a single speaker, varies according to emotion, style (carefully-articulated speech vs. more casual speech), speaking rate, dialect and prosodic context. This variability is the reason why even speaker-dependent ASR systems show appreciable degradation in performance when training and testing conditions are different. Variability can be even more pronounced when we factor in differences across speakers.

The other obstacle to robustness in ASR systems is noise corruption, which may be due to additive or convolutive noise arising from the environment, channel-interference or the encoding-decoding process. Different approaches have been explored for dealing with noise. One such approach is to enhance the speech signal by suppressing the noise while retaining the speech content with minimal distortion. Such a technique is used in the front-end of an ASR system prior to estimating the features as shown in Figure 1.2.

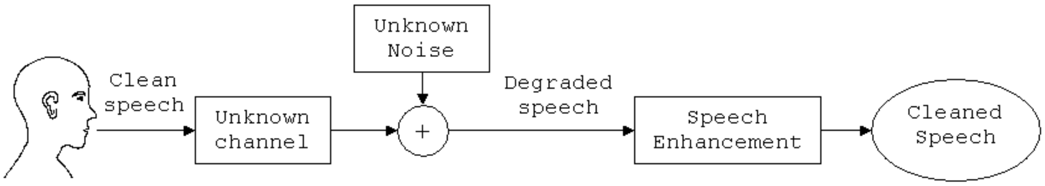


Figure 1.2 Sources that degrade speech recognition accuracy, along with speech enhancement that enhances the degraded speech to improve speech recognition robustness

1.2 How to incorporate robustness into speech recognition systems?

Figure 1.3 outlines the ASR system envisioned in this dissertation, which uses speech articulatory information in the form of articulatory trajectories and gestures to incorporate robustness into the ASR system. The front-end processing encodes the acoustic speech signal into acoustic features and performs operations such as mean and variance normalization, contextualization, etc. The speech inversion block transforms the acoustic features into estimated articulatory trajectories, which in turn are used along with the acoustic features in a gesture-based ASR system to perform word recognition.

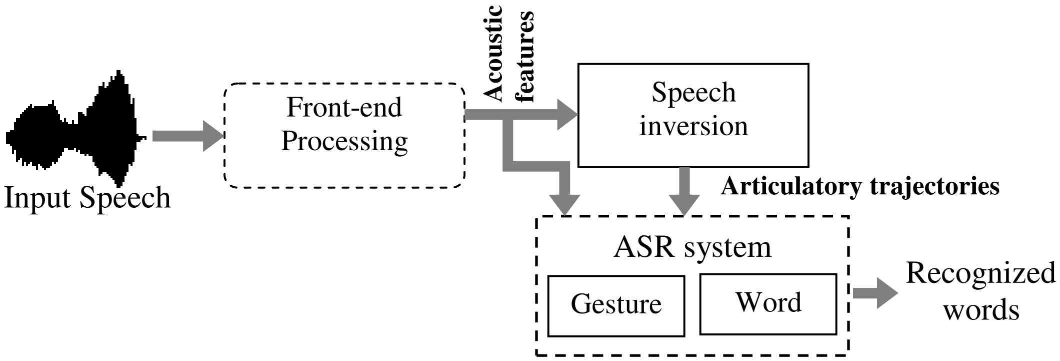


Figure 1.3 Architecture of Gesture based ASR system

In conversational speech, a high degree of acoustic variation in a given phonetic unit is typically observed across different prosodic and segmental contexts; a major part of which arises from contextual variation commonly known as coarticulation. Phone-based ASR systems represent speech as a sequence of non-overlapping phone units (Ostendorf, 1999) and contextual variations induced by coarticulation (Ohman, 1966) are typically encoded by unit combinations (e.g., tri- or quin-phone). These tri- or quin-phone based models often suffer from data sparsity (Sun & Deng, 2002). It has been observed (Manuel & Krakow, 1984; Manuel, 1990) that coarticulation affects the basic contrasting distinctive features between phones. Hence, an ASR system using phone-based acoustic models may be expected to perform poorly when faced with coarticulatory effects. Moreover triphone-based models limit the contextual influence to only the immediately close neighbors and, as a result are limited in the degree of coarticulation that they can capture (Jurafsky *et al*., 2001). For example, in casual productions of the word ‘strewn’, anticipatory rounding throughout the /str/ sequence can occur due to the vowel /u/. That is, coarticulatory effects can reach beyond adjacent phonemes and, hence, such effects cannot be sufficiently modeled by traditional tri-phone inventories.

In this study we propose that coarticulatory effects can be addressed by using an overlapping articulatory feature (or gesture) based ASR system. Articulatory phonology proposes the vocal tract constriction gestures of discrete speech organs (lips, tongue tip, tongue body, velum and glottis) as invariant action units that define the initiation and termination of a target driven articulatory constriction within the vocal tract. Articulatory phonology argues that human speech can be decomposed into a constellation of such constriction gestures (Browman & Goldstein, 1989, 1992), which can temporally overlap with one another. In articulatory phonology, gestures are defined in terms of the eight vocal tract constriction variables shown in Table 1.1 that are defined at five distinct constriction organs as shown in Figure 1.4. The tract variable time functions or the vocal tract constriction trajectories (abbreviated as TVs here) 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. For example the tract variable GLO and VEL are abstract measures that specify whether the glottis and velum are open/close, hence distinguishing for unvoiced/voiced and nasal/oral sounds. The TTCD and TBCD define the degree of constriction for tongue-tip and tongue-body and are measured in millimeters representing the aperture created for such constriction. TBCL and TTCL specify the location of the tongue-tip and tongue-body with respect to a given reference (F in Figure 1.4) and are measured in degrees. LP and LA define the protrusion of the lip and the aperture created by the Lip, and both are measured in millimeters. Gestures are defined for each tract variable. The tract variable time functions or trajectories (abbreviated as TVs here) are time-varying physical realizations of gestural constellations at the distinct vocal tract sites for a given utterance.

Figure 1.5 shows the gestural activations and TVs for the utterance “miss you” obtained from Haskins laboratories speech production model (aka TADA, Nam *et al*, 2004, see chapter 3 for details). A gestural activation is a binary variable that defines whether a gesture is active or not at a given time instant. In Figure 1.5 the gestural scores are shown as colored blocks, whereas the corresponding TVs are shown as continuous time functions.

Table 1.1 Constriction organs, vocal tract variables corresponding to the articulatory gestures

Constriction organ	Vocal tract variables
Lip	Lip Aperture (LA)
	Lip Protrusion (LP)
Tongue Tip	Tongue tip constriction degree (TTCD)
	Tongue tip constriction location (TTCL)
Tongue Body	Tongue body constriction degree (TBCD)
	Tongue body constriction location (TBCL)
Velum	Velum (VEL)
Glottis	Glottis (GLO)

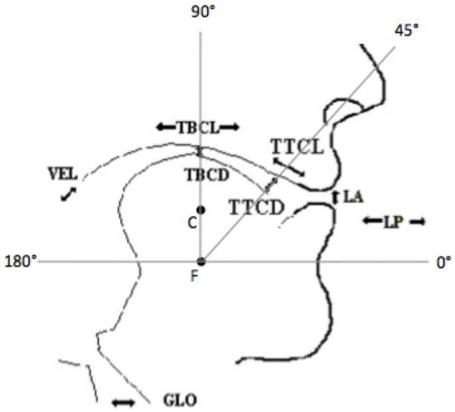


Figure 1.4. Vocal tract variables at 5 distinct constriction organs, tongue ball center (C), and floor (F) [Mermelstein, 1973; Browman & Goldstein, 1990]

Note in Figure 1.5, there are three TBCD gestures shown by the three rectangular blocks in the 5th pane from the top, whereas the VEL, TTCL, LA and GLO gestures shown in 3rd, 4th, 6th and 7th panes have only a single gesture. This is because the latter four gestures are responsible for only one constriction in the utterance 'miss you', LA and VEL for labial nasal /m/, GLO and TTCD for unvoiced tongue-tip critical constriction for consonant /s/, whereas TBCD is responsible for the vowels /IH/ in 'miss' and /Y/, /UW/ in 'you' that require narrow tongue body constrictions at mid-palatal, palatal and velic regions. The gestures can temporally overlap with one another within and across tract variables, which allows coarticulation and reduction to be effectively modeled.

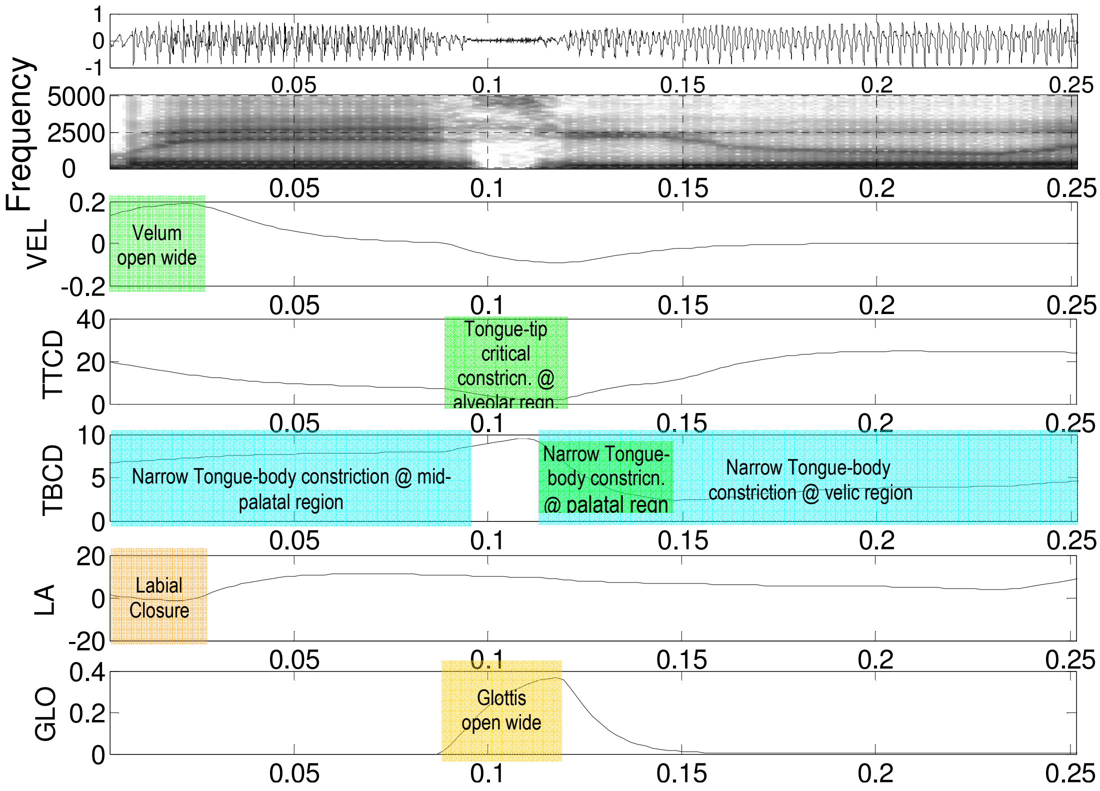


Figure 1.5 Gestural activations for the utterance “miss you”. The active gesture regions are marked by rectangular solid (colored) blocks. The smooth curves represent the corresponding tract variables (TVs)

Studying variability in speech using articulatory gestures also opens up the scope to better understand the relationship between acoustics and their corresponding articulation. Note that acoustic information relating to the production of speech sounds can sometimes be either hidden or, at the very least, quite subtle in the physical signal. For example, consider the waveforms, spectrograms and recorded articulatory information (obtained by placing transducers on the respective articulators) shown in Figure 1.6 for three pronunciations of “perfect memory” (Tiede *et al.*, 2001). These utterances were produced slowly with careful articulation, at a normal pace and at a fast pace by the same speaker. From the waveforms and spectrograms, we can see that the /t/ burst of the end of the word “perfect”, that is evident in the carefully articulated speech, is absent in the more fluent speech. In fact, whereas the /m/ of "memory" occludes the release of the /t/ in the normal-paced utterance, it occludes the release of the /t/ and the onset of the preceding /k/ in the fast spoken utterance. Due to the change in speaking rate, the degree of overlap between the gestures shown in the bottom three plots in Figure 1.6 are altered. As expected from the acoustics, the gesture for the lip closure of the /m/ is overlapped more with the tongue body gesture for the /k/ and the tongue tip gesture for the /t/ in the fast spoken utterance. However, the overall gestural pattern is the same. This result points to the invariance property of gestures. Given different variations of the same utterance, the degree of overlap between the gestures as well as the duration of each gesture might vary, but the overall gestural pattern will remain the same. Thus, while the acoustic information about the /k/and /t/ is not apparent in the fast spoken utterance (which is closest to what we expect in casual spontaneous speech), the articulatory information about these obstruents is obvious.

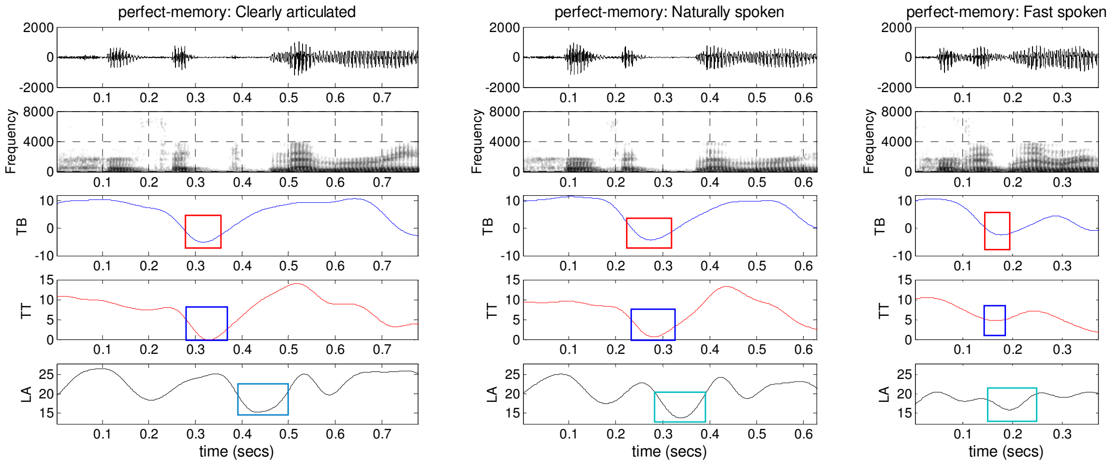


Figure 1.6 Waveforms, spectrograms, gestural activations and TVs for utterance 'perfect-memory' (Tiede *et al.*, 2001), when (a) clearly articulated (b) naturally spoken and (c) fast spoken. TB: vertical displacement for tongue-body transducer, TT: vertical displacement for tongue tip transducer and LA: lip aperture measured from upper and lower lip transducers

1.3 Objectives of this study

The goal of this study is to propose an ASR architecture inspired by articulatory phonology that models speech as overlapping articulatory gestures (Browman & Goldstein, 1989, 1992) and can potentially overcome the limitations of phone-based units in addressing variabilities in speech. To be able to use gestures as ASR units, they somehow need to be recognized from the speech signal. One of the primary goals of this research is to evaluate if articulatory gestures and their associated parameters can indeed be estimated from the acoustic speech signal. Some of the specific tasks performed in this research are stated below:

- In section 4.2, we present a model that recognizes speech articulatory gestures from speech (we name this model as the gesture-recognizer). We will explore different input conditions to obtain a better acoustic representation for articulatory gesture recognition. We will investigate the use of TVs as possible input and since we cannot expect to have prior knowledge about the TVs, we need to explore different ways to estimate TVs from a speech signal, motivating the task specified in the next bullet.
- In section 4.1, we explore different models (based on support vector regression, artificial neural networks, mixture density networks, etc.) to reliably estimate TVs from the speech signal (we name these models as the TV-estimators) and compare their performance to obtain the best model among them. Estimation of the TVs from the speech signal is a speech-inversion problem. Traditionally flesh-point articulatory information also known as pellet trajectories (Ryalls & Behrens, 2000; Westbury, 1994) has been used widely to perform speech inversion. In section 4.1.3.1 we will show that TVs are a better candidate for speech inversion than the pellet trajectories.
- In sections 4.3.2.1 and 4.3.2.2 we report the performance of the TV-estimator when the speech signal has been corrupted by noise.

- To analyze the suitability of TVs and gestures as a possible representation of speech for ASR systems, we will use the estimated TVs and the recognized gestures for performing ASR experiments with clean and noisy speech in section 4.3 and report their results.
- The experimental tasks specified above were all carried out using synthetic speech created in a laboratory setup. This approach is used because no natural speech database existed with gestural and TV annotations. Thus groundtruth TVs and gestural scores could only be obtained for synthetic speech. In chapter 5, we present an automated iterative time-warping algorithm that performs gestural score and TV annotation for any natural speech database. We annotate two databases: X-ray microbeam (XRMB [Westbury, 1994]) and Aurora-2 (Pearce & Hirsch, 2000) with gestural score and TV annotation and some analysis of the annotated data is presented.
- In chapter 6, we train the TV-estimator using the TV-annotated natural database and present the results. In section 6.1 we compare the speech inversion task on the XRMB data using TVs and pellet trajectories and show that TVs can be estimated more accurately than the pellet trajectories. Further, we show that the acoustic-to-articulatory mapping for the pellet trajectories are more non-unique than the TVs
- Finally in section 6.2 we propose and realize a gesture-based Dynamic Bayesian Network (DBN) architecture for an utterance recognition task, where the utterances consist of digit strings from the Aurora-2 database. The recognizer uses the estimated TVs and acoustic features as input, and performs utterance recognition on both clean and noisy speech.

Chapter 2: Background: Robust Approaches to ASR

Spontaneous speech typically has an abundance of variability, which poses a serious challenge to current ASR systems. Such variability has three major sources: (a) the speaker, introducing speaker specific variations such as dialectical - accentual - contextual variation, (b) the environment, introducing different background noises and distortions and (c) the recording device, which introduces channel variations and other signal distortions. In this dissertation we focus on (a) and (b). Usually contextual variability and noise-robustness are considered as two separate problems in ASR research. However while addressing speech variability in ASR systems, Kirchhoff (1999) and her colleagues (Kirchhoff *et al.*, 2002) showed that articulatory information can improve noise robustness while addressing speech variability due to coarticulation in speech. To account for variability of speech in ASR systems, Stevens (1960) suggested incorporating speech production knowledge into the ASR architecture. Incorporating speech production knowledge into ASR architecture is challenging because unlike acoustic information, speech production information (such as vocal tract shapes, articulatory configurations, their trajectories over time, etc.) is not explicitly available in usual ASR situations. Hence, the first logical step to introduce speech production knowledge into ASR is to estimate or recover such information from the acoustic signal. Two broad classes of articulatory information have been explored widely in literature: direct articulatory (recorded) trajectories and hypothesized articulatory features that are somehow deciphered from the acoustic signal. Landmark based systems were the offspring of both speech production and perception models, which targets to characterize linguistically important events. The different feature systems and approaches that aim to address speech variability and noise-corruption in ASR systems are detailed in this section.

2.1 Approaches that capture articulatory trajectories

The most direct way to capture articulatory information from speech is by placing transducers on different speech articulators and recording their movements while speech is generated. Such flesh-point articulatory trajectories had been exhaustively studied in the literature. Figure 2.1 shows the pellet placements for X-Ray MicroBeam (XRMB) dataset (Westbury, 1994). XRMB dataset contains recordings of articulator motions during speech production. The data is generated by tracking the motions of 2-3 mm diameter gold pellets glued to the tongue, jaw, lips, and soft palate. There are several other techniques to track articulatory events during speech, for example, Electromyography, Electropalatography (EPG), Electromagnetic Midsagittal Articulography (EMA) (Ryalls & Behrens, 2000) etc. Several studies have tried to estimate articulatory information from speech signal, a line of research commonly known as the 'acoustic-to-articulatory' inversion or simply speech inversion. Speech inversion or acoustic-to-articulatory inversion of speech has been widely researched in the last 35 years. One of the earliest and ubiquitously sited works in this area was by Atal *et al.* (1978), whose 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. At regular intervals they sampled the articulatory data to come up with 30,720 unique vocal tract configurations. For each configuration, they obtained the frequency, bandwidth and the amplitudes of the first five formants to define the corresponding acoustic space. Thus, given information in acoustic space, their approach would yield the corresponding vocal tract configuration.

Following the approach laid out Atal *et al.* (1978), Rahim *et al.* (1991, 1993) used an articulatory synthesis model to generate a database of articulatory-acoustic vector pairs. The acoustic data consisted of 18 Fast-Fourier Transform (FFT) derived cepstral coefficients, whereas the articulatory data is comprised of 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 the limitation to their approach was inadequate sampling strategy, as random sampling may select those physiologically-plausible articulatory configurations that may not be so common in typical speech. To address this fact Ouni & Laprie (1999) sampled on articulatory space such that the inversion mapping is piece-wise linearized. Their sampling strategy was based upon the assumption that the articulatory space is contained within a single hypercube, sampling more aggressively in regions where the inversion mapping is complex and less elsewhere. Shirai & Kobayashi (1986) proposed an analysis-by-synthesis approach, which they termed as Model Matching. In this approach real speech is analyzed to generate articulatory information and then the output is processed by a speech synthesizer such that it has minimal distance from the actual speech signal in the spectral domain. However, this approach severely suffered from computational overhead that led Kobayashi *et al.* (1985) to propose a two-hidden layer feed-forward MLP architecture that uses the same data as used by Shirai & Kobayashi, (1986), to predict the articulatory parameters. The approach in (Kobayashi *et al.*, 1985) was found to be 10 times faster than (Shirai & Kobayashi, 1986) and also offered better estimation accuracy. Regression techniques have been explored a number of times for speech inversion. Ladefoged *et al.* (1978) used linear regression to estimate the shape of the tongue in midsagittal plane, using the first three formant frequencies in constant vowel segments.

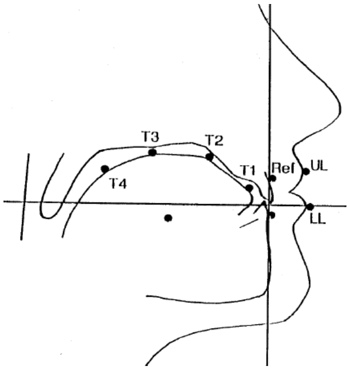


Figure 2.1 Pellet placement locations in XRMB dataset (Westbury, 1994)

Use of neural networks for speech-inversion has become much popular since the ubiquitously cited work by Papcun *et al.* (1992). They used MLPs to perform speech inversion to obtain three articulatory motions (y-coordinates for the lower lip, tongue tip and tongue dorsum) for six English stop consonants in XRMB. 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 the training time and maximizing the estimation performance. The network was trained using standard backpropagation algorithm. An important observation noted in their study was, trajectories of articulators considered critical for the production of a given consonant demonstrated higher correlation coefficients than for those who were considered non-critical to the production of that consonant. This observation was termed as the ‘Critical articulator phenomenon’. It should be noted here that this phenomenon may be better observed in TVs as opposed to the pellet-location based articulatory data as the critical articulation can be better defined by TVs modeling vocal-tract constriction than pellet traces. 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 further supported by Richmond (2001), who used Mixture

Density Networks (MDN) to obtain the articulator trajectories as conditional probability densities of the input acoustic parameters. He showed that the conditional probability density functions (*pdf*) of the critical articulators show very small variance as compared to the non-critical articulator trajectories. He also used Artificial Neural Networks (ANNs) to perform articulator estimation task and showed that the MDNs tackle the non-uniqueness issue of speech inversion problem more appropriately than the ANNs. Non-uniqueness is a critical issue related to acoustic-to-articulatory inversion of speech, which happens due to the fact that different vocal tract configurations can yield similar acoustic realizations, a most trivial example would be the difference between bunched and retroflex /r/ (Espy-Wilson *et al.*, 1999, 2000).

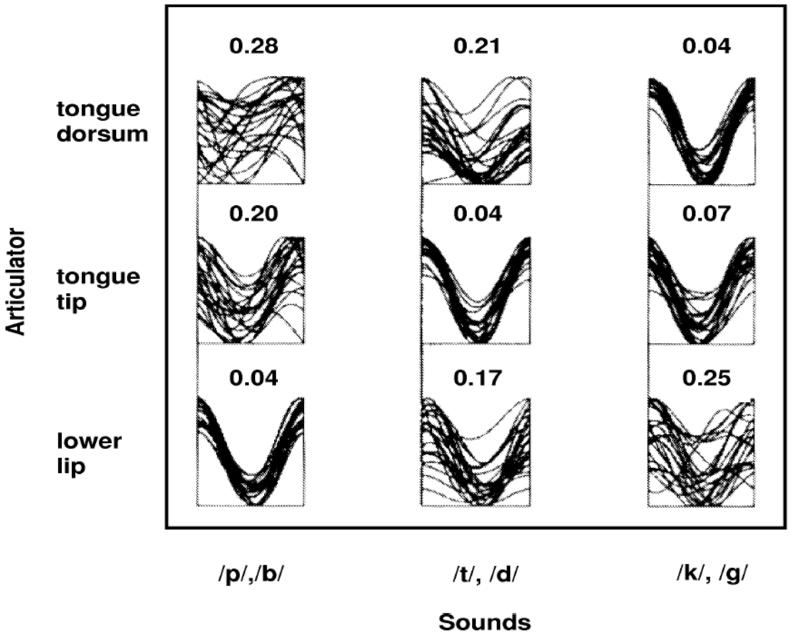


Figure 2.2 Trajectories (vertical movement) of critical and non-critical articulators. Three articulators: tongue dorsum, tongue tip and lower lip vertical trajectories are shown here for labial, coronal and velar sounds. Figure borrowed from Papcun *et al.* (1992)

The approach taken by Papcun *et al.* (1992) was further investigated by Zachs & Thomas (1994), however they used a different dataset than Papcun *et al.* (1992) and estimated eight articulatory channels, i.e., x and y coordinates for tongue tip, tongue body, tongue dorsum and lower lip. They used a new error function called "Correlation and Scaling Error" and showed a significant improvement in estimation accuracy using their error function as opposed to the default mean square error criteria in ANNs.

In a different study, Hogden *et al.* (1996) used a vector quantization to build a codebook of articulatory-acoustic parameter pairs. However their dataset was highly constrained containing 90 vowel transitions for a Swedish male subject in the context of two voiced velar oral stops. They built a lookup table of articulatory configurations and used the lookup table along with the codebook to estimate articulator positions given acoustic information. They reported an overall average Root Mean Square Error (RMSE) of approximately 2mm. A similar codebook approach was pursued by Okadome *et al.* (2000) who used a large dataset recorded from three Japanese male speakers. They also augmented the codebook search process by making use of phonemic information of an utterance. The average RMSE reported by their algorithm was around 1.6mm when they used phonemic information to perform the search process.

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

Frankel & King (2001) built a speech recognition system that uses a combination of acoustic and articulatory features as input. They estimated the articulatory trajectories using a recurrent ANN with 200ms input context window and 2 hidden layers. In their work they have used both the articulatory data obtained from direct measurements as well as from recurrent ANN estimation. They modeled the articulatory trajectories using linear dynamic models (LDM). These LDMs are segment specific, that is, each model describes the trajectory associated with each phone. Since the articulatory data used in their research lacked voicing information, they decided to use MFCC based feature set or exclusive features that captures zero crossing rate and voicing information (Frankel *et al.*, 2000). Phone models were trained using the expectation maximization (EM) rule. Phone classification was performed segment wise where the probability of the observations given the model parameters for each phone model was calculated. The phone classification accuracies from using estimated articulatory data did not show any improvement over the baseline MFCC based ASR system. However, using articulatory data from direct measurements in conjunction with MFCCs showed a significant improvement (4% in [Frankel *et al.*, 2000] and 9% in [Frankel & King 2001]) over the baseline system. They also observed the 'Critical articulator phenomenon' in their work and claimed that the knowledge about the critical and non-critical articulators may be necessary for an ASR system that relies upon articulator data. They claimed that recovering all the articulatory information perfectly over all the time should not be the goal of the speech-inversion module necessary for an ASR system; instead focus should be made to accurately estimate the critical articulators responsible for each segment of speech.

2.2 Phonetic features and their usage in ASR

Phonetic features are a set of descriptive parameters used in order to account for the phonological differences between phonetic units (Laver, 1994; Clements & Hume, 1995) of a language. The features may be based on articulatory movements, acoustic events or perceptual effects (Clark & Yallop, 1995). Ladefoged (1975) proposed a feature system where voicing is described as a glottal activity and has five values: glottal stop, laryngialized, voice, murmur and voiceless. Similarly Lindau (1978) proposed a feature system where the voicing or the glottal stricture is represented by different shapes of the glottis and are specified in terms of the values of glottal stop, creaky voice, voice, murmur and voiceless. A phonetic segment is defined as a discrete unit of speech that can be identified by a relatively constant phonetic feature(s). A given feature may be limited to a particular segment but may also be longer and are termed as the suprasegmental feature or may be shorter and are termed as the sub-segmental feature. Segments, usually phonological units of the language, such as vowels and consonants are of very short duration; typically a speech segment lasts approximately 30 to 300 msec. Utterances are built by linear sequence of such segments. Phonetic segments form a syllable, where syllables can also be defined in phonological terms. Different phonetic features have been proposed and different approaches introduced to obtain such phonetic features from speech signal. This section presents some of those approaches and presents their performance when applied to ASR.

2.2.1 Features capturing articulator location

The articulatory feature (AF) concept in literature parallels the "distinctive features" (DF) concept of phonological theory (Chomsky & Halle, 1968). Though there exists some strong similarity between the AFs and DFs, but there are some subtle differences too. DFs consist of both articulator-free and articulator-bound features (Stevens, 2002) defining phonological feature bundles that specify phonemic contrasts used in a language. On the contrary AFs define more physiologically motivated features based on speech production; hence they are fully articulator-bound features. Stevens (2002) proposed a lexical representation that is discrete in both how the words are represented as an ordered sequence of segments and how each of those segments is represented by a set of categories. Such discrete set of categories are motivated by acoustic studies of sounds produced from different manipulation of the vocal tract. For example, vowels typically are generated when the oral cavity is relatively open with glottal excitation. On the contrary consonants have a narrow constriction in the oral regions, the results are that the vowels usually have greater intensity than consonants and the low and mid frequency regions for consonants have weaker energy than the vowels. Reduced spectrum amplitude *a kin* to consonants can also be observed in case of glides (/w/ and /j/), where constriction is not created in the oral cavity but similar effects are produced due to the rise of the tongue dorsum producing a narrowing between the tongue and the palate, in case of /j/ or by rounding of lips in /w/. Stevens (2002) proposed that consonantal segments can be further sub-classified into three articulator-free features: continuant, sonorant and strident. For vowel, glide and consonant regions, articulator-bound features can be used, such as lips, tongue blade, tongue body etc., which determines which articulator is active for generating the sound at that specific region. Kirchhoff (1999) points out that some DFs such as *syllabic* and *consonantal* have the purpose of categorizing certain classes of speech sound but have no correlation or relationship to the articulatory motions. On the contrary the AFs are strong correlates of the articulatory space but have no direct functional dependency on acoustic space. ASRs that use DFs or acoustic-phonetic features, try to define high-level units, such as phones, syllables or words based on predefined set of such features for the language of interest.

Early attempts to exploit speech production knowledge in ASR systems were very limited in scope. From late 70s to early 90s of 20th century, most of the research efforts (Fujimura, 1986; Cole *et al.*, 1986; De Mori *et al.*, 1976; Lochschmidt, 1982) were focused on trying to decipher features from acoustic signal, which were largely acoustic-phonetic in nature. The CMU Hearsay-II system (Goldberg & Reddy, 1976) and the CSTR Alvey recognizer (Harrington, 1987) used acoustic-phonetic features. One of the earliest systems trying to incorporate AFs was proposed by Schmidbauer (1989), which was used to recognize speech in German language using 19 AFs that described the manner and place of articulation. The AFs were detected from preprocessed speech using a Bayesian classifier. The AF vectors were used as input to phonemic HMMs and an improvement of 4% was observed over the baseline for a small database. It was also observed that the AF features were robust against speaker variability and showed lesser variance of recognition accuracy for different phonemic classes as compared to the standard HMM-MFCC baseline. Self Organizing Neural Network (SONN) was used by Daalsgard (1992) and Steingrimsson *et al.* (1995) to detect acoustic-phonetic features for Danish and British English speech. The SONN output was used by a multivariate Gaussian mixture phone models for automatic label alignments. In a different study, Eide *et al.* (1993) used 14 acoustic-phonetic features for phonetic broad class classification and keyword spotting in American English speech. The features used in his research had both phonetic representation and articulatory interpretation. Using their feature set, they reported a classification accuracy of 70% for phoneme classification on TIMIT database. They showed significant improvement in performance when the baseline MFCC based system was combined with their feature set.

One of the earliest efforts to create a speech-production model inspired ASR system was by Deng (1992), where HMM states generated *trended*-sequence of observations, where the observations were piece-wise smooth/continuous. Deng *et al.* (1991, 1994[a, b]) and Erler & Deng (1993) performed an exhaustive study on articulatory feature based system, where they used 18 multi-valued features to describe place of articulation, vertical and horizontal tongue body movement and voice information. In their system they modeled the speech signal as rule-based combination of articulatory features where the features at transitional regions were allowed to assume any intermediate target value between the preceding and succeeding articulatory target values. They modeled each individual articulatory vector as HMM states and trained a single ergodic HMM, whose transition and emissions were trained using all possible vectors. They reported an average improvement of 26% over the conventional phone based HMM architecture for speaker independent classification task. Phone recognition for TIMIT dataset showed a relative improvement of at least 9% over the baseline system. For speaker-independent word recognition using a medium sized corpus, they reported a relative improvement of 2.5% over single-component Gaussian mixture phone recognizer.

A phonetic-feature classification architecture was presented by Windheuser *et al.* (1994), where 18 features were detected using a time-delay neural network. The outputs were used to obtain phoneme probabilities for ALPH English spelling database. Hybrid ANN/HMM architecture was proposed by Elenius *et al.* (1991, 1992) for phoneme recognition; where they compared spectral representations against articulatory features. For speaker independent phoneme recognition they reported that the articulatory feature based classifier performed better than the spectral feature based classifier; however for speaker dependent task the opposite was true.

King & Taylor (2000) used ANNs to recognize and generate articulatory features for the TIMIT database. They explored three different feature systems: binary features proposed by Chomsky & Halle (1968) based on *Sound Pattern of English* (SPE), traditional phonetic features defining manner and place categories, and features proposed by Harris (1994) that are based on Government Phonology (GP). The recognition rate of the three feature systems showed similar performance. In a different study Kirchhoff *et al.* (1999, 2002) used a set of heuristically defined AFs to enhance the performance of phone based systems. She showed that incorporating articulatory information in an ASR system helps to improve its robustness. The AFs used in her work, describes speech signal in terms of articulatory categories based on speech production models. The proposed AFs do not provide detailed numerical description of articulatory movements within the vocal tract during speech production; instead they represent abstract classes characterizing the most critical aspects of articulation in a highly quantized and canonical form (Kirchhoff, 1999). These AFs provide a representation level intermediate between the signal and the lexical units, for example: voiced/unvoiced, place and manner of articulation, lip-rounding etc. Acoustic signal was parameterized to acoustic features and a single hidden-layer MLP was used to derive the AFs given the acoustic features. She argued that the proposed AFs by itself or in combination with acoustic features will lead to increased recognition robustness against background noise. It was also demonstrated by Kirchhoff (1999) that the effectiveness of noise robustness of such a system increases with a decrease in the Signal-to-Noise ratio (SNR), which would be highly desirous from a robust ASR system. Her approach using articulatory AFs has shown success when used in conjunction with MFCCs in noisy conditions (Kirchhoff, 1999), based on this she inferred that AF and MFCC representation may be yielding partially complementary information and hence neither of them alone are providing better recognition accuracies than when both of them are used together.

ANNs have been extensively used in AF recognition from the speech signal. Wester *et al.* (2001) and Chang *et al.* (2005) proposed separate place classifiers for each manner class. Omar & Hasegawa-Johnson (2002) used a maximal mutual information approach to obtain a subset of acoustic features for the purpose of AF recognition. HMMs have also been researched widely for AF recognition. Metze & Waibel (2002) proposed context-dependent HMM phone models to generate an initial AF set, which were later replaced by a set of feature detectors that uses a likelihood combination at the phone or state level. In their research they showed a WER reduction from 13.4% to 11.6% on a Broadcast news database with a 40k dictionary. They also showed a reduction in WER from 23.5% to 21.9% for the Verbmobil task, which contains spontaneous speech.

Dynamic Bayesian Networks (DBN) has also been explored for the purpose of AF recognition. The major advantage of DBN is its capability to model explicitly the inter-dependencies between the AFs. Also a single DBN can perform both the task of AF recognition and word recognition, which further strengthens the claim for applicability of DBNs in AF based ASR system. One of the earlier works using DBN for the task of AF recognition was performed by Frankel *et al.* (2004). 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% and also noted a significant improvement in overall frame wise features simultaneously correct together from 47.2% to 57.8%. However tying AF features to phone level information overlooks the temporal asynchrony between the AFs. To address this issue an embedded training scheme was proposed by Wester *et al.* (2004), which was able to learn a set of asynchronous feature changes from data. Their system showed a slight increase in accuracy for a subset of the OGI number corpus (Cole *et al.*, 1995) over a similar model trained on phone-derived labels. Frankel & King (2005) proposed a hybrid ANN/DBN architecture, where the Gaussian Mixture Model (GMM) observations used by the DBNs are replaced by ANN posteriors. This hybrid ANN/DBN architecture combined the discriminative training power of ANN and the inter-feature dependency modeling capability of the DBN. The feature recognition accuracy reported in their paper for the OGI Number corpus was reported as 87.8%.

Livescu *et al.* (2007a) presented 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 for eight tiers of AFs. For transcription they began with phone alignments and used hybrid phone feature labeling to manually replace a canonical phone region with AFs. For the regions that were devoid of canonical phone information, they manually specified AFs based on information from Wavesurfer (2006). The resulting data consisted of 78 utterances drawn from SVitchboard (King *et al.*, 2005) which is a subset of the Switchboard corpus. Their work also shows inter-transcriber agreement and the degree to which they used the articulatory feature tiers. 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, Cetin *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. Also, 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. At the 2006 Johns Hopkins University Workshop, Livescu *et al.* (2007b) investigated the use of AFs for the observation and pronunciation models for ASR systems. They used the AF classifier outputs in two different ways (1) as observations in a hybrid HMM/ANN model and (2) as a part of the observation in a tandem approach. In this work they used both audio and visual cues for speech recognition and the models were implemented as DBNs. They used SVitchboard (King *et al.*, 2005) and the CUAVE audio-visual digits corpus to analyze their approach. They observed that the best ASR performance came from the tandem approach, where as the hybrid models though couldn't offer the best accuracy but required a very little training data. They predicted that hybrid model based approaches may hold promises for multi-lingual systems. Hasegawa-Johnson *et al.* (2007) exploited the asynchrony between phonemes and visemes to realize a DBN based speech recognition system. They noted that the apparent asynchrony between acoustic and visual modalities can be effectively modeled as the asynchrony between articulatory gestures corresponding to lips, tongue and glottis/velum. Their results show that combining visual cues with acoustic information can help reduce the WER at low SNR and the WER is found to further reduce if the asynchronies amongst gestures are exploited.

To address the issue of coarticulation modeling in speech recognition systems, Sun & Deng (1998) proposed an overlapping feature-based phonological model, which provides long-term contextual dependency. Influenced by the concept of gestural phonology (Browman & Goldstein, 1989, 1992) and autosegmental phonology (Goldsmith, 1990) they aimed to perform pronunciation or lexical modeling. Their framework is based on subphonemic, overlapping articulatory features where the rules governing the overlapping patterns are described by finite state automata. In such a framework, each state in the automaton corresponds to a bundle of features with specified relative timing information (Deng, 1997). They reported a word correct rate of 70.9% and word accuracy rate of 69.1% using bigram language model for the TIMIT dataset. They also proposed (Sun *et al.*, 2000 [a, b]) a data-driven approach to derive articulatory-feature based HMMs for ASR systems. They used University of Wisconsin's X-ray Microbeam database (Westbury, 1994) and created regression tree models for constructing HMMs. In their feature-based phonological model, patterns of overlapping features are converted to an HMM state transition network, where each state encodes a bundle of overlapping features and represents a unique articulatory configuration responsible for producing a particular speech acoustics. In their model asynchrony between the features are preserved. When adjacent features overlap with each other asynchronously in time, they generate new states which either symbolizes a transitional stage between two subsequent segments or an allophonic alteration due to contextual influence. They claimed that as their feature has long-time contextual dependency modeled appropriately in terms of bundle of overlapping features, hence should show improvements in ASR results over the phone-based models, as di- or tri-phone based models only incorporate short term or immediate phonemic contextual dependence. Their data-driven overlapping feature based system (Sun & Deng, 2002) showed an improvement in ASR performance for the TIMIT dataset, where they reported a phone correct rate of 74.7% and phone recognition accuracy of 72.95 as opposed to 73.99% and 70.86% from the conventional tri-phone system.

2.2.2 Landmark based feature detection

The Landmark based ASR models are inspired by the human speech production and perception mechanism. Landmark based ASR systems proposed by Stevens (2000b, 2002) use a feature based representation of the acoustic waveform and such a system helps to hypothesize locations of landmarks. The landmarks are identified as points in the speech signal corresponding to important events such as consonantal closures and release bursts. Some landmarks in an essence indicate articulator free features, such as continuants and sonorants. Based on the detected landmarks, various acoustic-phonetic cues, such as formant frequencies, hilbert envelop, duration of frication, spectral amplitudes, etc, are extracted around the landmark regions which are used to determine articulator-bound distinctive features, such as place of articulation, nasality etc. The hypothesized features are then compared against the feature based lexical entries corresponding to a word or a phone.

Several different implementations of landmark based systems exist but none of them has realized a full blown ASR system. Most of the research proposed in this field deals with some aspect of the landmark theory that is detecting the landmark regions, obtaining broad class information etc. Vowel landmarks were detected by Howitt (1999) using simple MLPs. Choi (1999) proposed a way to detect consonant voicing using knowledge based cues at manually-labeled landmarks. A landmark based ASR system has been proposed in Johns Hopkins summer workshop of 2004 (Hasegawa-Johnson *et al.*, 2005), which built three prototype ASR systems based on Support Vector Machines (SVM), Dynamic Bayesian Networks (DBN) and maximum entropy classification. They created a more feature-based representation of words as opposed to a phonetic one and compared their proposed models against the current state-of-the-art ASR model for conversational telephonic speech. Unfortunately, none of them were able to surpass the latter in terms of performance. They used an SVM based approach to detect both landmarks and the presence or absence of distinctive features. However they noted that their SVM based approach performed binary phone detection and classification with a very low error rate. They observed that a DBN based pronunciation model coupled with a SVM phonetic classifier was able to correctly label the underlying articulatory changes in the regions of pronunciation variation. They also noted that in their architecture it was possible to use a rescoring strategy that successfully chose salient landmark differences for alternate recognizer hypothesized words and performed landmark detection to obtain a better hypothesis.

Use of landmarks does not imply explicit use of speech production knowledge, but mostly reflects a hybridization between phone-based and articulatory feature based approach. The MIT-SUMMIT speech recognition system by Glass (2003) formalizes some of the landmark-based concepts proposed by Stevens (2002) in a probabilistic framework. In the SUMMIT system, potential phone boundary landmarks were located first and those were used by the phone-based dictionary to represent words. Different landmark detection algorithms (Chang & Glass, 1997; Glass, 1988] and acoustic cues (Halberstadt & Glass, 1998; Muzumdar, 1996] have been implemented in the SUMMIT system. SUMMIT operates either in the boundary based mode, where the phonetic boundary landmarks are explicitly modeled; or in a segment based mode, where the regions between the landmarks are modeled. Tang *et al.* (2003) proposed a two-stage feature based approach where they have used SUMMIT in a combined phone-feature setup for word recognition.

One of the first landmark systems that used SVMs for landmark detection was proposed by Juneja (2004) and Juneja & Espy-Wilson (2003 [a, b], 2008), where SVM discriminant scores were converted to likelihood estimates and a modified Viterbi scoring was done using a phonetic base-form dictionary, which was mapped to distinctive features. They named their system as the event-based system or the EBS. In their system, they hypothesized the speech recognition problem as a maximization of the joint posterior probabilities of a set of phonetic features and the corresponding acoustic landmarks (Juneja, 2004). SVM based binary classifiers recognizing manner features like syllabic, sonorant and continuant were used which performed the probabilistic detection of speech landmarks. The landmarks (Juneja, 2004) included stop bursts, vowel onsets, syllabic peaks, syllabic dips, fricative onsets and offsets, and sonorant consonant onsets and offsets. The SVM classifiers used knowledge based acoustic parameters (APs) which were acoustic-phonetic feature correlates. Their framework exploited two properties of the knowledge-based acoustic-phonetic feature cues: (1) sufficiency of the acoustic cues for a phonetic feature and (2) context-invariance of the acoustic-phonetic cues. They claimed that the probabilistic framework of their system makes it suitable for a practical recognition task and also enables the system to be compatible with a probabilistic language and pronunciation model. Their results claimed that their proposed system (Juneja, 2004; Juneja & Espy-Wilson, 2008) offered performance comparable to HMM-based systems for landmark detection as well as isolated word recognition.

2.3 Vocal Tract Resonances and Deep Architectures

Apart from features capturing articulatory motions, other sources of information such as vocal tract shapes and vocal tract resonances (VTR) has been used to capture the dynamics of natural speech. Deng *et al.* (1997) and Deng (1998) proposed a statistical paradigm for speech recognition where phonetic and phonological models are integrated with a stochastic model of speech incorporating the knowledge of speech production. In such an architecture the continuous and dynamic phonetic information of speech production (in the form of vocal tract constrictions and VTRs) is interfaced with a discrete feature based phonological process. It is claimed (Deng, 1998) that such integration helps to globally optimize the model parameters that accurately characterize the symbolic, dynamic and static components in speech production and also contribute in separating out the sources of speech variability at the acoustic level. Their work (Deng *et al.*, 1997) shows that synergizing speech production models with a probabilistic analysis-by-synthesis strategy may result in automatic speech recognition performance comparable to the human performance. Deng & Ma (2000) and Ma & Deng (2000) proposed a statistical hidden dynamic model to account for phonetic reduction in conversational speech, where the model represents the partially hidden VTRs and is defined as a constrained and simplified non-linear dynamical system. Their algorithm computes the likelihood of an observation utterance while optimizing the VTR dynamics that account for long term context-dependent or coarticulatory effects in spontaneous speech. In their work the hidden VTR dynamics are used as an intermediate representation for performing speech recognition, where much fewer model parameters had to be estimated as compared to tri-phone based HMM baseline recognizers. Using the Switchboard dataset they have shown reduction (Deng & Ma, 2000; Ma & Deng, 2000) in word error rates when compared with baseline HMM models. Togneri & Deng (2003) used the hidden-dynamic model to represent speech dynamics and explored EKF to perform joint parameter and state estimation of the model. Deng *et al.* (2004) proposed an efficient VTR tracking framework using adaptive Kalman filtering, and experiments on the Switchboard corpus demonstrated that their architecture accurately tracks VTRs for natural, fluent speech. In a recent study, Deng *et al.* (2006) showed that a structured hidden-trajectory speech model exploiting the dynamic structure in the VTR space can characterize the long-term contextual influence among phonetic units. The proposed hidden-trajectory model (Deng *et al.*, 2006) showed improvement in phonetic recognition performance on the TIMIT database for the four broad phone classes (sonorants, stops, fricatives and closures) when compared with the HMM baseline.

Deep Learning architectures (He & Deng, 2008) were introduced in ASR paradigm to address the limited capability of the HMM-based acoustic models for accounting variability in natural speech. The main drawback of HMM architectures are their first order Markov chain assumption and the conditional independence assumption. Deep Learning architectures have the capability to model streams of mutually interacting knowledge sources by representing them in multiple representation layers. A recent study by Mohamed *et al.* (2009) has proposed a Deep Belief Network (Hinton *et al.*, 2006) based acoustic model that can account for variability in speech stemming from the speech production process. A Deep

Belief Network is a probabilistic generative model consisting of multiple layers of stochastic latent variables (Mohamed *et al.*, 2009). Restricted Boltzmann machines (RBMs), owing to their efficient training procedure are used as the building block for Deep Belief Networks. Mohamed *et al.* (2009) performed a phone recognition task to the TIMIT corpus using MFCCs with delta (velocity) and delta-delta (acceleration) as the acoustic features and reported a phone error rate of 23%, compared to 25.6% obtained from Bayesian triphone HMM model reported by Ming & Smith (1998). They have also shown that their system offers the least phone error rate compared to some previously reported results. Another recent study by Schrauwen *et al.* (2009) proposed using a Temporal Reservoir Machines (TRM) which is a generative model based on directed graphs of RBMs. Their model uses a recurrent ANN to perform temporal integration of the input which is then fed to an RBM at each time step. They used the TRM to perform word recognition experiments on the TI46 dataset (subset of TIDIGITS corpus) and have used the Lyon passive ear model to parameterize the speech signal into 39 frequency bands. The least WER reported in their paper is 7%.

2.4 Noise Robust Approaches to Speech Recognition

Several approaches have been proposed to incorporate noise robustness into ASR systems, which can be broadly grouped into three categories: (1) the frontend based approach, (2) the backend based approach and (3) the missing feature theory.

Frontend based approaches usually aim to generate relatively contamination-free information for the backend classifier or model. Such approaches can be grouped into two sub-categories. First, the noisy speech signal is enhanced by reducing the noise contamination (e.g., spectral subtraction [Lockwood & Boudy, 1991], computational auditory scene analysis [Srinivasan & Wang, 2007], modified phase opponency model (MPO [Deshmukh *et al.*, 2007]), speech enhancement with auditory modeling using the ETSI system [Flynn & Jones, 2008], etc.), the enhanced signal is then parameterized and fed to the ASR system. Second, features effective for noise robustness are used to parameterize the speech signal before being fed to the ASR system (e.g., RASTAPLP [Hermansky & Morgan, 1994], Mean subtraction, Variance normalization and ARMA filtering (MVA) post-processing of cepstral features [Chen & Bilmes, 2007], cross-correlation features [Sullivan, 1996], variable frame rate analysis [You *et al.*, 2004], peak isolation [Strope & Alwan, 1997] and more recently the ETSI basic [2003] and the ETSI advanced [2007] frontends, etc.).

The backend based approach incorporates noise robustness into the backend of the ASR system, where the backend is typically a statistical model (usually a Hidden Markov Model (HMM)) for modeling different speech segments. The goal of the backend based systems is to reduce the mismatch between the training and the testing data. One such approach is to train the backend models using data that contain different types of noise at different levels (Kingsbury *et al.*, 2002). However a shortfall to such a system is the necessity of knowledge of all possible noise type at all possible contamination levels, which renders the training data immensely huge if not unrealizable. An alternative is to adapt the backend to the background noise. For instance, Parallel Model Combination (PMC [Gales & Young, 1996]) uses the noise characteristic and the relation between the clean and noisy speech signals to adapt the Gaussian mixture means and covariances of clean acoustic HMMs toward the true distributions of the noisy speech features. Usually such transformation is fairly accurate but computationally expensive because the model parameters need to be updated constantly for non-stationary noise. Maximum Likelihood Linear Regression (MLLR [Leggetter & Woodland, 1995]) performs model adaptation by rotating and shifting the Gaussian mixture means of clean HMMs using linear regression without using any prior knowledge of the background noise. Piecewise-Linear Transformation (PLT) was proposed by Zhang & Furui (2004) for a modified version of MLLR where different noise types are clustered based on their spectral characteristics and separate acoustic models are trained for each cluster at different Signal-to-Noise Ratios (SNR). During recognition, the best matched HMM is selected and adapted by MLLR.

The third approach is the missing feature theory (Cooke *et al.*, 2001; Barker *et al.*, 2000), which assumes that for noisy speech some spectro-temporal regions are usually so noisy that they can be treated as missing or unreliable. The missing feature approach tries to compute a time-frequency reliability mask to differentiate reliable regions from the unreliable ones where the mask can be binary (Cooke *et al.*, 2001) or real valued (Barker *et al.*, 2000). Once the mask is computed, the unreliable components are dealt with by two different approaches: (a) data imputation (Cooke *et al.*, 2001) where the unreliable components are re-estimated based on the reliable components and (b) marginalization (Cooke *et al.*, 2001) where only the reliable components are used by the backend for recognition. Bounded Marginalization (BM) was proposed in (Josifovski *et al.*, 1999) which generally outperforms "plain" marginalization. BM uses the knowledge that the unreliable data is bounded and the knowledge of such bounds is used to constrain the upper and lower bounds of the integral used for obtaining the likelihood of the incomplete data vector.

Use of articulatory information has also been found to improve noise robustness in ASR systems, though their actual use was motivated to account coarticulatory variation. Kirchhoff (1999) was the first to show that such information can help to improve noise-robustness of ASR systems as well. She showed that AFs in combination with MFCCs provided increased recognition robustness against the background noise (pink noise at four different SNRs). She concluded that the AFs and MFCCs may be yielding partially complementary information since neither alone provided better recognition accuracy than when both used together. In a different study, Richardson *et al.* (2003) proposed the Hidden Articulatory Markov Model (HAMM) that models the characteristics and constraints analogous to the human articulatory system. The HAMM is essentially an HMM where each state represents an articulatory configuration for each di-phone context, allowing asynchrony amongst the articulatory features. They reported that their articulatory ASR system demonstrated robustness to noise and stated that the articulatory information may have assisted the ASR system to be more attuned to speech-like information.

2.5 Speech Gestures as sub-word units

Variations in speech can be better modeled by using articulatory gestures that refer to spatiotemporal behavior of discrete constricting actions in the vocal tract (Browman & Goldstein, 1989, 1992). Articulatory phonology (Browman & Goldstein, 1989) views an utterance as a constellation of speech articulatory gestures, where the gestures may temporally overlap with one another and may get spatially reduced. Gestures are constriction (constriction-forming and releasing) action units produced by distinct constricting organs (lips, tongue tip, tongue body, velum and glottis) along the vocal tract.

Current ASR systems largely rely upon the contrastive features between the phonetic units to recognize one unit from another. Manuel & Krakow (1984) showed that the proximity of contrastive phonetic units affects coarticulation. Manuel (1990) examined vowel-to-vowel coarticulation across different languages and showed that it differs depending on how the languages divide the vowel space into contrastive units. It was observed that anticipatory coarticulation (when articulatory requirements of one phone are anticipated during the production of a preceding phone(s)) may produce contextually induced variability in the signal associated with the preceding phone(s). For example in a vowel-nasal sequence as in "*pan*", the velum typically begins (and may complete) its lowering movement associated with the nasal /n/, while the vocal tract is still open for the vowel /ae/ and well before the oral occlusion for the /n/ is achieved. These observations suggest that coarticulation results in spilling-over its effect to the neighboring phones. It is also observed (Manuel, 1990) that coarticulation affects the very primitives of contrast between phones;

hence an ASR system using mono-phone acoustic model may be expected to suffer adversely due to coarticulatory effects. To overcome the limitations of mono-phone acoustic models, bi-phone or tri-phone acoustic models have been proposed that considers a set of two or three neighboring phones to construct the acoustic model. However these di-phone or tri-phone-based ASR systems limit the contextual influence to only immediately close neighbors and require a significantly large training data to combinatorially generate all possible di-phone or tri-phone units. Such di-phone or tri-phone based models often suffer from data sparsity owing to the imbalance of available data for creating all possible di-phone or tri-phone models.

It has been observed that speakers generally limit coarticulation in a way that it does not destroy the distinctive attributes of gestures (Martinet, 1957, Manuel & Karkow, 1984; Manuel, 1990). These output constraints are found to be functionally dependent upon language-particular systems of phonetic contrast. It was also observed that the degree of anticipatory coarticulation (Manuel, 1990) varies from language to language and also by the proximity of contrastive phonetic units. In a study on coarticulatory stability in American English /r/, Boyce & Espy-Wilson (1997) observed the interaction between /r/ and surrounding segments and stated that the phonological and coarticulatory interaction between /r/ and its surrounding phones can be described as 'trajectory overlap' and 'sliding' of /r/ related characteristics to the neighboring regions which accounts for the articulatory plan for /r/.

Coarticulation is a property of action that can only occur when discrete actions are sequenced (Fowler, 2003), it has been described in a variety of ways: such as spreading of features from one segment to another or as assimilation. For example in case of '*strewn*', the coarticulatory effects of /u/ can cause some degree of anticipatory rounding throughout the /str/ sequence. This shows that coarticulatory effects can reach beyond adjacent phonemes and hence such effects are not covered by traditional tri-phone inventories. Fowler (2003) states that coarticulation can be tracked more transparently when articulatory activity is tracked, in such a case coarticulation is a temporal overlap of articulatory activity for neighboring consonants and vowels. In such an overlapping model, overlap can occur both in anticipatory (right-to-left) and carryover (left-to-right) direction. This phenomenon can be modeled by gestural overlap and is typically identified as coproduction. The span of such overlap can be segmentally extensive (Ohman, 1966; Recasens, 1984; Fowler & Brancazio, 2000) but may not be more than 250ms (Fowler & Saltzman, 1993). A consonantal duration can often be less than 100ms, which suggests that in consonantal context, coarticulatory effects can theoretically spill-over to more than a tri-phone context.

Chapter 3: Tools and Databases

In this study, speech variability is dealt with by modeling the speech signal as a bundle of overlapping articulatory gestures, where the degree and extent of overlap between the gestures are determined by those of coarticulatory effects. Speech gestures can be defined as constricting actions for distinct organs/constrictors along the vocal tract. The organs/ constrictors are the lips, tongue tip, tongue body, velum and the glottis. Each gesture is dynamically coordinated with a set of appropriate articulators. A word can be defined as a constellation of distinct gestures (gestural scores). For a given word's gestural score, the TAsk Dynamics Application model (TADA) developed at Haskins laboratories (Nam *et al.*, 2004) computes the inter-articulatory coordination and outputs the time functions of the vocal tract variables or TVs (both degree and location variables of the constrictors) and model articulator variables.

This dissertation aims to model coarticulation in terms of speech articulatory gestures. Unfortunately the spontaneous speech databases available for ASR do not come with any gestural specification; hence to obtain a proof of concept for our approach, TADA was used to generate a set of databases that contain synthetic speech along with their articulatory information in the form of articulatory gestures, TVs and pellet trajectories. These synthetic databases were used to perform a set of initial studies to ascertain whether articulatory gestures can be effectively recognized from the speech signal and the recognized gestures can further be a set of viable units for ASR. Finally, to confirm our observations made from our initial studies with synthetic speech, we performed similar experiments on natural speech, which requires a natural speech corpus with gestural and TV annotation. In order to annotate gestural scores and TVs for natural speech, we developed an iterative landmark-based time-warping procedure to time-warp synthetic speech onto a given natural speech. This technique is presented in section 5. The following subsection presents detail about the TADA model and the speech databases used in this dissertation.

3.1 The TAsk Dynamic and Applications Model

The TAsk Dynamic and Applications (TADA) model (Nam *et al.*, 2004) is Haskins laboratories articulatory speech production model that includes a task dynamic model and a vocal tract model. The task-dynamic model of speech production (Saltzman & Munhall, 1989; Nam *et al.*, 2004) employs a constellation of gestures with dynamically specified parameters (gestural scores), as a model input for an utterance. The model computes task-dynamic speech coordination among the articulators, which are structurally coordinated with the gestures along with the time functions of the physical trajectories for each vocal tract variable. The time functions of model articulators are input to the vocal tract model which computes the area function and the corresponding formants. Given English text or ARPABET, TADA generates input in the form of formants and TV time functions. The formants and pitch information were used by HLsyn™ (a parametric quasi-articulator synthesizer developed by Sensimetrics Inc., [Hanson & Stevens, 2002]) to produce a synthetic waveform. Figure 3.1 shows the flow-diagram of the TADA model and Figure 3.2 demonstrates how articulatory information (i.e., articulatory gestures, tract variables and pellet trajectories) is obtained from TADA.

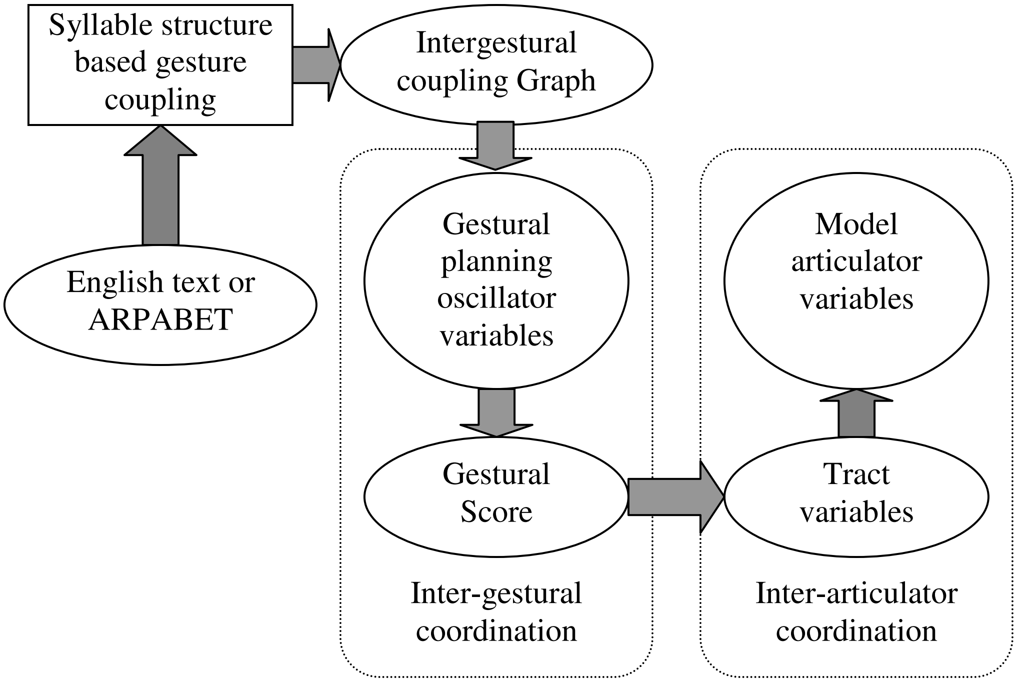

Syllable structure based gesture coupling
Intergestural coupling Graph
English text or ARPABET
Gestural planning oscillator variables
Model articulator variables
Gestural Score
Tract variables
Inter-gestural coordination
Inter-articulator coordination

Figure 3.1 Flow of information in TADA

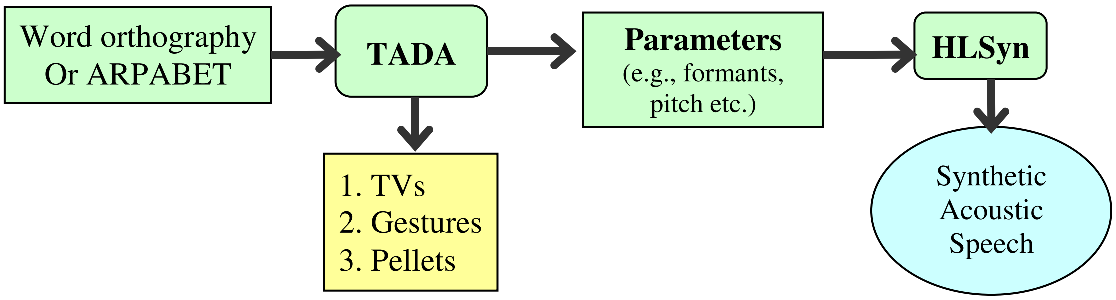

Word orthography Or ARPABET
TADA
Parameters
(e.g., formants, pitch etc.)
HLSyn
1. TVs
2. Gestures
3. Pellets
Synthetic Acoustic Speech

Figure 3.2 Synthetic speech and articulatory information generation using TADA and HLSyn

In the task dynamic model, gestures are defined with eight vocal tract constriction variables as shown in Table 3.1. The vocal tract time functions or TVs are time-varying physical realizations of gestural constellations at the distinct vocal tract sites for a given utterance. Figure 3.3 shows the gestural activations and TVs for the utterance “miss you” obtained from TADA. The larger square blocks in Figure 3.3 correspond to the gestural specifications for /m/, /i/, /s/, /y/ and /u/ in the utterance “miss you”. It can be seen in Figure 3.3 that the ‘narrow-palatal’ and the ‘narrow-velic’ TBCD gestures for the /y/ and /u/, respectively, overlap with one another from 0.1125s to 0.15s. In this region, the two gestures temporally overlap with each other in the same TV. This overlap results in blending of their dynamic parameters. The degree of blending between the gestures is defined by a blending parameter. When a gesture is active in each TV, it is distinctively specified by such dynamic parameters as constriction target, stiffness and damping. The gestures are allowed to temporally overlap with one another within and across TVs. Note that even when a TV does not have an active gesture, the resulting TV time function can be varied passively by another TV sharing the same articulator. For example, TTCD with no active gesture can also change (such changes are usually termed as passive movements of a TV) when there is an active gesture in TBCD because the tongue body and the tongue tip are coupled with one another. Figure 3.3 shows that even though TTCD does not have an active gesture from 0.125s to 0.25s, the corresponding TV moves passively since TBCD has an active gesture during that span.

TVs are defined by a set of uncoupled, linear, second order differential equations, shown in equation (1) (Saltzman & Munhall, 1989)

$$M\ddot{z} + B\dot{z} + K(z - z_0) = 0 \qquad (1)$$

where M, B and K are the task dynamic parameters of mass, damping coefficient and stiffness of each TV, and z and z_0 specify the target position of that TV. In the Task Dynamic model, a gesture is defined by the following parameters: (1) gestural activation, (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" (Byrd & Saltzman, 2003), (4) the damping parameter, which is typically set to "critical" in the gestural model (Byrd & Saltzman, 2003) to signify that there is no oscillatory overshoot or undershoot of the TVs 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 combined with one another. Out of these six gestural parameters, the mass and the damping parameters remain constant (i.e., use a default value by definition of the task dynamic model). The gestural activation and the stiffness parameters can be related to some extent.

Table 3.1 Constriction organ, vocal tract variables and involved model articulators

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

A gesture with a lower stiffness (e.g., a vowel) will have a longer activation interval. Similarly, gestures with a higher stiffness will have a shorter duration. The target parameter of a gesture is reflected by that gesture’s corresponding TV dynamics, i.e., the target value that the TV tries to attain.

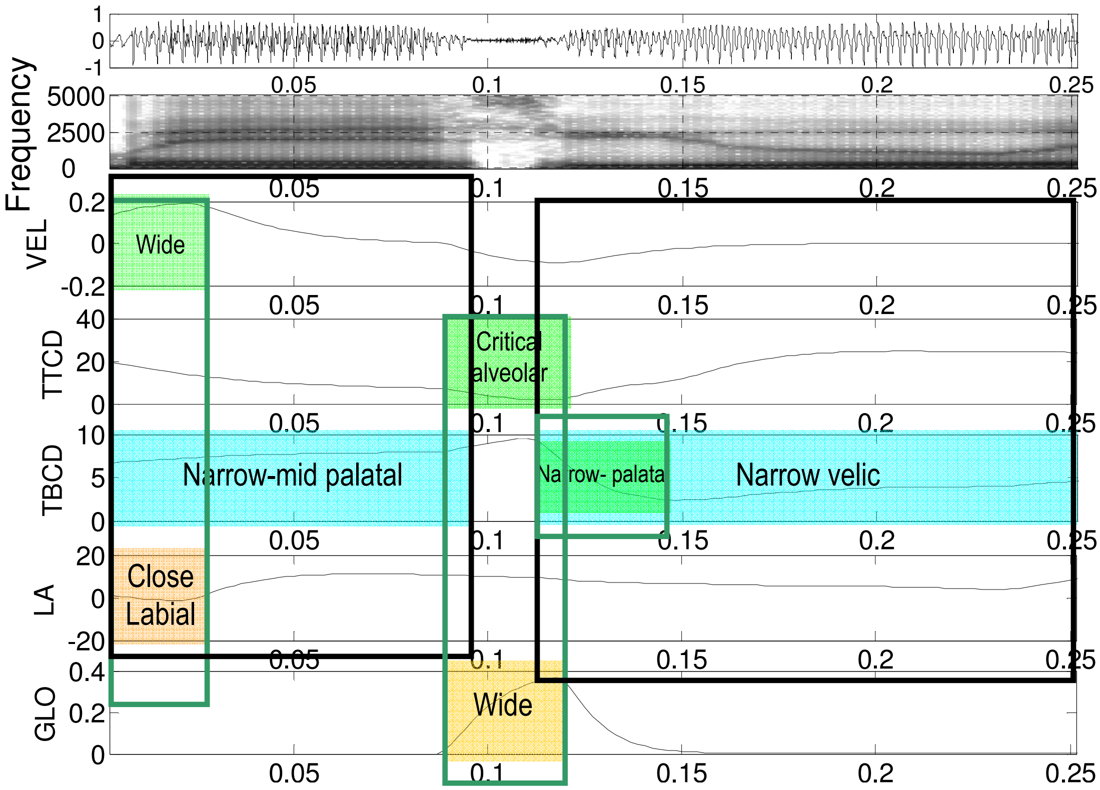


Figure 3.3 Gestural activations, TVs and approximate phone boundaries for the utterance "miss you". The active gesture regions are marked by rectangular solid (colored) blocks. The smooth curves represent the corresponding tract variables (TVs)

3.2 Synthetic database obtained from TADA and HLSyn

Three separate synthetic datasets were generated for this study. They are named as XRMB-SYN1, XRMB-SYN2 and AUR-SYN. All three databases were used for performing the initial studies reported in section 4, and they consist of TV trajectories, gestural scores, simulated pellet trajectory information (sampled at 5ms or 200Hz) and corresponding acoustic signals. Note that there are eight TV trajectories, one for each vocal tract variable shown in Table 3.1, and fourteen simulated pellet trajectories consisting of x and y co-ordinates for flesh-point locations T1, T2, T3, T4, UL, LL and Jaw which are shown in Figure 2.1. XRMB-SYN1 and XRMB-SYN2 contain isolated words taken from the XRMB

(Westbury, 1994), where XRMB-SYN1 is a subset of XRMB-SYN2. XRMB-SYN1 contains 363 while XRMB-SYN2 consists of 420 words. For both XRMB-SYN1 and XRMB-SYN2, 75% of the data were used as training samples, 10% as the validation set and the remaining 15% as the test set.

The third synthetic dataset AUR-SYN, was created to evaluate the noise robustness of the TV estimation process. This dataset is based on 960 utterances borrowed from the training corpus of the Aurora-2 (Pearce & Hirsch, 2000; Hirsch & Pearce, 2000). Although the training corpus (clean condition) of Aurora-2 has more than 8000 files, only 960 files were randomly chosen from them to build the AUR-SYN corpus. For these 960 files, the utterance, speaker's gender and their mean pitch (per file basis) were noted. The utterances were used by TADA to generate the TVs, gestural scores and the other necessary parameters required by HLsyn™. The parameters from TADA along with the mean pitch and gender information[1] were fed to HLsyn™ that generated the synthetic acoustic waveforms. The sampling rate of the TVs and gestural scores are the same as before. Seventy percent of the files from the AUR-SYN corpus were randomly selected as the training set and the rest were used as the test set. The test files were further corrupted with subway and car noise at seven different SNR levels similar to the Aurora-2 corpus.

3.3 The X-ray Microbeam database

The University of Wisconsin's X-Ray MicroBeam (XRMB) Speech Production database (Westbury, 1994) used in this study contains naturally spoken utterances both as isolated sentences and short paragraphs. The speech data were recorded from 47 different American English speakers (22 females and 25 males), where each speaker completed 56 tasks, each of which can be either read speech containing a series of digits, TIMIT sentences, or even as

[1] HLsyn in its default configuration doesn't require the knowledge of pitch and gender information; however for AUR-SYN these parameters were fed to HLsyn to create an acoustic waveform more similar to the waveforms in Aurora-2, from which the utterances were borrowed.

large as reading of an entire paragraph from a book. The sampling rate for the acoustic signals is 21.74 kHz. The data comes in three forms: text data consisting of the orthographic transcripts of the spoken utterances, digitized waveforms of the recorded speech and simultaneous X-ray trajectory data of articulator movements obtained from transducers (pellets) placed on the articulators as shown in Figure 2.1. The trajectory data were recorded for the individual articulators Upper Lip, Lower Lip, Tongue Tip, Tongue Blade, Tongue Dorsum, Tongue Root, Lower Front Tooth (Mandible Incisor) and Lower Back Tooth (Mandible Molar).

3.4 The Aurora-2 database

The Aurora-2 dataset (Pearce & Hirsch, 2000; Hirsch & Pearce, 2000) was created from the TIdigits database, which consists of connected digits spoken by American English speakers. The speech signal was sampled at 8 kHz and they are in binary raw format. There are three sections in this database, test set A, B and C; where sets A and B each have four subparts representing four different real-world noises (section A: subway, babble, car and exhibition; section B: restaurant, street, airport and train-station). Hence, altogether they have eight different noise types. Section C contains two subsections representing two noise types, one each from section A and B, but involving a different channel. As channel effects are not considered in this work, test-set C was ignored. Training in clean and the testing in a noisy condition is used in all the experiments reported in this thesis. A subset of 200 files selected randomly from each noise type at each SNR (having 1001 utterances) of the test set of Aurora-2 was selected as the development set and is termed as the “dev-set”. Note that, since the dev-set contain utterances borrowed from the test set, hence when the dev-set was used for estimating parameters of an architecture, the corresponding 200 utterances in the test set were ‘not used’ to test the architecture. Hence only the remaining 801 utterances were used.

Chapter 4: Initial study: Incorporating articulatory information for robust-ASR

Since Stevens (1960) pointed out that the anatomical or neuro-physiological representation of speech would more closely simulate the process of human speech perception in ASRs, various researchers have ventured into different approaches to create speech production and perception based ASR systems. One of the recent breakthroughs in realizing a speech production based ASR system (Livescu *et al.*, 2007b) proposed the use of articulatory features (AFs) for observation and pronunciation models. Kirchhoff (1999) and Kirchhoff *et al.* (2002) in a different study have demonstrated that AFs can also improve noise robustness of ASR systems. An overlapping articulatory feature database used by Sun *et. al* (2000b) to perform speech recognition showed an increase in recognition accuracy for the TIMIT database with respect to a baseline tri-phone HMM system.

This dissertation proposes to use speech articulatory gestures to model speech production. The AFs can be derived from phone labels and hence are synchronous with acoustic landmarks; whereas articulatory gestures are more intricately tied to the articulators. As a consequence, they are typically asynchronous with acoustic landmarks. Gestures also have been studied as the sub-word level entity for ASR tasks. However, due to the paucity of gestural specifications for a spontaneous speech corpus, such efforts have been very limited in scope. One of the initial efforts to incorporate TVs to generate gestural scores (described later in this section) was proposed by Zhuang *et al.* (2008). They proposed an instantaneous gestural pattern vector (GPV) and a statistical method to predict the GPVs from the TVs. The GPV encodes instantaneous information across all the tract variables given a gestural score, such as the constriction target and stiffness associated with gestural activation for each tract variable at that particular time.

Speech variability due to coarticulation may be addressed by articulatory phonology, which hypothesizes that human speech can be decomposed into a constellation of articulatory gestures. The advantage of articulatory phonology lies in the fact that it simultaneously captures both cognitive/discrete and physical/continuous characteristics of speech by posing constriction actions as the basic units. Since gestures are action units, they are intrinsically allowed to overlap with one another in time, as shown in Figure 4.1 (a) and (b). In this framework, coarticulatory variations are accounted for by gestural overlap in time and reduction in space. On the contrary, segmental or phonemic units occupy pre-allocated time slots so that they cannot fully account for such speech variations. Gestures on the other hand, can be modulated in their output, i.e. TVs, as a function of concurrent gestures or prosodic context while maintaining their intrinsic invariance. The phone-based model and the gesture based models are two different approaches to represent words in the lexicon. Their difference can be compared to "static" units versus "dynamic" units (Sun & Deng, 2002) or a concatenative approach versus a time-overlapping approach to represent the fundamental building blocks of speech utterances. Figure 4.1 shows why we believe that gesture-based ASR is more invariant against speech variability than ASR based on phones, di-phones or tri-phones. A comparison of the gestures in parts (a) and (b) show that the timing and degree of overlap in the gestures are very different for the carefully articulated "miss you" and the more casual production of "miss you". In part (a), the tongue tip (alveolar) constriction of the /t/ and the tongue blade (palatal) constriction for /y/ do not overlap. However, in part (b), these gestures overlap with one another considerably. As a result, the properties of the fricative change greatly. The word-final /s/ in part (a) has most of its energy above 4000 Hz as expected for an alveolar fricative. However, the fricative shown in part (b) has considerable energy starting as low as 2000 Hz. Its physical properties are more akin to a /sh/ than to a /s/. While the timing and degree of overlap between gestures vary due to changes in speech style and speech rate, the overall gestural pattern remains the same (i.e., the articulatory gestures and their sequencing in time), which highlights the invariance property of the speech gestures. Another advantage of the articulatory gesture based model would be its economical lexical representation (Tepperman *et al.*, 2009). Only 380 distinct GPVs were observed by Zhuang *et al.* (2008) for the database they used as compared to the thousands of tri-phone based models needed for a similar task. With 48 phonemes, there can be a possible set of 48*47*46 = 103776 tri-phones. However all tri-phone combinations are not valid. Usually, an exemplary database consists of 9580 tri-phones (Huang *et al.*, 2002). Use of articulatory gestures as sub-word units would enable an ASR system to account for speech variations as natural outcomes of simple modulations of gestural patterns, maintaining the unit's invariance and lexical distinctiveness. Figure 4.1 shows gestures as action units and how the degree of temporal overlap is easily expressed.

Usually coarticulation is defined as the assimilation of the place of articulation of one speech sound to that of an adjacent speech sound, or influence of one phone upon another during speech production. Often such an influence causes change in the distinctiveness of the phones which introduces variability in speech. Due to coarticulatory effects in fast speech, the articulators often fail to reach their place of articulation properly which leads to deviations in their acoustic signal from well articulated speech. In such cases, even if the articulator(s) fail to reach their respective target due to undershoot, still an effort for reaching the target should be visible in the articulatory gesture domain. In Figure 1.6, it can be seen that in a fast spoken 'perfect memory' utterance, the /t/ constriction fails to achieve its canonical target, and as a consequence failing to generate a proper /t/ burst in the acoustic output. However, in the articulatory regime, an effort toward an alveolar constriction is observed. This result shows that, due to coarticulation, gestures may be stretched or squeezed in time, but they should always be there no matter how adverse the coarticulation is; which is a direct consequence of the "invariance property of articulatory gestures". Hence, in phonetics, coarticulation is usually modeled as a transition from one gesture to another.

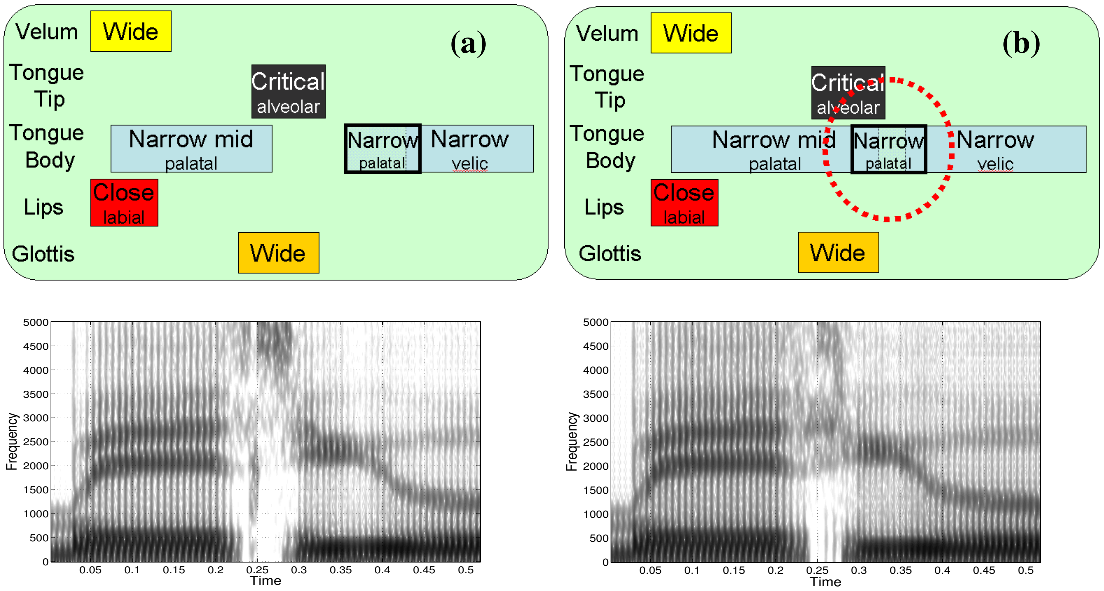


Figure 4.1 Gestural scores and spectrograms for the word (a) “miss. you” and (b) “missyou”. Note how the tongue-tip gesture for /s/ in ‘miss’ and tongue-body gesture /Y/ in ‘you’, overlap in time due to increase in speech rate (marked by the dotted circle) and correspondingly the frication energy extends till 2000Hz (with some visible formant structures) which makes the /s/ sound more as /sh/. However due to the relative invariance property of the gestures, the overall number of active gestures remain the same

The information flow in the task dynamic model depicted in Figure 3.1 and 3.2 shows that the TVs are obtained from the knowledge of the gestural scores in the forward model of TADA (using HLsyn); where the speech signal is synthesized from the knowledge of the articulator configurations. Given a speech signal, the requirement to obtain gestural scores would necessitate traversing in the opposite direction of Figure 3.1. In such case it will be reasonable to assume that the first step would be to estimate the TVs from the input speech. Finally the estimated TVs along with the acoustic waveform can be used together to estimate the gestural scores. As an initial attempt to recover gestural scores from TVs, Zhuang *et al.* (2008) proposed the GPV by sampling uniformly the gestural scores and its associated parameters in time. They used a statistical method to predict the GPVs from the TVs and obtained a prediction accuracy of 84.5% for the GPVs that have a higher frequency of occurrence. The potential advantages of estimating TVs in an intermediate stage prior to gesture recognition are twofold. First, gestures are tied to TVs in the sense that the gestural activations and their associated sets of dynamic parameters shape and control the dynamics of the TVs. Second, acoustic signals are continuous with higher bandwidth whereas speech gestures are discrete and quasi-stationary by definition having a much smaller bandwidth. Hence, it may be difficult and inaccurate to create a direct mapping between a high-bandwidth continuous regime and a locally stationary and discrete regime. On the other hand, TVs are continuous like the acoustic signal, but smoothly varying with low bandwidth like the gestural activation trajectories, and thus may be coupled well with both gestures and the acoustic signal. In other words, estimating TVs as an intermediate source of information prior to gesture recognition/recovery may provide an appropriate cross-representational bridge between the continuous and high bandwidth acoustic regime and the discrete articulatory gesture regime (i.e., gestural score). These facts suggest the necessity to perform estimation of TVs from the acoustic waveform prior to gesture recognition. Estimation of TVs from the acoustic waveform is essentially a ‘speech-inversion’ problem, which is known to be an ill-posed inverse problem as such an inversion from the acoustic space to the articulatory space is not only non-linear but also non-unique. The following section introduces the basic ideas of a speech-inversion problem and presents the different machine learning strategies used in this research to perform such an inverse task.

4.1 Estimating TVs from the Speech signal

The problem of estimating TVs from the acoustic parameters derived from the speech signal can be posed as a non-linear function estimation problem, where the TVs are represented as a non-linear function of the acoustic parameters. This nonlinear mapping between the acoustic parameters (derived from the acoustic waveform) and the TVs can be identified as an inverse problem. We can also think of this inversion as a time-series prediction problem. Speech inversion has been a field of active research in the last 40 years. The difference between the speech inversion task addressed in this proposal and the others discussed in the literature lies in the type of articulatory information used. The articulatory information used in previous studies were usually obtained from electromagnetic mid-sagittal articulography (EMMA) or electromagnetic articulography (EMA) (Ryalls & Behrens, 2000) data and were represented in terms of the cartesian coordinate displacements of pellets (transducers) placed on the articulators. Such pellet data is also known as flesh-point trajectories as they represent articulator flesh-point positional information in time. In this dissertation we will use pellet trajectory/data and flesh-point trajectory/data interchangeably. In contrast to pellet data, we focus on the TVs. The benefits of using TVs as opposed to the x and y coordinates of transducers are three fold. Firstly, as McGowan (1994) pointed out, the TVs specify the salient features of the vocal tract area functions more directly than the articulators. Secondly, as the TVs are relative measures as opposed to the absolute flesh point measures, they can effectively reduce the non-uniqueness problem in speech inversion. There may be one articulatory specification in terms of constriction degree and location specified by the TVs which can have many different sets of articulatory location (in the Cartesian coordinate space) that represent the same vocal tract constriction (McGowan, 1994). Finally, the TVs are generated by TADA (in its forward model) from gestural scores to synthesize speech thus, in the reverse model that generates gestural scores from the acoustic waveform it can be assumed that *a priori* knowledge about the TVs might help in obtaining the gestural scores, given the acoustic waveform. This link between the TVs and gestural scores is the reason why speech inversion using TVs is more appropriate for a gesture-based ASR architecture.

The study reported in this section aims to perform a detailed study of the inverse mapping between the acoustic waveform and TVs and finally estimate the gestural scores by using the synthetic data obtained from TADA. The key advantages of using this synthetic data are: (a) it is completely free from measurement noise and (b) we have absolute knowledge about its groundtruth TVs and gestural scores. We pose the design of the inverse model as a non-linear non-unique ill-posed regression problem. In the following sub-section, we briefly introduce the concept of speech inversion. Later, we will introduce the different machine learning techniques that we have explored along with a comparison of the speech inversion performance from using the TVs as opposed to the conventionally used pellet trajectories.

4.1.1 What is acoustic to articulatory speech inversion?

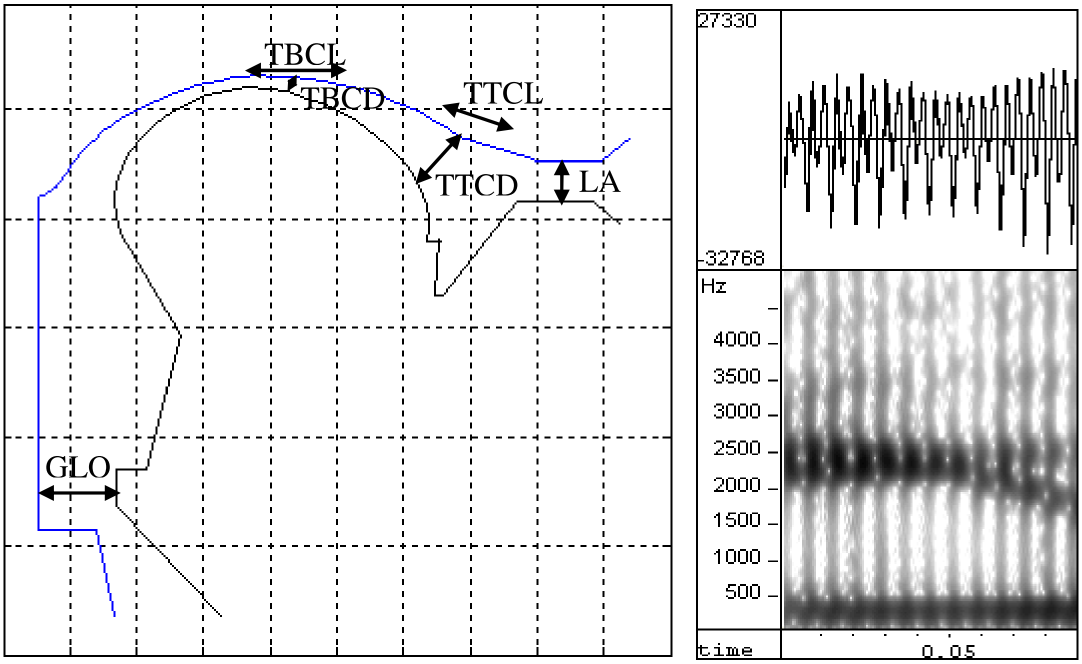
The configuration of the human vocal tract determines what speech sound is produced. This mechanism can be represented by a function *f*

$$f : x \to y \tag{2}$$

where *y* represents the speech signal, *x* represents the position of the articulators and *f* is the function that defines the mapping from the articulatory space to the acoustic space. Thus, given a vector $\bar{x}$ which is a specific articulatory configuration, we can obtain a specific speech output $\bar{y}$ when we know *f*. In most practical cases, we have the speech signal available to us with little or no articulatory data except what we can infer from the speech signal. Hence if we can define a function, *g*, such that

$$g : y \to x \tag{3}$$

then the articulatory configuration $\bar{x}$ can be obtained from the speech sample $\bar{y}$ using the function *g*. It can be observed that the function *g* is in fact the inverse of function *f*. Hence equation (3) represents the task of acoustic-to-articulatory speech inversion, i.e., given a speech signal we seek to obtain the articulatory configuration that created that speech signal. Figure 4.2 shows the vocal tract configuration for the phone '/y/' in 'miss-you'.



(a) (b)

Figure 4.2 (a) Vocal tract configuration (with the TVs specified) for the phone '/Y/' in 'miss you', (b) the corresponding time domain signal and the spectrogram

Figure 4.3 shows the human speech production mechanism and symbolically represents the speech inversion procedure. There can be many applications of acoustic-to-articulatory inversion, such as speech synthesis, speech coding, speech therapy, language acquisition, speech visualization, etc. Speech therapy deals with either speech training for subjects having difficulty in producing certain speech sounds or realizing a lip reading supplement to aid subjects with a hearing impairment. Finally, the most important application of speech inversion is in the area of robust speech recognition which has been researched actively in the last few years. Articulatory information provides information about the location, dynamics and constriction of the articulators, which can help in obtaining information such as vowel lengthening (Byrd, 2000) and prosodic stress (Cho, 2005). Such information can be exploited in an ASR system to improve its robustness against speech variability (King *et al.*, 2007; Frankel & King, 2001).

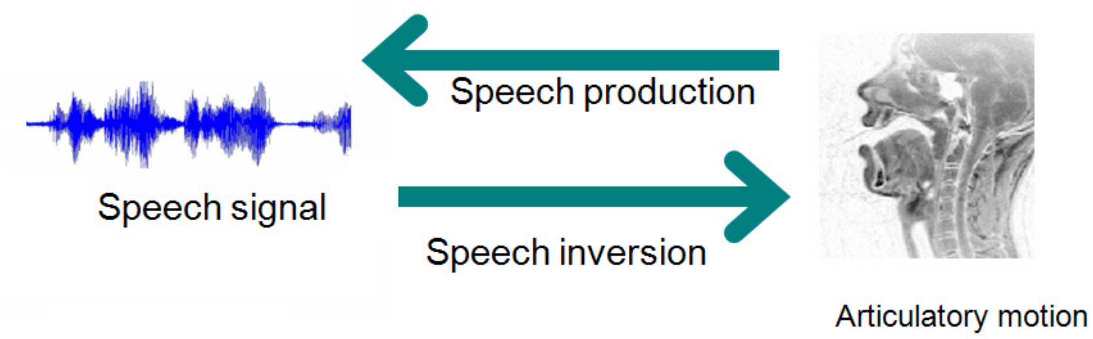


Figure 4.3 Speech Production: the forward path where speech signal is generated due to articulator movements. Speech Inversion: estimation of articulator configurations from the speech signal, commonly known as the "acoustic-to-articulatory inversion"

4.1.2 Realization of the inverse model

The challenge in the realization of the inverse model lies in the fact that the mapping from the speech signal to the TVs can be non-unique. This property renders the estimation of the non-linear function *g* in equation 3 as an ill-posed problem. Evidence from theoretical analysis, measurements from human articulatory data and also experimental analysis has indicated the existence of non-uniqueness in the functional relationship between speech and the articulatory data. The many possible articulatory configurations corresponding to a speech segment is often identified as the 'fibers' in articulatory space (Neiberg *et al*., 2008). This non-uniqueness in the inverse mapping from speech to the articulators arises when two or more different articulatory configurations are capable of producing the same (or very similar) sound(s). Hence given an acoustic waveform, which can be created from *C* different articulatory configurations, it becomes extremely difficult (if not impossible) to predict which of these *C* possible candidates generated the given speech.

Most efforts in speech inversion have focused on addressing the issue of non-uniqueness. In a study of non-uniqueness in speech-inversion, Neiberg *et al.* (2008) fitted data from acoustic and articulatory spaces to Gaussian mixture models (GMM) and studied the kurtosis and the Bhattacharya distance between the distributions to analyze the deviation of the modeled distributions from Gaussianity and the non-uniqueness related to articulatory configurations. They observed that stop consonants and alveolar fricatives are generally not only non-linear, but also non-unique; whereas dental fricatives are found to be highly non-linear but fairly unique. In their research, they found that the best possible piecewise linear prediction mapping cannot improve the mapping accuracy beyond a certain point. They also observed that incorporating dynamic information improved the performance, but did not completely disambiguate the one-to-many mapping paradox. A related and more recent study by Ananthakrishnan *et al*. (2009) modeled the probability distribution of the articulatory space conditioned on the acoustic space using GMMs and quantified the degree of non-uniqueness as the amount of spreading of the peaks in the conditional probability distribution. They showed that the non-uniqueness is higher for stop consonants, fricatives and nasals as compared to vowels, liquids and diphthongs.

Richmond, in his thesis (2001) visually demonstrated the non-uniqueness using articulatory probabilitygrams (a sample probabilitygram is shown in Figures 4.4 and 4.5). He trained a set of mixture density networks (MDNs), one each for each pellet trajectory. MDNs are essentially Multi-Layered Perceptrons (MLPs) that predict GMM parameters that provide the conditional *pdf*s of the pellet trajectories conditioned on the acoustic space. Richmond trained MDNs that predicted the parameters of 2 Gussian mixtures and witnessed a phenomenon similar to the ‘critical articulator’ phenomenon noted by Papcun *et al.* (1992). He also observed multi-modality in the inverse mapping, which he noted as the indication of non-uniqueness in the inverse mapping. Figures 4.4 and 4.5 shows an MDN probabilitygram borrowed from Richmond’s thesis (2001), which shows the tongue-tip y-axis trajectory and the lip aperture y-axis trajectory from the actual MOCHA data and also the corresponding *pdf* obtained from the MDN. It clearly shows that for consonants such as /s/, /sh/, /t/, where the tongue tip plays a critical role in pronunciation, the tt_y (Tongue Tip y-coordinate) channel shows less variance and hence is darker in the probabilitygram. On the other hand, the same channel for other consonants shows more variability as it is not critical for the production of those sounds.

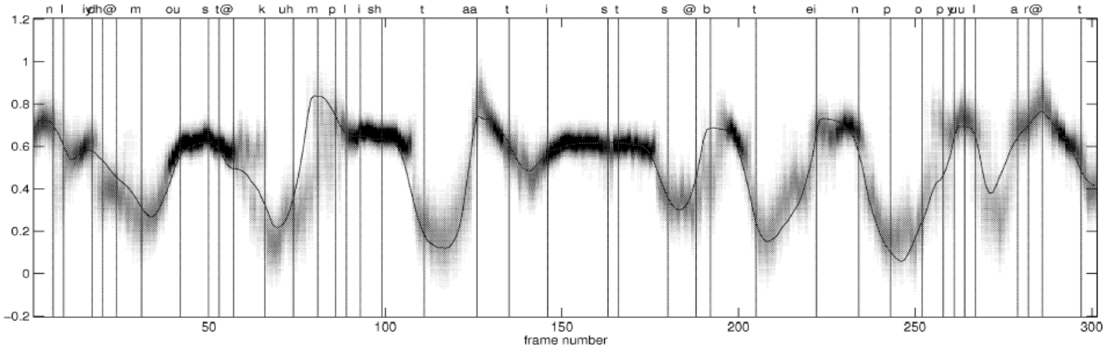


Figure 4.4 Overlaying plot of the Mixture Density Network (MDN) output (probabilitygram) and the measured articulatory trajectory (continuous line) for tt_y channel for the utterance “Only the most accomplished artists obtain popularity” from the MOCHA dataset (Wrench, 1999). Plot borrowed with permission from Richmond (2001)

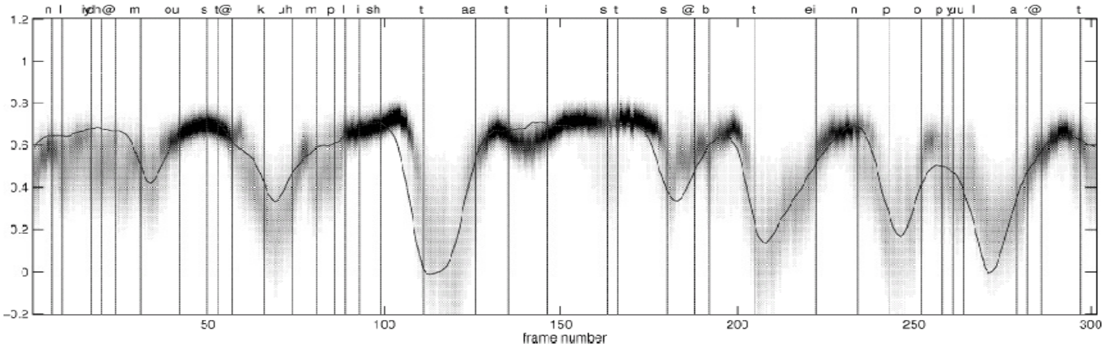


Figure 4.5 Overlaying plot of the Mixture Density Network (MDN) output (probabilitygram) and the measured articulatory trajectory (continuous line) for li_y (Lip incisor y-coordinate) for the utterance “Only the most accomplished artists obtain popularity” from the MOCHA dataset (Wrench, 1999). Plot borrowed with permission from Richmond (2001)

Non-uniqueness has also been studied by Dusan (2000). In most of the work related to studying non-uniqueness of the speech-inverse mapping, it was observed that a static solution (that is an instantaneous mapping for speech inversion) suffers largely from the non-uniqueness issue (Dusan, 2000). Incorporating dynamic information (Dusan, 2000, 2001; Richmond, 2001) about the acoustic data may help to disambiguate points of instantaneous one-to-many mappings, but would increase the difficulty of the non-linear mapping problem. Our initial results using feed-forward artificial neural networks (FF-ANNs) with a single hidden layer and 100 neurons show a significant improvement in the correlation score[2] between the actual and the reconstructed TVs from 0.853 to 0.958 for the Glottal TV (GLO) and 0.754 to 0.95 for the Velic TV (VEL). The instantaneous mel-frequency ceptral coefficients (MFCCs) were used as opposed to using the same with a contextualized window of 170ms. Table 4.1 compares the Pearson product moment correlation (PPMC) between the actual and reconstructed TVs obtained from using MFCCs with and without contextual information. PPMC indicates the strength of a linear relationship between the estimated and the actual trajectories and is defined as –

$$r_{PPMC} = \frac{N\sum_{i=1}^{N}\hat{\tau}_i\tau_i - \left[\sum_{i=1}^{N}\hat{\tau}_i\right]\left[\sum_{i=1}^{N}\tau_i\right]}{\sqrt{N\sum_{i=1}^{N}\hat{\tau}_i^2 - \left(\sum_{i=1}^{N}\hat{\tau}_i\right)^2}\sqrt{N\sum_{i=1}^{N}\tau_i^2 - \left(\sum_{i=1}^{N}\tau_i\right)^2}} \tag{4}$$

where, τ and $\hat{\tau}$ represents the actual and estimated TV vector and N represents their length.

To exploit the benefit of dynamic information, Toutios & Margaritis (2005a-b), Richmond (2001), Papcun *et al.*(1992) and many others constructed an input feature vector spanning a large number of acoustic frames, hence incorporating contextual information into the non-linear function optimization problem. Our approach to the speech inversion problem is similar to theirs in the sense that we explore popular non-linear function approximation techniques using dynamic information (i.e., contextual information) in the acoustic space and we term this model as the direct inverse model.

[2] From now on ‘correlation’ refers to the Pearson Product Moment correlation (PPMC) between the actual function and the estimated function.

Table 4.1. Correlation of the TVs obtained from instantaneous mapping versus mapping using contextual information in the acoustic space, using an ANN with a single hidden layer with 100 neurons

TV	MFCC w/o context (instantaneous mapping)	MFCC with a context of 170ms
GLO	0.8534	0.9577
VEL	0.7536	0.9504
LA	0.6477	0.8483
LP	0.5636	0.7387
TBCL	0.8365	0.9418
TBCD	0.7252	0.8994
TTCL	0.7710	0.9119
TTCD	0.7045	0.8858

Several machine learning techniques have been implemented in the literature for the task of speech inversion. Toutios & Margaritis (2005a-b) used Support Vector Regression (SVR) to estimate EMA trajectories for the MOCHA database and their results were found to be quite similar to that of the ANN based approached presented by Richmond (2001). ANN is widely known for its versatility in nonlinear regression problems. However, they fall short in ill-posed regression problems where the ill-posedness is due to a one-to-many mapping. To address the one-to-many mapping scenarios, Jordan & Rumelhart (1992) proposed the supervised learning with distal teacher or distal supervised learning (DSL) and Bishop (1994) proposed Mixture density networks. Based on MDN, Richmond (2007) proposed the Trajectory Mixture Density Network (TMDN) model for speech-inversion. While SVR and ANN based approaches fall in the category of *direct inverse models*, the DSL and the TMDN approaches can be identified as *indirect inverse models*. This section introduces the various machine learning techniques that we explored in our initial speech inversion experiments using the synthetic data specified in section 3.

4.1.2.1 Hierarchical Support Vector Regression

The Support Vector Regression (Smola & Scholkhopf, 2004) is an adaptation of Vapnik's Support Vector Classification algorithm (Vapnik, 1998) to the regression case. Given a set of *N* training vectors x_i and a target vector *t* such that $t_i \in \mathbb{R}$, the SVR algorithm seeks to find an optimal estimate (in terms of Structural Risk Minimization) for the function $t = g(x)$, which has at most ε deviation from the actually obtained targets t_i for all the training data and at the same time is as flat as possible. The ε-SVR algorithm defines that estimate as

$$g(x) = \sum_{i=1}^{N} \left(\alpha_i^* - \alpha_i\right) k(x_i, x) + \beta \qquad (5)$$

where $k(\ ,\ )$ is the kernel used, β is the bias terms and α_i, α_i* are the coefficients obtained from the solution of the quadratic problem

$$\max\left[W(\alpha, \alpha^*) \mid 0 \le \alpha, \alpha^* \le C; i = 1:N; \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) = 0 \right] \qquad (6)$$

where

$$W\left(\alpha, \alpha^*\right) = \sum_{i=1}^{N} \left[(\alpha_i^* - \alpha_i) t_i - \varepsilon(\alpha_i^* + \alpha_i) \right] - \frac{1}{2} \sum_{i,j=1}^{N} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) k(x_i, x_j)$$

The constant *C* is the trade-off between the flatness of *g* and the amount up to which deviations larger than ε are tolerated in the solution. $C > 0$ and $\varepsilon \ge 0$ are parameters that are user-defined. *C* can be as high as infinity, while usual values for ε are 0.1 or 0.01. The kernel function $k(\ ,\ )$ is used to transform the data into a high dimensional space to induce non-linearity in the estimate function. SVR performs non-linear regression by projecting the data into a high dimensional space via *k*(,) and then performs linear regression in that space. We have used Radial Basis Function (RBF) kernel with user-defined *γ* parameter

$$k(x,y)=\exp(-\gamma\|x-t\|^2) \tag{7}$$

4.1.2.2 Feedforward Artificial Neural Networks (FF-ANN)

Since Papcun *et al.* (1992) used MLPs (layered ANNs using perceptron rule) to estimate articulatory trajectories for six English stop consonants, the potential of ANNs for the speech inversion task has been enthusiastically investigated. Zachs & Thomas (1994) and Richmond (2001) have studied the potential of ANNs for performing speech inversion. Once trained, ANNs require relatively low computational resources compared to other methods both in terms of memory requirements and execution speed (Mitra *et al.*, 2009a, 2010a; Richmond, 2001). ANN has the advantage that it can have *M* inputs and *N* outputs; hence a complex mapping of *M* vectors into *N* different functions can be achieved. In such an architecture, the same hidden layers are shared by all the outputs (shown in Figure 4.6), which endows the ANNs with the implicit capability to exploit any cross-correlation that the outputs may have amongst themselves (Mitra *et al.*, 2009a, 2010a). Note that the articulatory trajectories are often correlated with one another, for example the tongue tip and the tongue body are mechanically coupled with one another; hence any movement in the tongue body will also result in a movement in the tongue tip and vice-versa. ANNs can exploit such correlations due to the reason stated above. The FF-ANNs were trained with backpropagation using scaled conjugate gradient (SCG) algorithm (Moller, 1993).

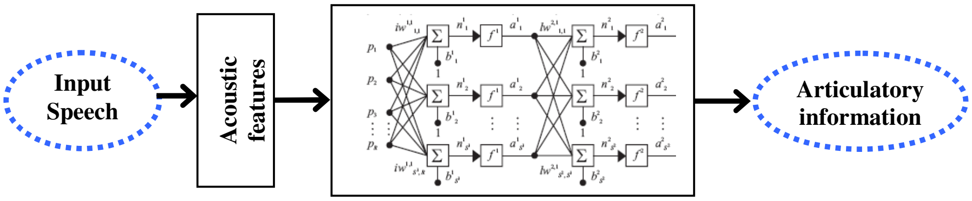


Figure 4.6 Architecture of the ANN based direct inverse model

4.1.2.3 Autoregressive Artificial Neural Networks (AR-ANN)

The estimated articulatory trajectories from SVR and FF-ANN based direct inverse models were found to be corrupted by estimation noise. Human articulator movements are predominantly low pass in nature (Hogden *et al.*, 1998) and the articulatory trajectories usually have a smoother path, defined by one that does not have any Fourier components over the cut-off frequency of 15 Hz. Nonlinear AR-ANN shown in Figure 4.7, has a feedback loop connecting the output layer with the input, which helps to ensure smoother trajectories for the articulatory motions. The output of AR-ANN can represented as –

$$\hat{y}(t) = g(\hat{y}(t-1), \hat{y}(t-2), ..., \hat{y}(t-d), u(t)) \tag{8}$$

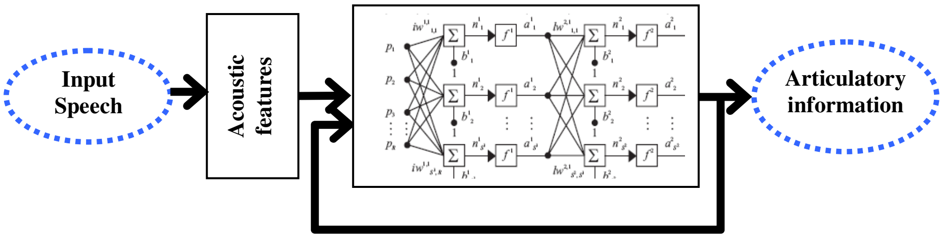


Figure 4.7 Architecture of the AR-ANN based direct inverse model

The AR-ANN has its own disadvantages: (i) the architecture has to be trained with dynamic-backpropagation or backpropagation in time, which is computationally very expensive, (ii) a single architecture cannot be trained easily for all the articulatory trajectories[3]; hence individual AR-ANNs have to be trained for each articulatory trajectory.

Both FF-ANN and AR-ANN are trained based on minimization of the sum-of-squares error approach. Given a set of training and target data set [*x*, *t*] and a set of neurons with weights and biases defined by *w* and *b* respectively, the sum-of-squares error is defined by

$$E_{SE}(w,b) = \frac{1}{2N} \sum_{i=1}^{N} \sum_{k=1}^{c} \left[g_k\left(x^i, w, b\right) - t_k^i \right]^2 \tag{9}$$

[3] This may be because the dynamics of the different trajectories are different in nature and may not correlate so strongly with one another.

where $g_k\left(x^i,w,b\right)$ defines the network output, where the network is defined by weights *w* and biases *b*. Considering a dataset of infinite size, i.e., *N*→∞, (9) can be written as

$$E_{SE}(w,b)=\lim_{N\to\infty}\frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{c}\left[g_k\left(x^i,w,b\right)-t_k^i\right]^2 \tag{10}$$

$$E_{SE}(w,b)=\frac{1}{2}\sum_{k=1}^{c}\iint\left[g_k\left(x,w,b\right)-t_k\right]^2 p(t,x)dtdx \tag{11}$$

The minimization of the error function E_{SE} with respect to $g_k(x,w,b)$ gives the following [3]

$$\frac{\partial E_{SE}}{\partial g_k\left(x,w,b\right)}=0 \tag{12}$$

Using (12) it can be shown that

$$g_k\left(x,w^*,b^*\right)=\mathrm{E}\left[t_k \mid x\right] \tag{13}$$

where E[A|B] is the conditional expectation of A conditioned on B, w^* and b^* are the weights and biases of the network after training. Hence (13) shows that networks that are optimized based on sum-of-squares approach generate average of the target data points conditioned on the input. Hence Direct inverse models obtained from supervised learning algorithms resolve one-to-*M* (where *M* > 1) inconsistencies by averaging (Bishop, 1994; Jordan & Rumelhart, 1992) across all the *M* candidates. If the set of *M* possible candidates form a non-convex set, then the average of the *M* candidates does not necessarily belong to that set, hence the solution obtained is not necessarily the correct inverse solution.

4.1.2.4 Distal Supervised Learning (DSL)

To address the issues with conventional supervised learning architectures for one-to-many mapping cases, Jordan & Rumelhart (1992), proposed Supervised Learning with a Distal Teacher or DSL. In the DSL paradigm there are two models placed in cascade with one another: (1) the forward model (which generates acoustic features given the articulatory trajectories, hence M-to-1 mapping) and (2) inverse model (which generates the articulatory trajectories from acoustic features, hence 1-to-M mapping). Given a set of [x_b, y_b] pairs, DSL first learns the forward model, which is unique but not necessarily perfect. DSL learns the inverse model by placing it in cascade with the forward model as shown in Figure 4.8.

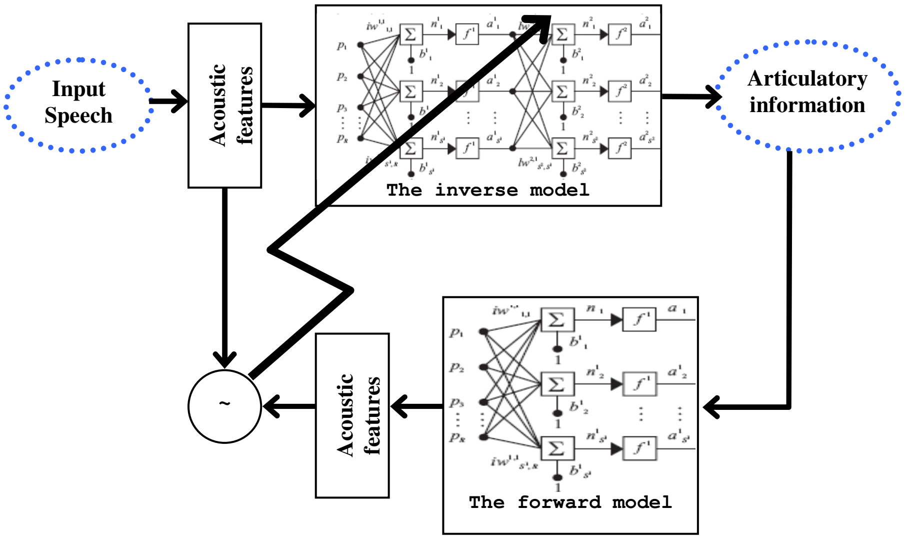


Figure 4.8 The Distal Supervised Learning approach for obtaining acoustic to TV mapping

The DSL architecture can be interpreted as an 'analysis-by-synthesis' approach, where the forward model is the synthesis stage and the inverse model is the analysis stage. In the DSL approach, the inverse model is trained (its weights and biases updated) using the error that is back-propagated through the forward model whose previously learned weights and biases are kept constant.

Considering a forward mapping between an input vector *x* and an output vector *y*, using a vector of network weights and biases, *w* and *b*, the relationship can be expressed as –

$$\hat{t} = g(x,w,b) \tag{14}$$

Learning the forward model is based on the following cost function

$$L = \frac{1}{2} E\left[(t-\hat{t})^T (t-\hat{t}) \right] \tag{15}$$

where *t* is the desired target for a given input. For the inverse model, Jordan & Rumelhart (1992) defined two different approaches, a local optimization approach and an optimization along the trajectory approach. The local optimization approach necessitates using an online learning rule, whereas the optimization along trajectory requires recurrency in the network (hence, error minimization using backpropagation in time), both of which significantly increase the training time and memory requirements. In this work we propose a global optimization approach, which uses the tools of DSL as proposed in (Jordan & Rumelhart, 1992), but instead uses batch training in the feedforward network. The cost function that the DSL tries to minimize is represented as

$$J = \frac{1}{2N}\sum_{k=1}^{N}\left[(t_k^* - t_k)^T (t_k^* - t_k)\right] \tag{16}$$

where N is the total number of training samples, t_k is the target vector for the k^{th} training sample, and t^*_k is the actual target output from the network. The weight update rule is as follows

$$w[n+1] = w[n] - \eta \nabla_w J_n \tag{17}$$

where η is the learning rate, $w[n]$ represents the weights of the network at time index n. The gradient can be obtained from (16) using the chain rule,

$$\nabla_w J_n = \frac{1}{N}\sum_{k=1}^{N}\left(-\frac{\partial x_k^{\ T}}{\partial w}\frac{\partial t^{*T}_{k,n}}{\partial x_k}(t_k - t^*_{k,n})\right) \tag{18}$$

where $t^*_{k,n}$ is the estimated target vector for the k^{th} training sample at the n^{th} time instant.

4.1.2.5 Trajectory Mixture Density Networks (TMDN)

Mixture density networks (MDNs [Bishop, 1994]) combine the conventional feedforward ANNs with a mixture model. In MDN architectures the ANN maps from the input vector x to the parameters of a mixture model (shown in Figure 4.9) to generate a conditional *pdf* of the target t conditioned on the input x. Typically a Gaussian mixture models (GMM) is used in the MDN setup because of their simplicity and the fact that a GMM with appropriate parameters can approximate any density function. A Gaussian kernel is represented as

$$k_i\left(t \mid x\right) = \frac{1}{(2\pi)^{0.5c}\sigma_i\left(x\right)^c}\exp\left[-\frac{\left\|t-\mu_i\left(x\right)\right\|^2}{2\sigma_i\left(x\right)^2}\right] \tag{19}$$

where *x* and *t* are the input and the target vector, $\mu_i(x)$ is the center of the i^{th} kernel and $\sigma_i(x)$ is the spherical covariance (this assumption can be relaxed by considering either a diagonal or a full covariance) for each Gaussian kernel and *c* is the input dimension. In this setup, the probability density of the target data conditioned on the input using a GMM with *m* mixtures can be represented as

$$p\left(t \mid x\right) = \sum_{i=1}^{m}\alpha_i(x)k_i(t \mid x) \tag{20}$$

where $\alpha_i(x)$ is the prior probability and $k_i(t|x)$ is the conditional probability density given the i^{th} kernel. To satisfy the following conditions for the prior probabilities

$$\sum_{i=1}^{m}\alpha_i(x) = 1 \ and \ 0 \le \alpha_i(x) \le 1 \tag{21}$$

The following 'softmax' function is used to define $\alpha_i(x)$ (Bishop, 1994)

$$\alpha_i = \frac{\exp(z_i^{\alpha})}{\sum_{l=1}^{m}\exp(z_l^{\alpha})} \tag{22}$$

where z_i^{α} is the ANN output corresponding to the prior probability for the i^{th} mixture of the GMM component. The variances and means of the GMM model are related to the ANN outputs as follows

$$\sigma_j = \exp(z_j^{\sigma}) \ and \ \mu_{jk} = z_{jk}^{\mu} \tag{23}$$

where z_i^{σ} and z_i^{μ} are the ANN outputs corresponding to the variance and the mean of the j^{th} mixture. The MDN is trained by minimizing the following cost function

$$E_{MDN} = -\ln\left[\sum_{i=1}^{m}\alpha_i(x^N)k_i(t^N \mid x^N)\right] \tag{24}$$

As seen in Figure 4.9, the ANN part of MDN generates the GMM parameters which are used to estimate the cost function E_{MDN}. The cost function E_{MDN} is minimized with respect to the ANN weights and biases.

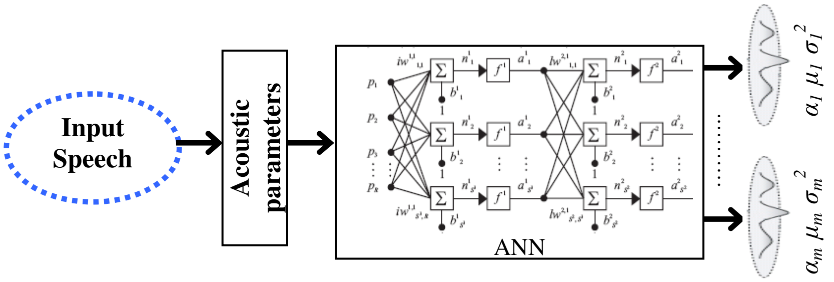


Figure 4.9 The MDN architecture

The derivative of the cost function is evaluated separately with respect to the priors, means and variances of the mixture model that are back-propagated through the network to yield the derivative of the cost function with respect to the network weights and biases, more details available at (Bishop, 1994). The standard MDN architecture provides the conditional probability density of the targets conditioned on the input. To estimate the articulatory trajectories from the conditional probability densities, a maximum likelihood parameter generation (MLPG) algorithm was proposed by Tokuda *et al.* (2000). The MLPG algorithm was used with MDN architecture by Richmond (2007) and the resulting architecture was named as the trajectory MDN or (TMDN). In TMDN architecture, the target vector is augmented with dynamic information to yield a vector sequence O as shown below.

$$O = [o_1^{\mathrm{T}}, o_2^{\mathrm{T}}, ...o_n^{\mathrm{T}}, ..., o_T^{\mathrm{T}}]^{\mathrm{T}}, \text{ where } o_n = [t_n^{\mathrm{T}}, \Delta t_n^{\mathrm{T}}, \Delta\Delta t_n^{\mathrm{T}}]^{\mathrm{T}} \qquad (25)$$

In our work the dynamic target vectors are calculated as

$$\Delta t_n = \sum_{\tau=-T/2}^{T/2} w(\tau)\, t_{n+\tau} \qquad (26)$$

$$\Delta\Delta t_n = \sum_{\tau=-T/2}^{T/2} w(\tau)\, \Delta t_{n+\tau}$$

where (T+1) is the total duration of the window and the window is defined as

$$\begin{aligned} & w(\tau) = m(\tau)\omega_{ham}(\tau) \\ & where,\ m(\tau) = \begin{cases} -1, & if\ \tau < 0 \\ +1, & otherwise \end{cases} \\ & and,\ \omega_{ham}(\tau) = \left[0.54 - 0.46\cos\left(\frac{2\pi\tau}{\mathrm{T}}\right) \right] \end{aligned} \tag{27}$$

where $\omega_{ham}(\tau)$ is a hamming window. The vector O can be related to the target vector by the following relation, where the details about the transformation matrix W can be found from Tokuda *et al.* (2000) and Toda *et al.* (2007).

$$\begin{aligned} & O = WT \\ & T = [t_1, t_2, ..., t_N]^{\mathrm{T}} \\ & W = [w_1, w_2, ..., w_N]^{\mathrm{T}} \end{aligned} \tag{28}$$

In TMDN architectures the augmented feature vector O is used to train the MDN models, where O is derived from the target vector T using the transformation matrix W. The MDN in such a case gives the following conditional density P(o_n | x_n). For the simplest case, where the GMM in the MDN has a single mixture, the target trajectory is generated by maximizing P(O | λ) or P(WT | λ) with respect to T as shown in (29), where λ is the mixture sequence.

$$\frac{\partial \log P(WT \mid \lambda)}{\partial T} = 0 \tag{29}$$

A set of linear equations are generated (detailed derivation given in Tokuda *et al.* [2000]) from (29), as

$$W^{\mathrm{T}}\Sigma^{-1}WT = W^{\mathrm{T}}\Sigma^{-1}M^{\mathrm{T}} \tag{30}$$

where

$$\begin{aligned} & \Sigma^{-1} = diag\left[\Sigma_{\lambda_1}^{-1}, \Sigma_{\lambda_2}^{-1}, ...\Sigma_{\lambda_n}^{-1}, ...\Sigma_{\lambda_N}^{-1}\right] \\ & M = \left[\mu_{\lambda_1}^{\mathrm{T}}, \mu_{\lambda_2}^{\mathrm{T}}, ...\mu_{\lambda_n}^{\mathrm{T}}, ...\mu_{\lambda_N}^{\mathrm{T}}\right]^{\mathrm{T}} \end{aligned} \tag{31}$$

μ_{λ_1} and $\Sigma_{\lambda_1}^{-1}$ are the 3x1 mean vector and the 3x3 diagonal covariance matrix (for a single mix GMM). Solving (30) for *T* gives the required maximum likelihood trajectory. For MDNs with multiple mixtures, the approximation with suboptimal mixture sequence technique discussed by Toda *et al.* (2007) is used.

4.1.2.6 Kalman smoothing

The estimated articulatory trajectories were found to be corrupted with estimation noise from all except the AR-ANN model. It was observed that smoothing the estimated articulatory trajectories improved estimation quality and the correlation and reduced the root mean square error (RMSE). This is a direct consequence of the observation made by Hogden *et al.* (1998), which claimed that articulatory motions are predominantly low pass in nature with a cut-off frequency of 15 Hz. This led us to introduce a Kalman smoother based post-processor in the architectures discussed above. Since articulatory trajectories are physical quantities, they can be approximately modeled as the output of a dynamic system. For the proposed architecture, we selected the following state-space representation

$$\begin{aligned} y_n &= Fy_{n-1} + w_{n-1} \\ t_n &= Hy_n + v_n \end{aligned} \tag{32}$$

with the following model parameters

$$\begin{aligned} F = \begin{bmatrix} 1 & \Gamma \\ 0 & 1 \end{bmatrix} &\text{ and } H = \begin{bmatrix} 1 & 0 \end{bmatrix} \\ y_0 &\sim \mathrm{N}(y_0, \hat{y}_0, \Sigma_0) \\ w_n &\sim \mathrm{N}(w_n, 0, Q) \\ v_n &\sim \mathrm{N}(v_n, 0, R) \end{aligned} \tag{33}$$

where Γ is the time difference (in ms) between two consecutive measurements, $y_n=[y_n^p \;\; y_n^v]^T$ is the state vector and contains the position and velocity of the articulatory trajectories at time instant n. t_n is the estimated articulatory trajectory which is considered as noisy observation of the first element of the state y_n. The variables w_n and v_n are process and measurement noise, which have zero mean, known covariance Q and R, and are considered to be Gaussian. The goal is to find the smoothed estimate of the state $y_{n\,|\,N}$ given the observation sequence $T = \{t_1, t_2, t_3\ldots,t_N\}$, that is:

$$y_{n|N} = E[\,y_n \mid t_1, t_2 \ldots, t_N\,] \tag{34}$$

Although F and H are known parameters of the state space representation, the unknown parameter set $\Theta = \{Q, R, \hat{y}_0, \Sigma_0\}$ should be learned from the training set. After learning the unknown parameter set, $\Theta = \{Q, R, \hat{y}_0, \Sigma_0\}$ the smoothed state $y_{n\,|\,N}$ is estimated by the Kalman Smoother.

4.1.3 Speech Inversion Experiments and Results

We begin our speech-inversion experiments by comparing the performance of TV estimation with pellet trajectory estimation and will show that the TVs can be estimated more accurately than the pellet trajectories. Next, we perform a detailed analysis of TV estimation using the different machine learning approaches specified in section 4.1.2.

In the experiments presented in this section XRMB-SYN2 was used as the data. The speech signal was parameterized as acoustic parameters (APs) and mel-frequency cepstral coefficients (MFCC). APs (Juneja, 2004; Chen & Alwan, 2000; Seteven *et al.*, 1999] are knowledge based acoustic-phonetic feature sets that provide phonetic information, such as formant values, pitch information, mean Hilbert envelope, energy onsets and offsets, and periodic and aperiodic energy in different subbands (Deshmukh *et al.*, 2005). A complete list of the APs is provided in Appendix A. The APs were measured using a 10ms window with a frame rate of 5ms. For the APs, the feature dimension was much higher compared to the MFCCs; 40 different APs were selected, where the selection was primarily knowledge based, supported by analyzing the correlation information of the APs with the respective TVs. For the MFCCs, 13 cepstral coefficients were extracted. Each of these acoustic features was measured at a frame rate of 5ms (time-synchronized with the TVs) with window duration of 10ms. The acoustic features and the target articulatory information (the TVs and the simulated pellet trajectories) were z-normalized and then scaled such that their dynamic range was confined within [-0.95, +0.95], except for SVR where the dynamic range is scaled between [-1, +1]. In order to incorporate dynamic information into the acoustic space, the input features were contextualized in all the experiments reported in this section. The feature contextualization is defined by the context-window parameter $\hat{C}$, where the current frame (with feature dimension *d*) is concatenated with $\hat{C}$ frames from before and after the current frame (with a frame shift of 2 or time shift of 10ms), generating a concatenated feature vector of size $(2\hat{C}+1)d$. From our empirical studies (Mitra *et al.*, 2009b), we have identified that the optimal context parameter $\hat{C}$ for the MFCCs is 8 (context duration of 170ms) and for the APs is 9 (context duration of 190ms). These values will be used in the experiments presented here.

The shape and dynamics of the estimated articulatory trajectories were compared with the actual ones using three quantitative measures: the root mean-squared (rms) error (RMSE), mean normalized rms error (Katsamanis *et al.*, 2009) and the Pearson product-moment correlation (PPMC) coefficient. The RMSE gives the overall difference between the actual and the estimated articulatory trajectories and is defined as

$$RMSE = \sqrt{\frac{1}{N}(\tau - \hat{\tau})^T(\tau - \hat{\tau})} \tag{35}$$

where $\hat{\tau}$ represents the estimated TV vector and τ represents the actual TV vector having *N* data points. The RMSE provides a performance measure in the same units as the measured articulatory trajectories. PPMC has been defined before in equation (4). Some of the TVs have a different measuring unit (e.g., TBCL and TTCL are measured in degrees) from the pellet trajectories (all pellet trajectories are measured in mm). Thus, to better summarize the inversion performance for all articulatory trajectories, we use the non-dimensional mean normalized RMSE, RMSE$_{nrm}$ (Katsamanis *et al.*, 2009) and its average, RMSE$_{nrm\_avg}$ defined by

$$RMSE_{nrm\_i} = \frac{RMSE_i}{\sigma_i} \tag{37}$$

$$RMSE_{nrm\_avg} = \frac{1}{T}\sum_{i=1}^{T} RMSE_{nrm\_i} \tag{38}$$

where *T* is the number of articulatory trajectories considered (8 for TVs and 14 for pellet trajectories).

4.1.3.1 Comparing TV and pellet trajectory estimates

TMDN has been used by Richmond (2007) to estimate articulatory pellet trajectories for the multichannel articulatory MOCHA dataset (Wrench, 1999). Results reported by Richmond (2007) indicate that TMDN offers much better accuracy over ANN for pellet trajectory estimation. Using a similar approach as Richmond (2007), we trained individual MDN models for each articulatory trajectory, where the articulatory trajectories were augmented with static, delta and delta-delta features as shown in (25). The MDN was built such that it generated the parameters of a GMM model with diagonal covariance matrix; yielding the parameters for a 3-dimensional Gaussian mixture (one dimension for each feature stream of static, delta and delta-delta features). The models were trained with 1 to 4 mixture components, but increasing the number of mixtures did not show any appreciable improvement of the results in our case; hence we will be presenting the results from the single mixture MDN only. The MDNs were built with a single hidden layer architecture, where the number of neurons in the hidden layer was optimized using the validation set. Table 4.2 shows the optimal number of neurons in MDN for each articulatory trajectory for each acoustic feature type. The networks were trained with the SCG algorithm using a maximum of 4000 training iterations. After the MDNs were trained, the MLPG algorithm was run ad-hoc on the resulting sequence of MDN generated *pdf*s for the validation set. The RMSE between the estimated and the groundtruth articulatory trajectory was used as the validation error.

The mean of the static features generated by the MDN should be equivalent to the output of a single hidden layer ANN (Richmond, 2007) having linear activation functions, as noted from (13); these outputs are considered as single-hidden layer ANN outputs. The TMDN as well as the ANN outputs for each articulatory trajectory were processed with a Kalman smoother and the results are shown in Table 4.2. The Kalman smoother was found to improve the PPMC on an average by 3% for both TVs and pellets.

Table 4.2 Optimal number of neurons for each articulatory trajectory for 1-mix MDN

TVs	MFCC	AP	Pellets	MFCC	AP
GLO	60	45	ULx	15	45
VEL	90	60	ULy	90	90
LA	60	45	LLx	60	90
LP	15	45	LLy	105	30
TBCL	105	30	JAWx	90	75
TBCD	45	15	JAWy	15	105
TTCL	60	60	TTx	105	15
TTCD	60	30	TTy	75	60
			TDx	30	15
			TDy	45	30

In addition, 3-hidden layer FF-ANN architectures with tan-sigmoid activation were implemented for both the TVs and pellet trajectories. The FF-ANN architectures had as many output nodes as there are articulatory trajectories (eight trajectories for TVs and 14 trajectories for pellet data). Single 3-hidden layer FF-ANN architecture was realized for each articulatory information type (i.e., TVs and Pellet trajectories) and for each feature type (MFCC or AP). The number of neurons in each hidden layer was optimized by analyzing the RMSE from the validation set. During the optimization stage we observed that the performance of the articulatory trajectory estimation improved as the number of hidden layers was increased. It may be the case that additional hidden layers incorporate additional non-linear activation functions into the system, which may have increased the potential of the architecture to cope with the high non-linearity inherent in the speech-inversion process. However the number of hidden layers was confined to three because (a) the error surface becomes more complex (with many spurious minima) as the number of hidden layers are increased, thereby increasing the probability that the optimization process finds a local minimum and (b) increasing the number of hidden layers increases the training time as well as the network complexity. The optimal ANN architectures for the MFCCs and APs were found to be 150-100-150 and 250-300-250[4], where the numbers represent the number of neurons in each of the three hidden layers. The 3-hidden layer FF-ANNs were trained with a training epoch of 5000 and the estimated trajectories were processed with a Kalman smoother. Post processing with Kalman smoothing decreased the RMSE on an average by 9%.

Table 4.3 shows the $RMSE_{nrm\_avg}$ and PPMC of all the TVs and Pellet trajectories from the 3 approaches discussed above. Note that lower RMSE and higher PPMC indicate better performance of the estimation. Table 4.3 shows that overall, the 3-hidden layer FF-ANN offered both lower RMSE and higher PPMC in both TV and pellet estimation tasks compared to the TMDN and 1-hidden layer ANN. Some of the TVs involve articulator

[4] The optimal number of neurons in the hidden layers was found to be very similar for TV and pellet estimation for a given acoustic feature; hence we have used the same configuration for both the types of speech inversion task.

movements that should be observed in particular pellet trajectories, whereas the others are not comparable to the pellet data at all. For example, the TV GLO represents the vibration of the vocal folds thereby distinguishing voiced regions from unvoiced ones. There is no such information present in the pellet trajectories as it is almost impossible to insert pellet transducers on the vocal chords. The TV-pellet sets that are closely related to one another are as follows – {LP: ULx, LLx}; {LA: ULy, LLy}, {TTCL, TTCD: TTx, TTy} and {TBCL, TBCD: TDx, TDy}. Table 4.4 lists the obtained PPMC for the related TV and pellet trajectory estimates from the 3-hidden layer FF-ANN when MFCCs are used as the acoustic features.

There are several important observations from Table 4.3: (a) overall the TV estimates offered better PPMC coefficients and mean normalized rms error (*$RMSE_{nrm\_avg}$*) than the pellet trajectories, (b) TMDN always showed improvement over the 1-hidden layer ANN model having the same number of neurons with linear activation function and (c) the 3-hidden layer FF-ANN with non-linear activation showed overall the best performance.

Table 4.3 Performance comparison between TV and pellet trajectory estimation

	TVs				Pellets trajectories			
	MFCC		*AP*		*MFCC*		*AP*	
	$RMSE_{nrm\_avg}$	$PPMC_{avg}$	$RMSE_{nrm\_avg}$	$PPMC_{avg}$	$RMSE_{nrm\_avg}$	$PPMC_{avg}$	$RMSE_{nrm\_avg}$	$PPMC_{avg}$
1-hidden ANN	*0.462*	0.881	*0.465*	0.886	*0.507*	0.838	*0.507*	0.849
TMDN	*0.443*	0.891	*0.456*	0.891	*0.493*	0.846	*0.499*	0.854
3-hidden FF-ANN	*0.313*	0.948	*0.317*	0.944	*0.410*	0.889	*0.407*	0.898

Table 4.4 Comparison of PPMC between relevant articulatory pellets and TVs for 3-hidden layer ANN using MFCC

TVs	PPMC	Pellets	PPMC
LP	0.927	LLx	0.788
		ULx	0.918
LA	0.894	LLy	0.889
		ULy	0.738
TTCL	0.951	TTy	0.945
TTCD	0.949	TTx	0.929
TBCL	0.968	TDy	0.974
TBCD	0.962	TDx	0.969
Avg	0.942	*Avg*	0.894

Observations from Table 4.3 are further confirmed in Table 4.4, which shows that for the best performing architecture, that is the 3-hidden layer ANN, the estimated TVs overall offered higher PPMC coefficient as compared to the relevant pellet trajectory estimates. It should be pointed out here that the average PPMC for the 3-hidden layer FF-ANN shown in Tables 4.3 and 4.4 are not the same, as Table 4.3 shows the average across all the TVs / pellets and Table 4.4 shows the average across only the relevant set of TVs/pellets as specified above. The results are indicative of the fact that the TVs can be estimated more accurately from the speech signal than the pellet trajectories. Two reasons may explain this difference. Firstly, according to McGowan (1994), the TVs specify acoustically salient features of the vocal tract area functions more directly than the pellet information. Secondly, the TVs (i.e. the constriction location and degree) are intrinsically relative measures, whereas the pellet trajectories provide arbitrary flesh-point location information in the 2D Cartesian coordinate system and are required to go through normalization (Richmond, 2001). Since the normalization process is sensitive to the nature of data, the relative nature of the information is not effectively captured. It should be noted, however, that such pellet-trajectory-associated problems were not overly severe in our experiment because, unlike the case of natural speech, there were no distortion in the data (as the data was synthetically generated using TADA) introduced by intra- and inter-speaker variability. Finally, note that the better performance of TVs does not seem to hold for the tongue body TVs. This can be possibly attributed to the different roles played by the tongue body in speech. The tongue body TVs are controlled primarily for vowels which do not usually involve very narrow constrictions in the vocal tract (although velar consonants (e.g. /k/ and /g/) do employ it). It can thus be said that TVs are superior for representing articulations with narrow constrictions (consonants), since such constrictions will have a disproportionate influence on the acoustics (Stevens, 2000a). For example, the TB constriction for a coproduced vowel will produce little modulation of the acoustics of stop closure or fricative noise, while a consonantal constriction will have a very large influence, determining if there is silence or turbulence. Also note that our main goal in retrieving articulatory trajectory information is to incorporate them for the purpose of articulatory gesture estimation. Since articulatory gestures are action units that inherently define constriction location and degree along the vocal tract, it can be surmised that the TVs would be more appropriate intermediate entities between acoustic observations and articulatory gestures rather than the flesh-point pellet trajectories. Thus, even if the pellet-trajectories are recovered more accurately than the TVs (which are not found to be the case here) they could not be expected to perform as well as the TVs in the estimation of articulatory gestures.

4.1.3.2 TV Estimation

In this section, we will provide a more detailed analysis of the TV estimation processes. Apart from the machine learning approaches explored in the last section, we will examine SVR, AR-ANN and finally DSL for TV estimation and then compare their performance with that of the MDN and FF-ANN architectures presented in the last section.

Hierarchical SVR

The task of speech inversion can be viewed as a non-linear regression problem, which can be performed using a hierarchical SVR framework (Mitra *et al.*, 2009a). In the SVR framework, speech is parameterized as MFCCs and APs and then contextualize as stated in section 4.1.3. Please note that for only the experiments involving hierarchical SVRs, the synthetic dataset XRMB-SYN1 was used. Separate SVR models with RBF kernel were trained for each TV, where the set of APs[5] for each model was selected based upon their relevance. We observed that certain TVs (TTCL, TBCL, TTCD and TBCD) are known to be functionally dependent upon other TVs, while the remaining TVs (GLO, VEL, LA and LP) are relatively independent and can be obtained directly from the acoustic features. This dependency is used to create the hierarchical architecture shown in Figure 4.10. From the results of the validation set the optimal value of C in equation (7) was found to be 1.5 and γ in equation (8) was set equal to 1/d based on results reported by Toutios & Margaritis (2005a) and Weston *et al.* (2003), where d = dimension of the input feature set.

[5] The number of pertinent APs for each TV is shown in (Mitra *et al.*, 2009) and the full list of those APs are given in Appendix A

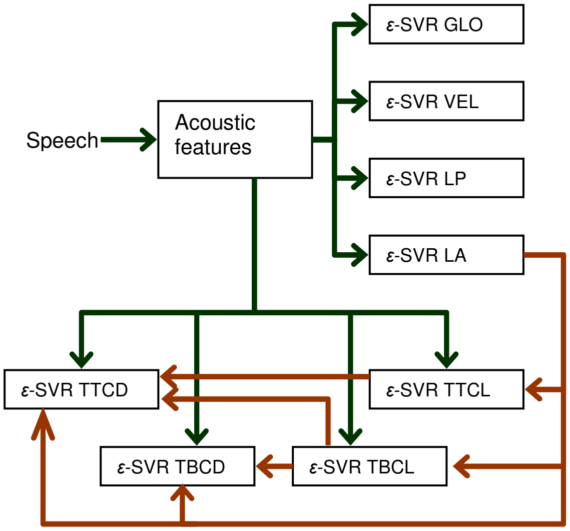


Figure 4.10 The hierarchical ε-SVR architecture for generating the TVs

AR-ANN

The estimated TVs from TMDN, FF-ANN and SVRs were found to be fairly noisy, which necessitated the use of Kalman smoother post-processing. As articulatory movements are inherently low pass in nature, maintaining smoother trajectories is a desired outcome in the speech inversion task. Using autoregressive architecture can be suitable for such an application, as the feedback loop may help to retain the smoothness of the estimated trajectories. Individual AR-ANN models were trained separately for each of the TVs. A 2-hidden layer AR-ANN model with tan-sigmoid activation, SCG training (using 5000 training epochs) with dynamic backpropagation was used. The number of neurons in each hidden layer was optimized and for all the models the number of neurons within each hidden layer was confined within 25 to 200. A unit delay[6] was used in each of the AR-ANN architectures. The TV estimates from the AR-ANNs were found to be fairly smooth, hence were not post processed with the Kalman smoother.

[6] Multiple delays were also tested, but were not found to yield appreciable improvement in performance.

DSL architecture

A single DSL architecture was trained for all the eight TV trajectories for each acoustic feature set of MFCCs and APs. The forward models were created using single hidden-layer FF-ANN and trained using the SCG algorithm. The number of neurons in the hidden layer was optimized using the rms error over the validation set. The inverse models were built using a 3-hidden-layer network and the number of neurons in each layer was optimized using the rms error on the validation set. The DSL models were trained using a gradient descent learning algorithm (with a variable learning rate), momentum learning rule (momentum = 0.9) and mean squared predicted performance error (Jordan & Rumelhart, 1992) with regularization as the optimization criteria (regularization parameter = 0.4). The number of neurons in the forward model was 350 and 400 and in the inverse model were 150-100-150 and 250-300-250 for the MFCCs and APs respectively.

Comparison of TV estimation architectures and their performance

The TV estimation results from TMDN, 3-hidden layer FF-ANN, SVR, AR-ANN and DSL are shown in Figures 4.11 - 4.13 for both APs and MFCCs. It can be observed from the plots that the 3-hidden layer FF-ANN architecture overall offered superior performance over the other approaches, closely followed by the DSL technique. For LA, DSL always performed better than the 3-hidden layer FF-ANN. The worst performance was observed from the SVR and the AR-ANN architectures. The feedback loop in the AR-ANN architecture helps to maintain the inherent smoothness of the articulatory trajectories but at the same time can be a source of progressive error introduction. If the AR-ANN model makes a significant error at any time instant, that error gets fed back to the system, resulting in progressive error in subsequent estimates. The TMDN results though were not as good as the 3-hidden layer FF-ANN, but were much better most of the time than the SVR and AR-ANN architectures.

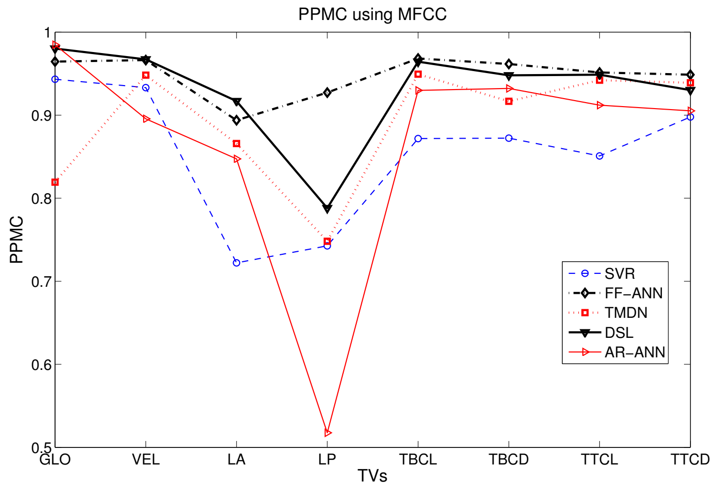


Figure 4.11 PPMC for TV estimation from different architectures using MFCCs

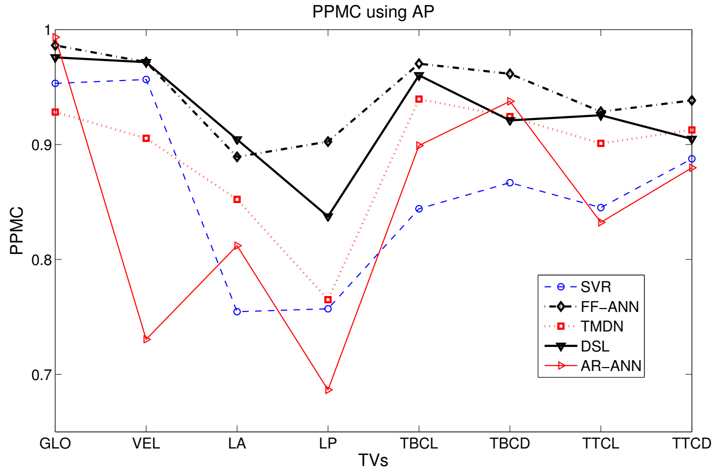


Figure 4.12 PPMC for TV estimation from different architectures using APs

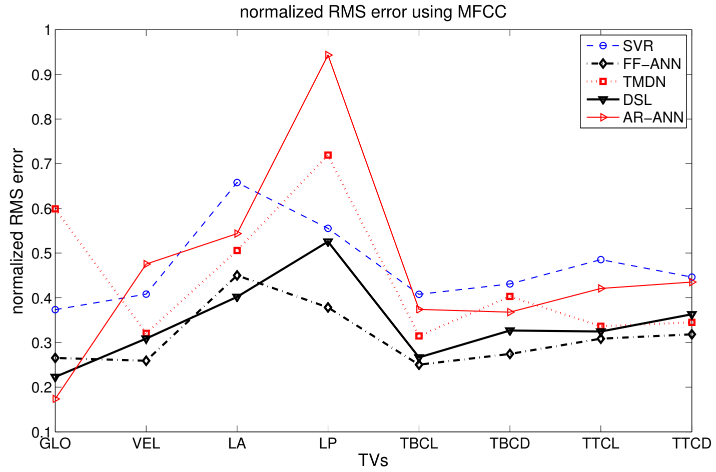


Figure 4.13 Normalized RMSE for TV estimation from different architectures using MFCCs

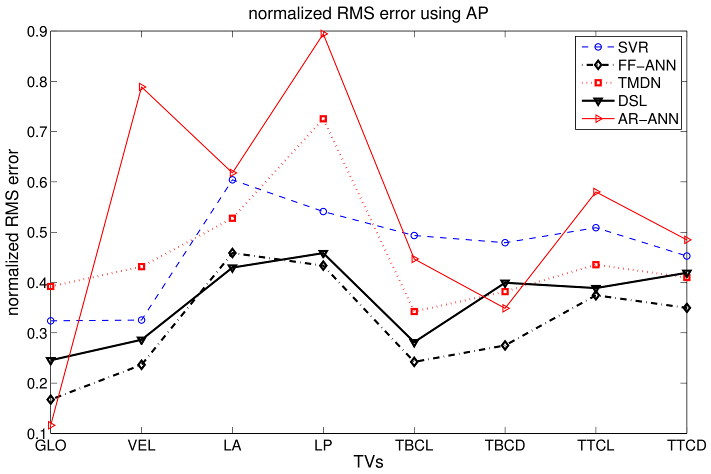


Figure 4.14 Normalized RMSE for TV estimation from different architectures using APs

Tables 4.5-4.8 presents the PPMC and RMSE obtained from the different machine learning architectures for TV estimation using acoustic features MFCCs and APs. As noted from Tables 4.7 and 4.8, different TVs have different measuring units and dynamic ranges; hence accordingly the RMSE needs to be interpreted. For example GLO and VEL have a very small dynamic range compared to others and hence very small RMSE. On the contrary, TBCL and TTCL are measured in degrees and have a larger dynamic range compared to others, hence their RMSE is in degrees and their values are larger than others.

Tables 4.5-4.8 show that the APs most of the time offered better accuracy for GLO and VEL, whereas for the other TVs, the MFCCs provided better results. The APs have specific parameters for detecting voicing (e.g., periodic and aperiodic energies at different subbands) and nasalization (Ratio of the energy in BW [0 to 320Hz] and energy in BW [320 to half the sampling rate] measured in dB, [Pruthi, 2007]). Thus, GLO and VEL are better captured using the APs.

Table 4.5 PPMC from the different TV-estimation architectures using MFCC as the acoustic feature

	SVR	FF-ANN	AR-ANN	DSL	MDN
GLO	0.943	0.965	0.985	0.980	0.819
VEL	0.933	0.966	0.896	0.967	0.948
LA	0.722	0.894	0.847	0.917	0.866
LP	0.743	0.927	0.518	0.788	0.748
TBCL	0.872	0.968	0.930	0.964	0.949
TBCD	0.872	0.962	0.932	0.948	0.917
TTCL	0.851	0.951	0.912	0.949	0.942
TTCD	0.898	0.949	0.905	0.930	0.939

Table 4.6 PPMC from the different TV-estimation architectures using AP as the acoustic feature

	SVR	FF-ANN	AR-ANN	DSL	MDN
GLO	0.953	0.986	0.993	0.976	0.928
VEL	0.957	0.972	0.730	0.972	0.905
LA	0.755	0.889	0.812	0.904	0.852
LP	0.757	0.903	0.687	0.837	0.765
TBCL	0.844	0.970	0.899	0.960	0.940
TBCD	0.867	0.962	0.938	0.921	0.924
TTCL	0.845	0.929	0.832	0.926	0.901
TTCD	0.888	0.938	0.880	0.905	0.913

Table 4.7 RMSE from the different TV-estimation architectures using MFCC as the acoustic feature

	SVR	FF-ANN	AR-ANN	DSL	MDN
GLO	0.043	0.031	0.020	0.026	0.069
VEL	0.027	0.017	0.032	0.021	0.021
LA	2.334	1.596	1.928	1.426	1.795
LP	0.538	0.366	0.913	0.509	0.696
TBCL	11.322	6.946	10.383	7.400	8.734
TBCD	1.591	1.013	1.358	1.206	1.488
TTCL	7.707	4.896	6.682	5.153	5.338
TTCD	3.277	2.337	3.197	2.667	2.534

Table 4.8 RMSE from the different TV-estimation architectures using AP as the acoustic feature

	SVR	FF-ANN	AR-ANN	DSL	MDN
GLO	0.037	0.019	0.013	0.028	0.045
VEL	0.022	0.016	0.052	0.019	0.029
LA	2.142	1.627	2.192	1.524	1.872
LP	0.524	0.420	0.866	0.444	0.702
TBCL	13.699	6.724	12.405	7.813	9.502
TBCD	1.768	1.015	1.287	1.475	1.410
TTCL	8.081	5.946	9.210	6.173	6.912
TTCD	3.324	2.568	3.560	3.080	3.010

The different architectures described above targeted different aspects of the speech inversion process. For example, AR-ANN targeted the inherent smoothness (low-frequency nature) of the TVs and the DSL and TMDN architecture were designed to explicitly address the non-uniqueness involved in speech inversion. The 3-hidden layer FF-ANN targeted the non-linearity of the speech inversion task. The better performance of the 3-hidden layer FF-ANN suggests that non-linearity may be the most critical aspect of TV estimation from the speech signal. The non-linearity in the FF-ANNs is imparted by the tan-sigmoid activations used in the hidden layers. We observed that increasing the number of hidden layers in the FF-ANN architecture resulted in an increase in the PPMC and a simultaneous decrease in the RMSE, as shown in Table 4.9, where the FF-ANN had eight output nodes (one for each TV). From Table 4.9 it can be seen that increasing the number of hidden layers increased the PPMC consistently for all but LP.

Table 4.9. PPMC for FF-ANNs with different number of hidden layers for MFCC

	GLO	VEL	LA	LP	TBCL	TBCD	TTCL	TTCD
1-hidden layer	0.942	0.951	0.872	0.928	0.956	0.946	0.929	0.928
2-hidden layer	0.960	0.961	0.885	0.925	0.967	0.960	0.940	0.939
3-hidden layer	0.965	0.966	0.894	0.927	0.968	0.962	0.951	0.949

From these observations, we re-iterate Qin *et al.*'s (2007) claim that non-uniqueness may not be a critical problem for speech inversion although their work was focused on pellet-trajectory based speech inversion. McGowan (1994) stated that the non-uniqueness in the acoustic-articulatory mapping may be reduced for the TVs compared to the pellet trajectories as there can be one articulatory specification (in terms of constriction degree and location) in TV-space which can have many different sets of articulatory location (in Cartesian coordinates) in Pellet-trajectory space, that represent the same vocal tract constriction. Hence for TVs we can expect a further (if at all any) reduction in non-uniqueness for the speech inversion task. It is well known that speech to articulatory inversion is a primarily non-linear problem (Richmond, 2001) and this fact could be the driving force behind the success of the 3-hidden layer FF-ANN. The DSL approach uses a similar architecture as the 3-hidden layer FF-ANN, but its inability to match the performance of the latter can be due to the inaccuracy in the forward model. As pointed out before, the DSL topology is more like an analysis-by-synthesis architecture, where the performance of the synthesis part entirely depends upon the accuracy of the forward model. To ensure a highly accurate forward model, exhaustive data is typically required to ensure the forward model has examples of all possible pairs of articulatory data and acoustic observation. However in a real-world scenario such exhaustive data may not be always practical rendering the inaccuracy of the forward model. An example of the predicted trajectories from the 3-hidden layer FF-ANN for five different TVs (VEL, LA, TBCL, TBCD, TTCL and TTCD) is shown in Figure 4.15 for the synthetic utterance 'a ground'. It can be seen that the raw trajectories from the FF-ANN architecture are much noisier and the Kalman-smoothing helped to reduce that noise effectively.

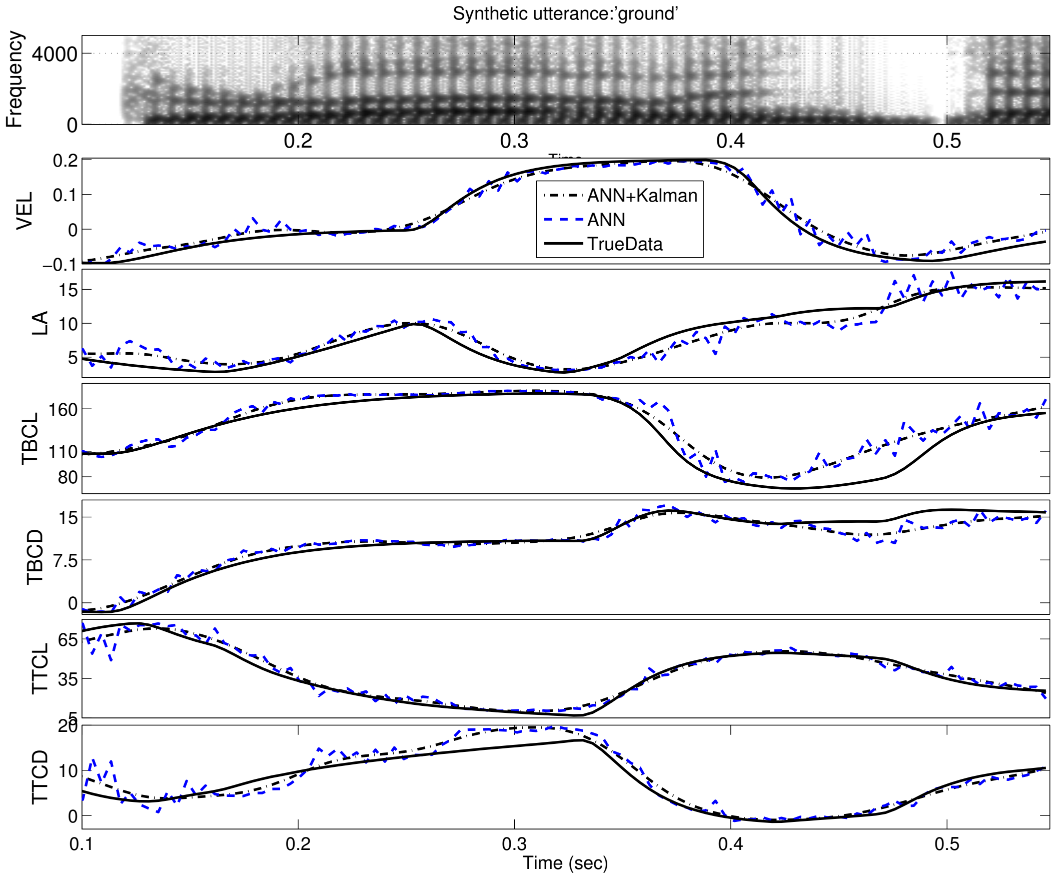


Figure 4.15 Actual and estimated TVs from ANN and ANN+Kalman using MFCC as the acoustic feature

4.1.4 Speech Inversion: *Observations*

In this section we observed using a TADA generated synthetic dataset that TV estimation can be done with overall better accuracy than estimation of articulatory pellet trajectories. This result suggests that the TVs may be a better candidate than the pellet trajectories, for articulatory feature based ASR systems. Analysis of different approaches to TV estimation suggests that for the synthetic dataset we used, non-linearity is the governing factor rather than non-uniqueness. We draw this conclusion since the 3-hidden layer FF-ANN architecture, which models well the nonlinearity inherent in speech inversion, offered much better accuracy over the other competing approaches. The 3-hidden layer FF-ANN is simpler to construct and even simpler to execute when trained, hence it would be an ideal candidate for TV estimation in a typical ASR architecture or a gesture based ASR system envisioned in this dissertation.

4.2 Recognizing Articulatory Gestures from the Speech signal

This section describes a speech articulatory gesture recognizer that recognizes articulatory gestures from the speech signal. Recollect that the speech gestures are constriction actions produced by distinct constricting organs of the vocal tract. Once a given gesture is activated, it generates the TVs that represent the degree and/or location of constriction of the associated constricting organs according to its set of corresponding dynamic parameters (target position, stiffness etc.). Recognizing gestures for a given utterance involves recovering gestural activations and their dynamic parameters. Due to the lack of any natural speech database containing such gestural information our initial experiments were performed on a synthetic speech dataset XRMB-SYN2 presented in section 3. For gesture recognition we proposed a cascaded neural network architecture for recognizing articulatory gestures from speech, where gestural activations are recognized in the first stage using an auto-regressive neural network, and the dynamic parameters associated with the activated gestures are recognized in the second stage using a feed-forward neural network.

4.2.1 Why Gesture recognition?

Note that in a typical ASR situation the only available observation is the acoustic speech signal and neither the articulatory gestures nor the TVs are readily available, hence they have to be estimated from the speech signal. Several studies have tried to obtain/annotate gestural information from the acoustic speech signal. Sun & Deng (2002) proposed an automatic annotation model of gestural scores, where the model itself was trained with manually annotated gestural scores. They showed improvement in ASR performance by using their overlapping feature-based phonological model defined by general articulatory dynamics. Gestural activation recovery (where the gestural activations represent the time interval when a gesture is active) from the acoustic signal has been performed by Jung *et al.* (1996) using a temporal decomposition (TD) method (Atal, 1983) on multi-channel articulatory trajectories. TD (Atal, 1983) models a set of speech parameters for an utterance by a sequence of overlapping target functions and their corresponding target vectors. Jung *et al.* (1996) used TD to construct a set of target functions from data-derived basis functions. The resultant target functions and weights for each basis function were used to derive the gestural score, which were applied to various CVC syllables embedded in frame sentences. However, their task was restricted to the recovery of only gestural activations and did not consider gestural dynamic parameters. The dynamic parameters of active gestures such as the stiffness and target are crucial to distinguish utterances in a gesture-based lexicon (Browman & Goldstein, 1992). The stiffness helps to distinguish consonants from vowels: the motion for consonants, which is parameterized as a gesture with higher stiffness, is faster than that of vowels. The targets provide spatial information about the location and degree of a constriction. For example, in case of /s/ as in ‘miss’ (shown in Figure 3.3), the tongue-tip gesture will have a critical constriction degree at the alveolar ridge, with an 'alveolar' TTCL target and TTCD target near 0 mm. Hence, estimating only gestural activations is not sufficient for lexical access. To address this problem, Zhuang *et al.* (2009) proposed the GPVs that are instantaneous single time slice realizations of gestural activations and their corresponding dynamic parameters, as recognition units. They proposed a tandem ANN-GMM model that predicts the GPVs from *a priori* knowledge of TVs and reported that the GPVs were correctly recognized 80% of the time and word recognition rate was 85% (Zhuang *et al.*, 2009) using the estimated GPVs for a dictionary of 139 words. However, the drawback of performing GPV recognition is that the number of possible GPVs for a speech database with a large dictionary size can potentially be huge, necessitating a large number of GPV models to be learned and evaluated. Moreover, not all GPVs occur with similar frequencies, introducing data sparsity issues similar to those encountered with tri-phone models. In addition, the GPV recognizer in Zhuang *et al.*, (2009) assumed *a priori* knowledge of the TVs without estimating them from the speech signal, which may not be practical for a typical ASR system. In a different study Tepperman *et al.* (2009) used an HMM-based iterative bootstrapping method to estimate gestural scores but their approach was limited to a small dataset.

In this section, we propose a new model that recognizes gestures directly from acoustic speech signals and aims to provide a proof of concept that, gestures indeed can be obtained from speech (for which we used a synthetic speech corpus) with a quantified degree of accuracy. Contrary to Zhuang, *et al.* (2009), gesture recognition is not performed as a frame-wise instantaneous GPV recognition; instead the task is broken into two subcomponents as two stages of a cascaded architecture: (a) recognizing gestural activation intervals in the first stage and (b) estimating the dynamic parameters for the active gesture intervals in the second stage. Separate gesture-recognition models were built for each tract variable (e.g., LA, TBCL, GLO, etc). We further examine whether TVs estimated from the acoustic signal can improve gesture recognition when combined with acoustic information. Please recollect here from section 4.0 that the potential advantages of estimating TVs in an intermediate stage before gesture recognition are twofold. First, gestures are tied to TVs in the sense that the gestural activations and their associated sets of dynamic parameters shape and control the dynamics of the TVs. Second, acoustic signals are continuous with higher bandwidth whereas speech gestures are discrete and quasi-stationary by definition having much smaller bandwidth. Hence, it may be difficult and inaccurate to create a direct mapping between a high-bandwidth continuous regime and a locally stationary and discrete regime. On the other hand, TVs are continuous like the acoustic signal, but smoothly varying with low bandwidth like the gestural activation trajectories, and thus may be coupled well with both gestures and acoustic signal[7]. In other words, estimating TVs as an intermediate source of information prior to gesture recognition/recovery may provide an appropriate cross-representational bridge between the continuous and high bandwidth acoustic regime and the discrete articulatory gesture regime (i.e., gestural score).

The goal of the study presented in this section is to develop a methodology, grounded in articulatory phonology, to recognize speech gestures, in the acoustic waveform. The recognized gestural information can be utilized as potential sub-word units in a full-blown ASR (which is the goal of this dissertation). It is therefore important to explore optimal ways of recognizing gestures and have quantitative knowledge about how accurately they can be recognized, which is addressed in this section. This study is important for several reasons. First, although gestures might be used as hidden variables in a full-blown system, finding an optimal way of recognizing them explicitly is crucial to the design and implementation of the entire recognition system. Second, gesture recognition results from synthetic data can be used as a baseline to evaluate those obtained for natural speech in the future.

4.2.2 The Gesture Recognizer

Recognizing gestures entail obtaining gestural scores (i.e., gestural activation intervals, targets, and stiffnesses) from an acoustic signal parameterized with MFCCs, APs and/or TV information. We pursued four approaches to gesture recognition from speech that differed with respect to the types of inputs used as shown by Figure 4.16(a-d). Approach-1 used the acoustic features only (i.e., the MFCCs or the APs); approach-2 used only the TVs estimated from acoustic features; approaches 3 and 4 both use TVs along with acoustic features, with the former using estimated and the latter using groundtruth TVs. Note here that the 3-hidden layer FF-ANN based TV-estimator presented in the last section has been used for estimating

[7] As evidenced by our prior research (Mitra e*t al.*, 2009, 2010), TVs can be estimated satisfactorily from the speech signal, indicating that TVs may be coupled well with the corresponding acoustic waveform.

the TVs for the gesture recognition task. The acoustic parameterization was matched for the TV-estimator and the gesture recognizers, i.e., the MFCC-based gesture recognizer used the MFCC based TV-estimator and likewise for the APs.

For all of the above four approaches we adopted a 2-stage cascade model of ANNs (shown in Figure 4.17), where gestural activation (onset and offset) information is obtained in the first stage using a non-linear autoregressive (AR) ANN, and gestural parameter estimation (target and stiffness parameters) is performed in the second stage using an FF-ANN. Note that a separate cascaded gesture-recognition model was trained for each tract variable (e.g., LA, TTCD, etc) using all of the four input combinations (shown in Figure 4.16). Altogether 4 cascade models were trained for each tract variable, except for GLO and VEL[8].

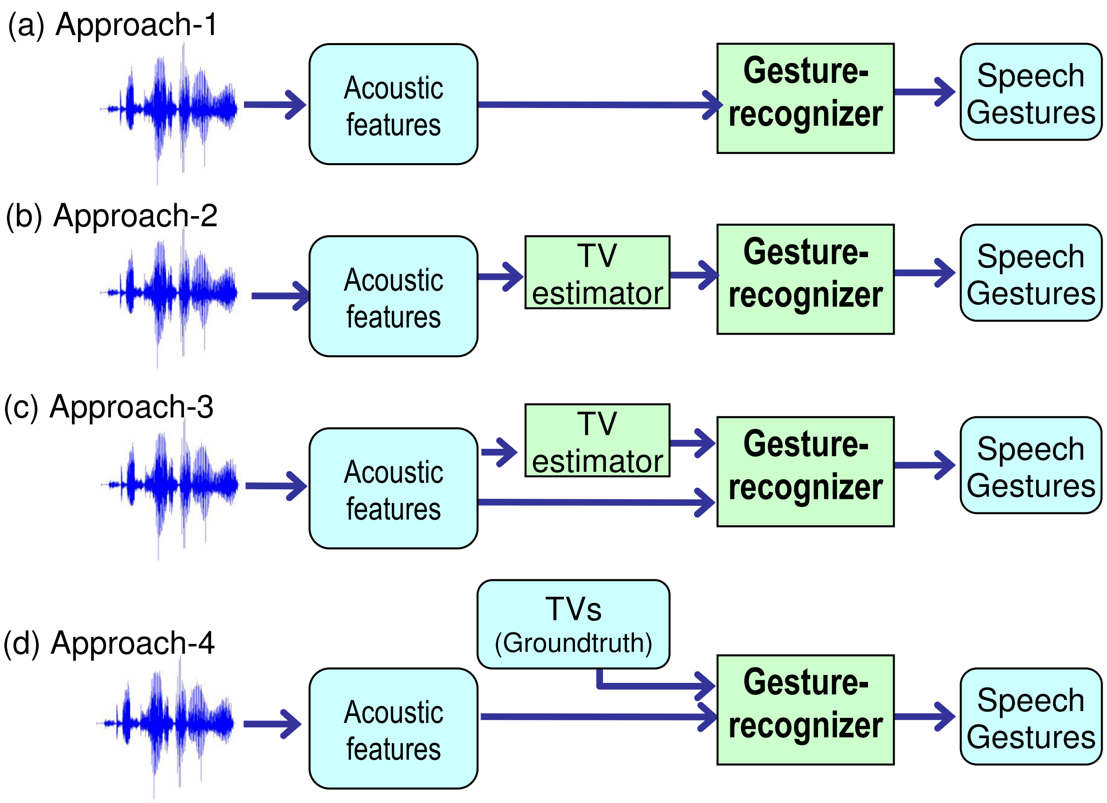


Figure 4.16 The Four approaches for Gesture recognition

[8] We observed that using only acoustic features GLO and VEL can be recognized with an accuracy of around 99%, hence for them only approach-1 in Figure 4.16 was implemented.

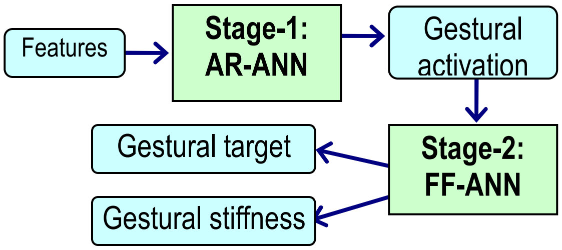


Figure 4.17 The 2-stage cascaded ANN architecture for gesture recognition

Gestural activation is discrete and quasi-stationary in nature. That is, gestural activation at any instant of time *i* can have only one of two discrete states: $S_i \in \{0,1\}$, with $S_i = 1$ when active, and $S_i = 0$ when inactive. Once a gesture is active or inactive it maintains that state for a given interval of time (at least 50ms to at most 300ms), which implies that instantaneous switching between the two states does not occur and the gesture can be considered quasi-stationary. We model this quasi-stationarity by incorporating memory into the gestural activation detection process, using a recurrent feedback loop characteristic of AR-ANN (Demuth *et al.*, 2008). Memory is used to remember the sequence of prior activation states (S_{t-1}, S_{t-2}, … $S_{t-\Delta}$) and that information along with the current acoustic observation $u(t)$ is used to predict the activation state S_t for the t^{th} time instant. As shown by equation (39)

$$S_t = f_{AR-ANN}(S_{t-1}, S_{t-2}, ..S_{t-\Delta}, u(t)) \tag{39}$$

where $f_{AR\text{-}ANN}$ represents the nonlinear AR-ANN network. Note that the autoregressive memory serves to effectively prevent instantaneous switching between activation states.

The second stage of the gesture recognition model uses an FF-ANN to predict gestural dynamic parameters: constriction targets and gestural stiffnesses (Saltzman & Munhall, 1989; Browman & Goldstein, 1992) during the active gestural intervals. Obtaining gestural dynamic parameters is essentially a function estimation problem where the parameters target and stiffness can theoretically have any real value and FF-ANNs can be trained to approximate any function (with a finite number of discontinuities) (Demuth *et al.*, 2008; Lapedes & Farber, 1988). We considered 10 different tract variables for the gesture-recognition model: LP, LA, TTCL, TTCD, TBCLC, TBCLV, TBCDC, TBCDV, VEL and GLO. Note that, since tongue body gestures are shared by velar consonants and vowels with distinct timescales (fast for consonants and slow for vowels), the original TBCL and TBCD tract variables used in TADA were partitioned into consonant (TBCLC and TBCDC) and vowel (TBCLV and TBCDV) sub-tract variables. The acoustic features (MFCCs or APs) used as the input to the cascaded ANN were temporally contextualized in a similar manner as was done for TV estimation (described in section 4.1.3) and the optimal context windows for each stage were found to vary for different tract variables.

Note that both GLO and VEL gestures are specified separately using a much simpler procedure than is used for the other tract variables. These tract variables (GLO and VEL) are independent unlike other tract variables interacting with one another due to their articulatory dependency. Similarly, all gestures for both GLO and VEL are assumed to have only one target and stiffness value, unlike gestures in the other tract variables. We observed that using approach-1 (i.e. using just the acoustic features as inputs) for GLO and VEL provided a recognition accuracy of nearly 99%; hence we have not explored the other three approaches for these two tract variables.

4.2.3 Gesture Recognition Experiments and Results

In this section we present and compare the performances of the different types of gesture-recognizers outlined in Figure 4.16 using the XRMB-SYN2 database. The four different sets of gesture-recognition models were constructed for each of the 8 gestures (LP, LA, TTCL, TTCD, TBCLC, TBCLV, TBCDC and TBCDV). Please recollect that for the GLO and VEL gestures, only approach-1 was constructed. The network configurations (i.e., input contextual information, number of neurons and the delay chain in the feedback path of the AR-ANN) were optimized separately for each TV's set of models for the first stage (i.e. the AR-ANN) and the second stage (i.e. the FF-ANN) in the cascaded architecture using the development set of XRMB-SYN2. The networks in both the stages contained a single hidden layer with tan-sigmoid activation functions, and were trained using the SCG algorithm with a training epoch of 2500 iterations. The performance of the gesture recognizers was evaluated by first quantizing the gestural parameters obtained from the second stage based on a quantization code[9] constructed from the training set, and then computing a frame-wise gesture recognition accuracy score using equation (40)

$$\mathrm{Re}c.\ Acc.\ = \frac{N-S}{N} \times 100 \tag{40}$$

where N is the total number of frames in all the utterances and S is the number of frames having at least one of the three gestural parameters (activation, target and stiffness) wrongly recognized. Figure 4.18 presents the overall gesture recognition accuracy (averaged across the 8 different gestures ignoring GLO and VEL) obtained from the four approaches using MFCCs and APs as the acoustic feature.

[9] The number of quantization levels used to perform quantization of the gestures GLO, VEL, LA, LP, TTCL, TTCD, TBCLV, TBCDV, TBCLC and TBCDC are 6, 4, 8, 10, 14, 16, 10, 10, 4 and 4 respectively

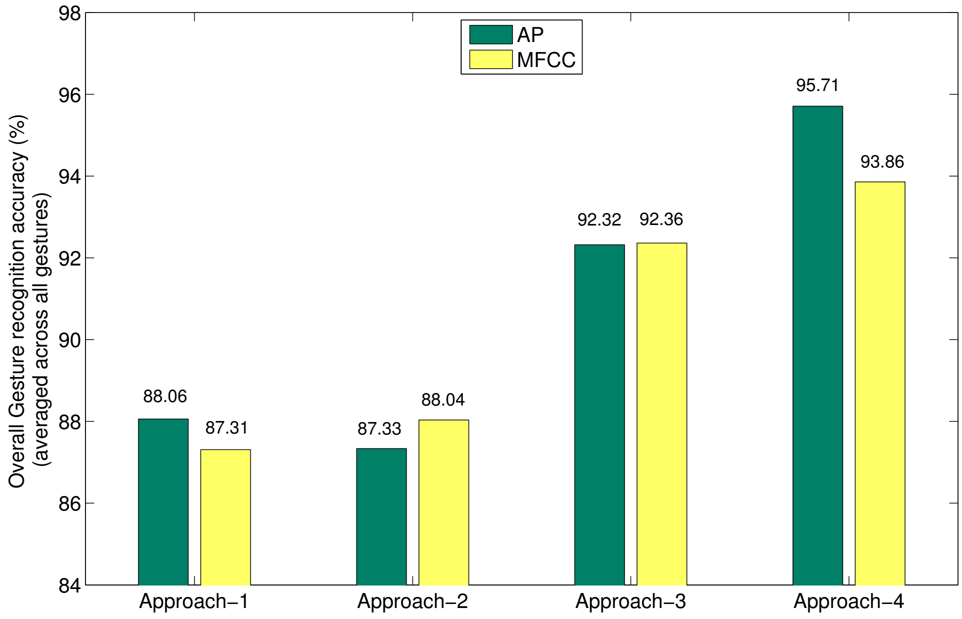


Figure 4.18 Average gesture recognition accuracy (%) obtained from the four approaches (1 to 4) using AP and MFCC as acoustic feature.

Figure 4.18 presents the following interesting observation:

(1) Approach-4 offers the best recognition accuracy for both MFCC and AP. This is expected as approach-4 uses the groundtruth or actual TVs. However, since we cannot assume *a priori* knowledge of the actual TVs, approach-4 cannot be feasibly applied for ASR of actual speech utterances. Nevertheless, approach-4 provides the theoretical accuracy ceiling that would be expected if we could have an absolutely accurate TV-estimator in approach-3.

(2) For approach-4, using APs as the acoustic feature gives higher recognition accuracy than using MFCCs, which may indicate that APs provide a better acoustic parameterization than MFCCs for gesture recognition.

(3) Approach-1 uses only the acoustic features, i.e., APs or MFCCs for gesture recognition, and as observed from Figure 4.18, APs show overall higher recognition accuracy than the MFCCs, confirming the statement made in (2).

(4) Approach-2 uses only the estimated TVs and as observed in Figure 4.18, MFCCs offer better recognition accuracy than APs. The reason for this result lies in Table 4.5-4.8, which shows that TV estimation using the MFCCs is better than TV estimation using the APs for five TVs (LA, LP, TBCD, TTCL and TTCD) out of the eight. Hence overall the MFCC based TV estimates are relatively more accurate than the APs, resulting in the MFCC based model in approach-2 to show superior performance than the AP based one.

(5) For approach-3, the AP and the MFCC based system gave almost similar recognition accuracies. While the MFCC based TV-estimator is more accurate, the APs offer better acoustic parameterization and these two counter-balances each other to show similar performance in approach-3.

(6) Approach 1, 2 and 3 are more realistic gesture-recognition architectures for ASR application, as only the acoustic features are considered as the observable and TVs in approach 2 and 3 are estimated from acoustic features. Amongst these 3 approaches, approach-3 offered the best recognition accuracy indicating that estimating TVs for gesture recognition is indeed beneficial, as we have speculated. Approach-3 is analogous to the use of tandem features used in ASR (Frankel *et al.*, 2008) where an ANN is used to perform a non-linear transform of the acoustic parameters to yield the estimated TVs, which in turn helps to improve the recognition of gestural scores when used in conjunction to the acoustic parameters. Note that the improvement caused by TVs cannot be just due to the increased number of input parameters. If that was the case, then APs would be far superior to MFCCs in approach-1.

Given these observations, we can state that the cascaded neural network gesture recognizer using acoustic features and estimated TVs as input will recognize gestures relatively more accurately than when only acoustic features or estimated TVs are used as the input. Figure 4.19 presents the recognition accuracies obtained for all gestural types, where approach-1 is only used for GLO and VEL and approach-3 is used for all of the remaining gestures. The figure shows that using approach-1 the GLO and VEL gestures were recognized quite well (accuracy > 98%). This observation is encouraging as it indicates that it may indeed be relatively simple to estimate parameters for these gestures from synthetic speech. APs offered better recognition accuracy for the GLO, VEL, TBCL-V, TBCD-V and TBCD-C gestures; this was expected as the APs have specific features for capturing voicing (the periodic and aperiodic information using the approach specified in (Deshmukh *et al.*, 2005)) and nasalization (using AP's proposed in [Pruthi, 2007]) information, whereas the MFCCs have none. However, since some AP's rely on formant information and since formant tracking using noisy speech becomes increasingly difficult and unreliable with decreasing SNR, the AP based gesture recognition models will not likely be a reliable choice for recognizing gestures from noisy speech.

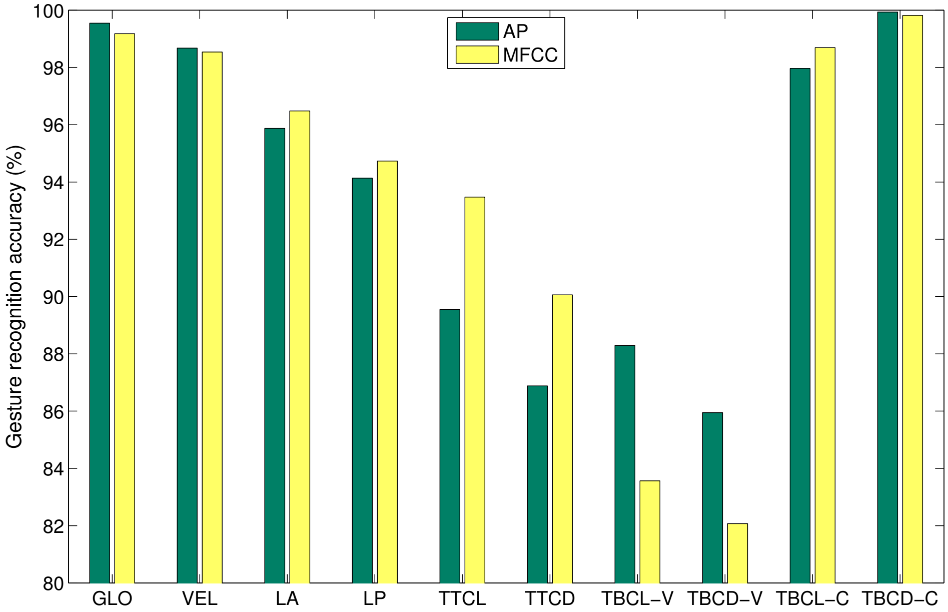


Figure 4.19 Gesture recognition accuracy (%) obtained for the individual gesture types using the cascaded ANN architecture, where the inputs for GLO and VEL were acoustic features only (i.e., AP or MFCC) while for the remainder, the input was defined by the concatenation of estimated TVs and acoustic features

Table 4.10 presents the optimal configuration for the 2-stage cascaded gesture recognition model for each gestural type. Note that in the two stages of the cascaded model, different optimal context window lengths were found for gestural activation and parameter detection. The Δ in Table 4.10 represents the order of the delay chain in the feedback path of the AR-ANN architecture used for gestural activation detection.

Table 4.10 Optimal configuration for gesture recognition (activation and parameter) using Approach-1 for GLO and VEL and Approach-3 for the rest

	AP			*MFCC*		
	Activation detection		*Parameter estimation*	*Activation detection*		*Parameter estimation*
Gesture	Δ	*Context (ms)*	*Context (ms)*	Δ	*Context (ms)*	*Context (ms)*
GLO	4	170	210	5	190	210
VEL	4	150	210	4	130	210
LA	3	90	210	10	90	210
LP	4	90	290	9	90	290
TTCL	4	90	210	4	90	210
TTCD	7	190	210	4	190	230
TBCLV	4	130	290	4	170	290
TBCDV	9	150	290	7	190	290
TBCLC	4	150	210	10	190	210
TBCDC	4	150	210	4	170	210

Note that for a given gesture, the optimal input feature context window for activation detection (i.e., for AR-ANN) is smaller compared to that used for gestural parameter estimation (i.e., for FF-ANN). This might be because the recognizer could not effectively recognize a gesture's specified target until the corresponding TV reaches its target (requiring a larger window of observation) whereas activation can be recognized by simply detecting a constricting motion on a TV (requiring a smaller observation window). Also the acoustic feature context windows for gesture-recognition are different than those used for TV estimation, where the optimal context window using MFCCs and APs was found to be 170 ms and 190 ms respectively. Hence three following factors may have contributed to the superior performance of approach-3 relative to approaches 1 and 2 (as observed in Figure 4.18):

(1) Approach-3 has the benefit of using three context windows (one each for TV estimation, activation detection and parameter estimation), and the concomitant power of the multi-resolution analysis they provide.

(2) Approach-3 uses two streams of input information: (a) acoustic features and (b) estimated TVs, whereas approach 1 & 2 uses only one of those two.

(3) Finally, as stated earlier (section 1), acoustic signals have higher bandwidth whereas speech gestures are discrete units which are quasi-stationary by definition, having bandwidth close to zero. Hence trying to create a direct mapping between them will be prone to errors, for which approach-1 may not have been as successful as approach-3. TVs are smoothly varying trajectories (with bandwidth lower than the acoustic waveform but higher than gestures) that are not only coupled strongly with gestures but are also coupled well with the acoustic signal, hence using them as an intermediate information turns out to be a better strategy.

4.2.4 Gesture Recognition: *Observations*

In this section we presented a cascaded neural network architecture for recognizing gestures from the acoustic waveform and evaluated different input conditions to obtain the best implementation. We have tested gestural recognition using four different sets of input information: (1) acoustic signals, (2) estimated TVs, (3) acoustic signals and estimated TVs and finally (4) acoustic signals and groundtruth TVs. While the first three approaches are more realistic in terms of ASR, the last one assumes prior TV knowledge. We explored the fourth approach to provide information regarding the maximum recognition accuracy that can be achieved given the TV estimator is accurate. Amongst the first three approaches, the third approach offered the best recognition accuracy, offering at least 4% improvement in performance than either of the first two approaches. Thus, we claim that incorporating estimated TVs as tandem-features can ensure higher accuracy for gesture recognition.

4.3 ASR experiments using TVs and gestures

Prior studies have shown that articulatory information, if extracted properly from the speech signal, can improve the performance of automatic speech recognition systems. We have shown in the last sections that articulatory information in the form of TVs and articulatory gestures can be obtained from the acoustic speech signal. The study presented in this section uses estimated articulatory information in the form of TVs and gestural scores in conjunction with traditional acoustic features and performs word recognition tasks for both noisy and clean speech. In this section we will show that incorporating articulatory information can significantly improve word recognition rates when used in conjunction with the traditional acoustic features.

4.3.1 Articulatory information for noise-robust ASR

Incorporating speech production knowledge into ASR systems was primarily motivated to account for coarticulatory variation. Kirchhoff (1999) was the first to show that such information can help to improve noise-robustness of ASR systems as well. She (1999) and her colleagues (2002) used a set of heuristically defined AFs, which they identified as pseudo-articulatory features. Their AFs represent the speech signal in terms of abstract articulatory classes such as: voiced/unvoiced, place and manner of articulation, lip-rounding, etc. However their AFs do not provide detailed numerical description of articulatory movements within the vocal tract during speech production. They showed that their AFs in combination with MFCCs provided increased recognition robustness against the background noise, where they used pink noise at four different SNRs. They concluded that the AFs and MFCCs may be yielding partially complementary information since neither alone provided better recognition accuracy than when both used together. In a different study, Richardson *et al.* (2003) proposed the Hidden Articulatory Markov Model (HAMM) that models the characteristics and constraints analogous to the human articulatory system. The HAMM is essentially an HMM where each state represents an articulatory configuration for each di-phone context, allowing asynchrony amongst the articulatory features. They reported that their articulatory ASR system demonstrated robustness to noise and stated that the articulatory information may have assisted the ASR system to be more attuned to speech-like information.

In this section we demonstrate that articulatory information in the form of TVs estimated from the speech signal can improve the noise robustness of a word recognizer using natural speech when used in conjunction with the acoustic features. In section 4.1 we have shown that the TVs can be estimated more accurately compared to the pellet trajectories and we demonstrated that estimation of the TVs from speech is predominantly a non-linear process. In this section we will re-train the 3-hidden layer FF-ANN TV-estimator and the gesture-recognizer models presented in section 4.1 and 4.2, using the AUR-SYN data. The trained models will finally be deployed on the natural utterances of Aurora-2 database, to estimate and recognize their corresponding TVs and gestural scores. The models were retrained using AUR-SYN as the acoustics in AUR-SYN is phonetically similar to the

Aurora-2 acoustics that we will be using for the ASR experiments. Please note here that MFCCs are used as the acoustic parameterization for these set of experiments and not the APs, as some AP's rely on formant information and formant tracking for noisy speech can potentially become difficult and unreliable with a decrease in SNR. Since the Aurora-2 database contains noise contaminated speech utterances, the APs may not be a reliable parameterization for the noisy speech utterances, especially for those having very low SNRs.

Our work presented in this section is unique in the following ways:

(a) Unlike the results reported by Frankel *et al.* (2000, 2001) we do not use flesh-point measurements (pellet trajectories) of the different articulators. Instead, we are using the vocal tract constriction trajectories or TVs, which are less-variant than the pellet trajectories (McGowan, 1994; Mitra *et al.*, 2010a). None of the work available in the literature evaluated the articulatory information (in the form of TVs) estimated from the speech signal under noisy conditions. In the present study, we show that TVs can be estimated more robustly from noise-corrupted speech compared to pellet trajectories and also that the estimated TVs do a better job than pellet trajectories when applied to word recognition tasks under noisy conditions.

(b) The work presented by Frankel *et al.* (2000, 2001) used LDM at different phone contexts to model the articulatory dynamics for clean speech, whereas we are using the TV estimates (without any phone context) directly into an HMM based word recognizer for the recognition task.

(c) Kirchhoff *et al.*'s work (1999, 2002) though uses articulatory information for noise robust speech recognition; their AFs do not capture the dynamic information about articulation but describe only the critical aspects of articulation. They are mostly hypothesized or abstract discrete features derived from acoustic landmarks or events and are not directly obtained from actual articulatory events. On the contrary, TVs provide actual articulatory dynamics in the form of location and degree of vocal tract constrictions in the production system.

(d) Kirchhoff *et al.*'s work dealt with only pink noise at four SNR levels (30dB, 20dB, 10dB and 0dB), whereas we report our results on eight different real-world noise types (subway, car, babble, exhibition, train-station, street, airport and restaurant) at six different SNRs (20dB, 15dB, 10dB, 5dB, 0dB and -5dB). Richardson *et al.* (2003) used hypothetical AFs obtained from a diphone context. Their noise robustness experiment was very limited in scope, and used stationary white Gaussian noise at 15dB SNR only.

(e) In our study, articulatory information is used across different acoustic feature sets and front-end processing methods to verify whether the benefits observed in using such articulatory information are specific to particular features or are consistent across features.

We justify the selection of TVs as opposed to the pellet trajectories by performing ASR experiments (both in noisy and clean conditions) using the TVs and the pellet-trajectories and comparing the noise-robustness witnessed in the ASR results from the two.

(f) Finally, we use the recognized gestures for performing word recognition experiments using both clean and noisy utterances and report the results in this section. Note that the only prior use of gestures for ASR was reported by Zhuang *et al.* (2009). However, in that study the ASR task was performed on a synthetic corpus identical to XRMB-SYN1. The study reported by Sun & Deng (2002) presents ASR results obtained from an overlapping articulatory feature based phonological model akin to the articulatory gestures, however their experiments were not performed under noisy conditions. Hence the experiments presented in this section for the first time explicitly uses articulatory gestures for performing word recognition task on both clean and noisy utterances.

4.3.2 ASR experiments and results using TVs and Gestures

We aim to test the possibility of using the estimated TVs and gestures as possible inputs to the word recognition task on Aurora-2 (Pearce & Hirsch, 2000) and examine whether they can improve the recognition accuracies in noise. The details of the experiments are described in the following subsections. In section 4.3.2.1, we first present the TV estimation results for the synthetic speech data for clean and noisy conditions. In section 4.3.2.2, we then apply the synthetic-speech-trained TV-estimator on the natural utterances of Aurora-2 to estimate their corresponding TVs. In section 4.3.2.3, we perform word recognition experiments using the estimated TVs and Gestures as inputs, and compare their performances when combined with traditional acoustic features (MFCCs and RASTAPLP) or other front-end processing methods.

4.3.2.1 TV Estimation in clean and noisy condition for AUR-SYN (synthetic speech)

The performance of the FF-ANN based TV-estimator is evaluated using the quantitative measures: PPMC and RMSE, shown in equations (4) and (35). The FF-ANN TV-estimator used in the experiments presented in this section was trained with the training set of AUR-SYN and the results are obtained using the test-set. Table 4.11 presents RMSE and PPMC of the estimated TVs for the clean set of AUR-SYN with and without using the Kalman smoothing. Table 4.11 shows that using the Kalman smoother helped to reduce RMSE and increase PPMC for the clean test set.

Figures 4.20 and 4.21 show RMSE and PPMC plots, respectively, of the estimated TVs at different SNRs from the test set of AUR-SYN corrupted with subway noise. As SNR decreases, the RMSE of the estimated TVs increases and their PPMC decreases, this indicates that the estimation deteriorates with decrease in SNR. Using Kalman smoothing results in lower RMSE and higher PPMC at a given SNR. The car noise part of the AUR-SYN test-set shows a similar pattern.

Table 4.11 RMSE and PPMC for the clean speech from AUR-SYN

	No-smoothing		Kalman smoothed	
	RMSE	PPMC	RMSE	PPMC
GLO	0.0196	0.9873	0.0191	0.9880
VEL	0.0112	0.9874	0.0101	0.9900
LA	1.0199	0.9654	0.9054	0.9734
LP	0.2257	0.9795	0.1986	0.9841
TBCL	2.2488	0.9966	2.0097	0.9973
TBCD	0.4283	0.9882	0.3841	0.9907
TTCL	2.9758	0.9806	2.8108	0.9830
TTCD	1.2362	0.9893	1.1722	0.9905

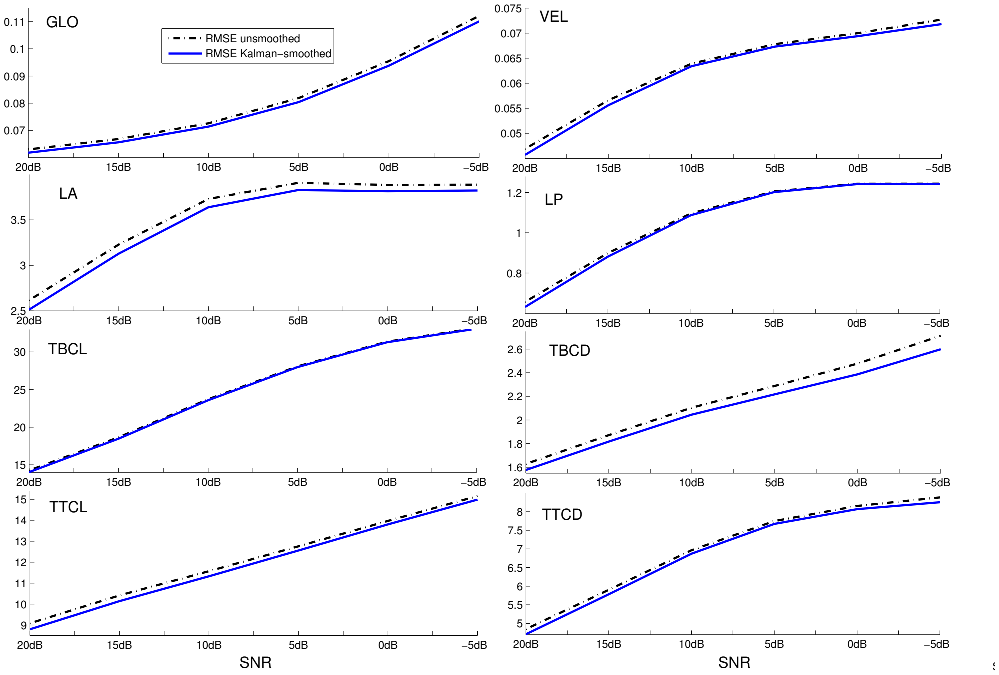


Figure 4.20 RMSE of estimated TVs for AUR-SYN (synthetic speech) at different SNRs for subway noise

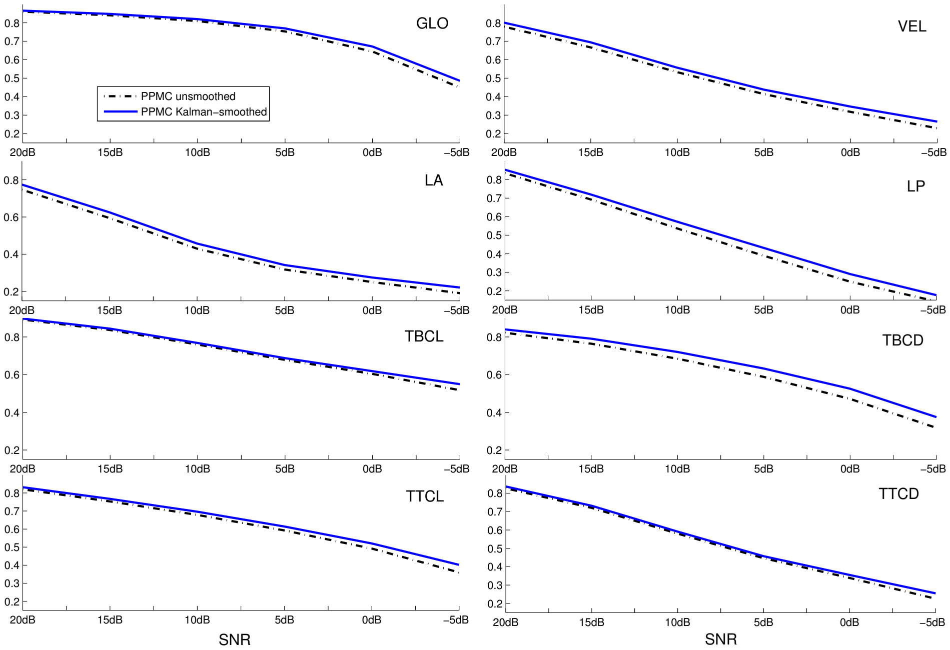


Figure 4.21 PPMC of estimated TVs for AUR-SYN (synthetic speech) at different SNRs for subway noise

4.3.2.2 TV Estimation in clean and noisy condition for Aurora-2 (natural speech)

The FF-ANN TV-estimator presented in the last section (which was trained with the clean synthetic speech from AUR-SYN) was used to estimate TVs for the natural speech of the Auroa-2 database. The raw estimated TVs were then Kalman-smoothed. Since there is no known groundtruth TVs in Aurora-2, RMSE and PPMC cannot be computed directly. Instead we compared the unsmoothed or Kalman-smoothed estimated TVs from different noise types and levels to the corresponding unsmoothed or Kalman-smoothed estimated TVs from clean utterances, to obtain the relative RMSE and PPMC measures. Figures 4.22 and 4.23 show that the relative RMSE increases and the PPMC decreases as SNR decreases for the subway noise section of Aurora-2, and Kalman smoothing helps to improve the relative RMSE and the PMMC. Note that the TV estimates for the natural utterances showed a relatively lower PPMC compared to those of the synthetic utterance (see Figures 4.20 and 4.21). This may be due to the mismatch between the training data (synthetic data of AUR-SYN) and testing data (natural utterances of Aurora-2).

Figures 4.24 and 4.25 show how the estimated TVs from natural speech look compared to those for the synthetic speech. Figure 4.24 shows the groundtruth TVs (GLO, LA, TBCL, TTCL and TTCD) and the corresponding estimated TVs for the synthetic utterance ‘two five’ from AUR-SYN for clean condition, 15dB and 10dB SNR subway noise contaminated speech. Figure 4.25 shows the same set of TVs estimated from the natural utterance ‘two five’ from Aurora-2 for clean condition, 15dB and 10dB SNR. Note that, since we do not know the groundtruth TVs for this natural utterance, it cannot be shown in the plot.

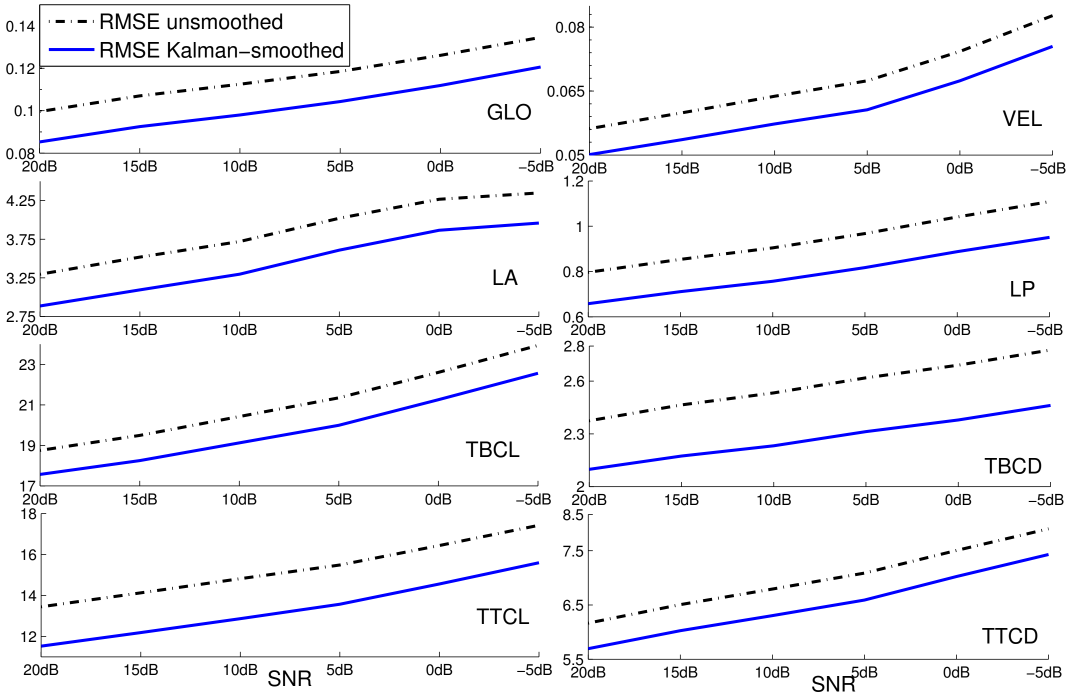


Figure 4.22 RMSE (relative to clean condition) of estimated TVs for Auora-2 (natural speech) at different SNRs for subway noise

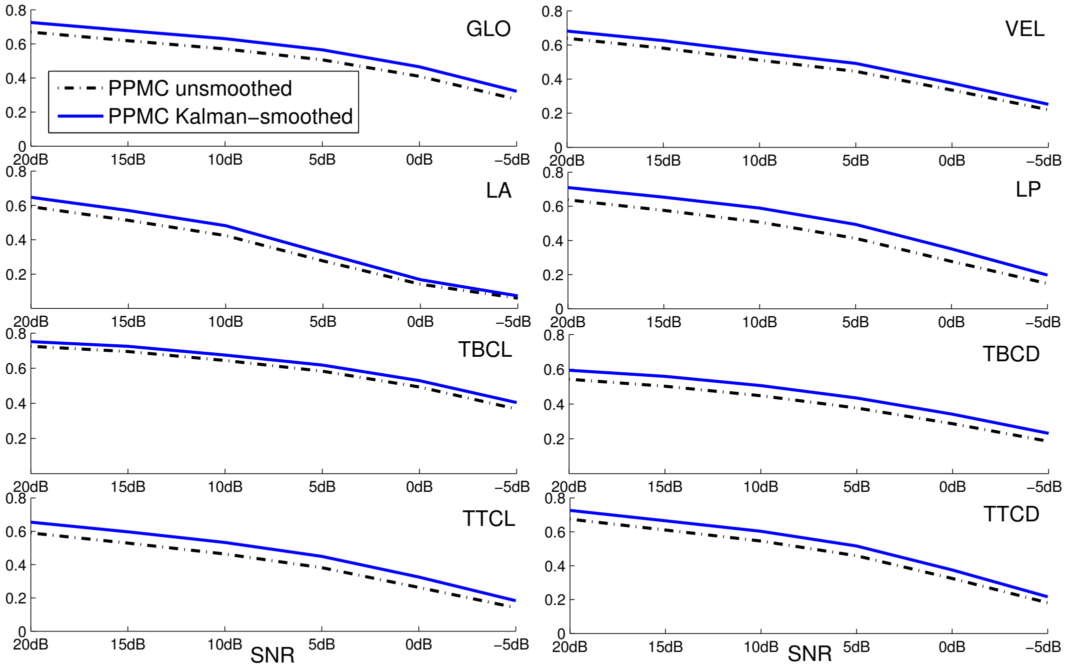


Figure 4.23 PPMC (relative to clean condition) of estimated TVs for Auora-2 (natural speech) at different SNRs for subway noise

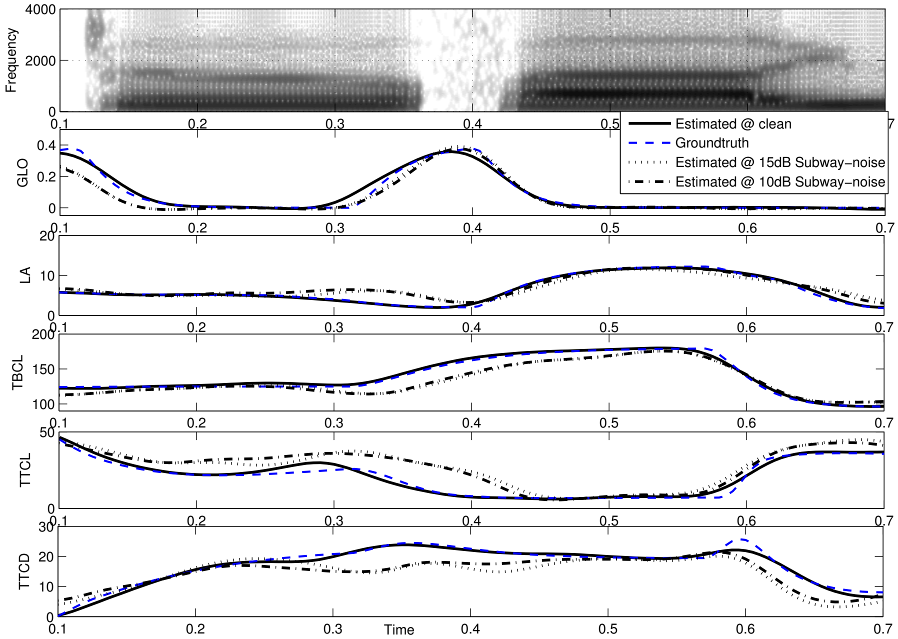


Figure 4.24 The spectrogram of synthetic utterance ‘two five’, along with the ground truth and estimated (at clean condition, 15dB and 10dB subway noise) TVs for GLO, LA, TBCL, TTCL and TTCD

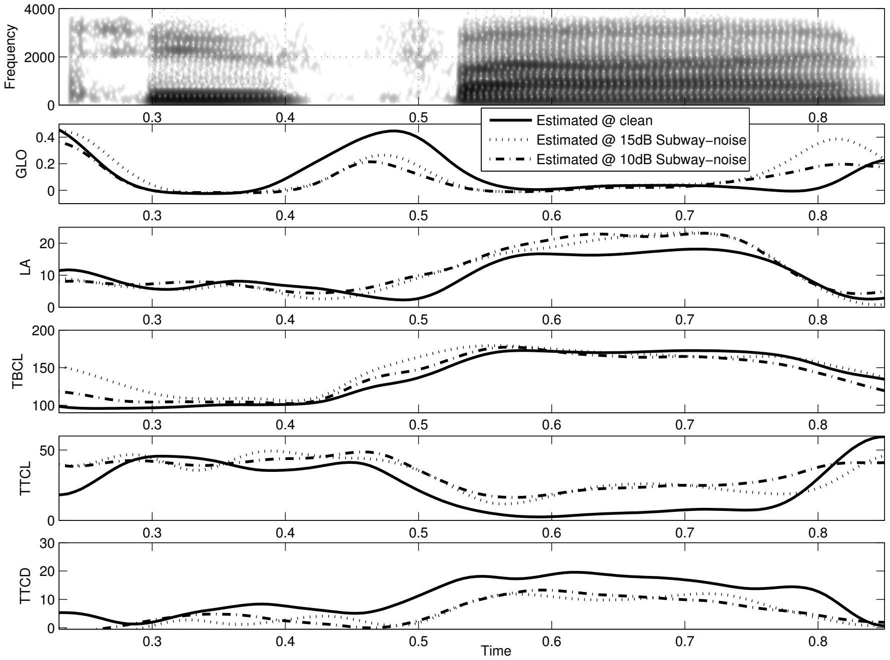


Figure 4.25 The spectrogram of natural utterance ‘two five’, along with the estimated (at clean condition, 15dB and 10dB subway noise) TVs for GLO, LA, TBCL, TTCL and TTCD

Comparing Figure 4.24 and 4.25 we observe that the estimated TVs for both the natural and synthetic speech show much similarity in their dynamics at clean condition; with noise addition the dynamic characteristics of the trajectories deviate from those in the clean condition.

In section 4.1.3.1 and (Mitra *et al.*, 2010a), we showed that TVs can be estimated relatively more accurately than flesh-point pellet trajectories for clean synthetic speech. To further validate the TV's relative estimation superiority over pellet trajectories for noisy speech, we trained a 3-hidden layer FF-ANN pellet-estimation model using TADA-simulated pellet trajectories from the AUR-SYN data. Seven pellet positions were considered: Upper Lip, Lower Lip, Jaw, and four locations on the Tongue; since each position was defined by its x- and y- coordinates, this gave rise to a 14 dimensional data trajectory which we named as Art-14. The pellet trajectory estimation model was deployed on the test set of the Aurora-2 data and the estimated pellet trajectories were smoothed using a Kalman filter. Fig. 4.26 shows the average relative PPMC across all the components of the Kalman-smoothed TV and pellet trajectory estimates for the subway noise section of Aurora-2.

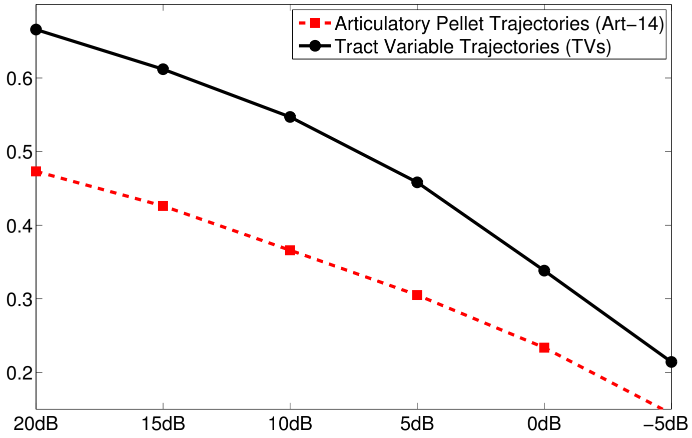


Figure 4.26 Average PPMC (relative to clean condition) of the estimated TVs and pellet trajectories (after Kalman smoothing) for Auora-2 (natural speech) at different SNRs for subway noise

It can be observed from Fig. 4.26 that the TV estimates offer a higher average relative PPMC at all noise levels compared to the pellet-trajectory estimates, indicating the relative noise-robustness of the TVs.

4.3.2.3 Noise robustness in word recognition using estimated TVs

In this section, we present the ASR experiments using the estimated TVs as inputs, to examine if they can improve the ASR noise-robustness. We employed the HTK-based speech recognizer distributed with the Aurora-2 (Hirsch & Pearce, 2000; Pearce & Hirsch, 2000), which uses eleven whole word HMMs with three mixture components per state and two pause models for 'sil' and 'sp' with six mixture components per state. The ASR experiment was based on training in clean condition and testing on multi-SNR noisy data. The following subsections report ASR results obtained from using the estimated TVs in different input conditions.

Use of TVs and their contextual information in ASR

We first examined if variants of TVs, or their Δs[4] can improve ASR performance, and tested four different feature vectors[10] as ASR inputs: (a) TVs (b) TVs and their velocity coefficients (TV+Δ)[11], (c) TVs and their velocity and acceleration coefficients (TV+Δ+Δ^2) and (d) TVs and their velocity, acceleration and jerk coefficients (TV+Δ+Δ^2+Δ^3). Figure 4.27 shows their word recognition accuracies along with a baseline defined by using the MFCC feature vector[12].

[10] The dimension of TV and each of its Δs is 8.

[11] Δ, Δ^2 and Δ^3 represent the first, second, and third derivatives, respectively.

[12] The dimension of MFCC feature vector is 39: 12 MFCC + energy, 13 Δ and 13 Δ^2.

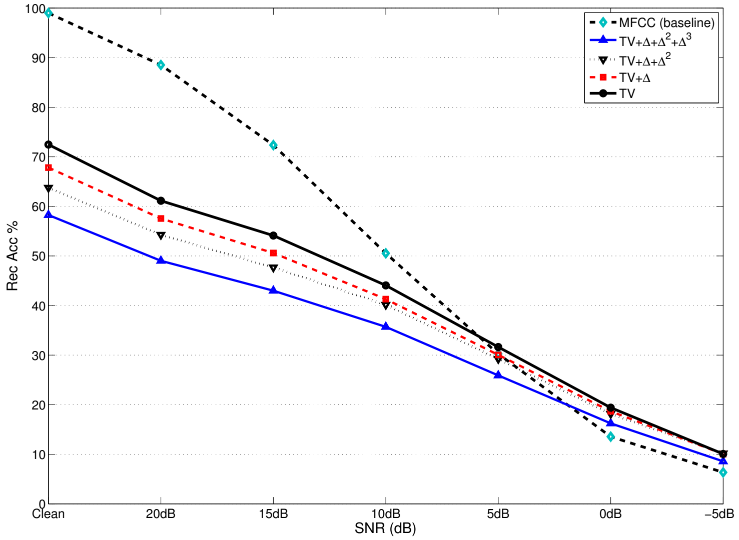


Figure 4.27 Average word recognition accuracy (averaged across all the noise types) for the baseline and TVs with different Δs

The recognition accuracy from using TVs and/or their Δs in the clean condition is much below the baseline recognition rate, which indicates that TVs and their Δs by themselves may not be sufficient for word recognition. However at 0dB and -5dB, TVs and their Δs offered better accuracy over MFCCs (significance was confirmed at the 1% level[13], using the significance-testing procedure described by Gillick & Cox (1989). Our observation for the clean condition is consistent with Frankel *et al.*'s observation (2000, 2001) that using estimated articulatory information by itself resulted in much lower recognition accuracy as compared to acoustic features. We also observed that TVs' contextual information (their Δs) in conjunction with TVs did not show better accuracies than TVs alone (at the 5% significance level[13]). This may be because the TV-estimator already uses a large contextualized (context window of 170ms) acoustic observation (as specified in section 4.1.3) as the input; hence, the estimated TVs by themselves should contain sufficient contextual information and further contextualization may be redundant.

[13] The detailed significance test results are shown in Appendix B.

TVs in conjunction with the MFCCs

Frankel *et al.* (2000, 2001) noticed a significant improvement in recognition accuracy when the estimated articulatory data was used in conjunction with the cepstral features, which we also have observed (Mitra *et al.*, 2009c). We used the MFCCs along with the estimated TVs for the ASR experiments. Here we considered three different models by varying the number of word (digit) mixture components per state from 2 to 4, identified as "Model-2mix", "Model-3mix" and "Model-4mix", where "Model-3mix" is the baseline model distributed with Aurora-2. Figure 4.28 compares the recognition accuracy[14] of MFCC+TV from the different word models to the baseline accuracy using MFCC only. Adding TVs to MFCCs resulted in significant improvement in the word recognition accuracy compared to the baseline system using MFCCs only. The improvement is observed at all noise levels for all noise types. Note the baseline here is the result from the Model-3mix[15], which showed the best performance amongst the models using MFCC+TV as shown in Figure 4.28. Also in Figure 4.28 we show the performance of the 14 flesh-point pellet trajectories (Art-14) when used in addition to the MFCCs, where the back-end uses 3-mixture components per state. Figure 4.28 clearly shows the superiority of TVs over Art-14 for improving the noise-robustness of a word-recognizer. Although Art-14 is found to improve the noise robustness over the MFCC baseline, it fails to perform as well as the TVs.

[14] The recognition accuracy here is averaged across all the noise types.

[15] We used this model for the rest of this dissertation.

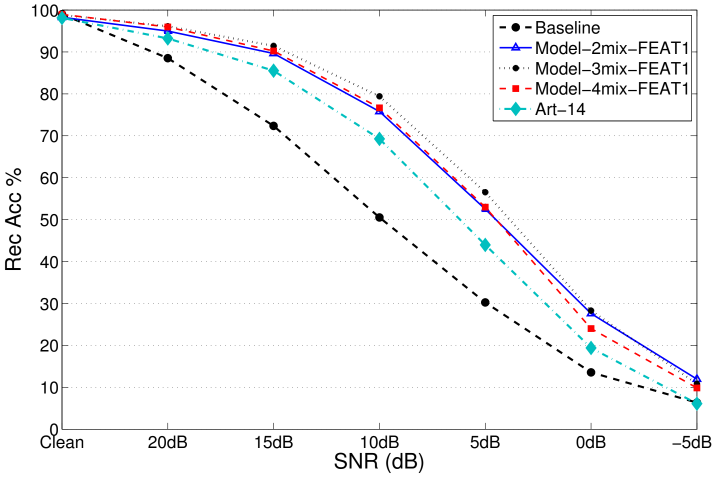


Figure 4.28 Average word recognition accuracy (averaged across all the noise types) for the baseline, MFCC+TV using the three different number of Gaussian mixture components per state, and MFCC+Art14 using a 3 Gaussian mixture component per state model

Speech enhancement

This section examines how speech enhancement will interact with the use of TV estimates and MFCCs. We used the preprocessor based MPO-APP[16] speech-enhancement architecture described in (Mitra *et al.*, 2009d) to enhance the noisy speech signal from Aurora-2. Four different combinations of MFCC and TV estimates were obtained depending upon whether or not their input speech was enhanced[17]. Figure 4.29 presents the average word recognition accuracies obtained from these four different feature sets. Similar to the results in the last section, we notice that articulatory information (in the form of TVs) can increase the noise robustness of a word recognition system when used with the baseline-MFCC features.

Indeed, TV estimates from enhanced speech exhibited poorer performance than TVs from noisy speech. This can be due to the fact that the MPO-APP based speech enhancer

[16] MPO: Modified Phase Opponency and APP: Aperiodic-Periodic and Pitch detector. The MPO-APP (Deshmukh *et al.*, 2007) speech enhancement architecture was motivated by perceptual experiments.

[17]The $\text{MFCC}_{\text{MPO-APP}}$ and the $\text{TV}_{\text{MPO-APP}}$ are the MFCCs and TVs that were obtained after performing MPO-APP enhancement of the speech signal.

(Deshmukh *et al.*, 2007) models speech as a constellation of narrow-band regions, retaining only the harmonic regions while attenuating the rest. The voiceless consonants (which are typically wideband regions) are most likely to be attenuated as a result of MPO-APP enhancement of speech. Given the attenuation of unvoiced regions in the enhanced speech, the TV-estimator may have difficulty in detecting the TVs properly at unvoiced consonant regions.

In Figure 4.29, the best accuracy is found in MFCC+TV from clean condition to 15dB, and $\text{MFCC}_{\text{MPO-APP}}$+TV from 10dB to -5dB. Such a system can be realized by using the preprocessor-based MPO-APP architecture prior to generating the baseline MFCC features only for SNRs lower than 15dB, which is named as [$(\text{MFCC+TV})_{\text{SNR}\geq 15\text{dB}}$ + $(\text{MFCC}_{\text{MPO-APP}}\text{+TV})_{\text{SNR}<15\text{dB}}$] feature set (Mitra *et al.*, to appear). Note the preprocessor-based MPO-APP (Mitra *et al.*, 2009d) has an inbuilt SNR-estimator in its preprocessing module which has been used to perform speech enhancement only if the detected SNR is < 15dB. Figure 4.30 compares [$(\text{MFCC+TV})_{\text{SNR}\geq 15\text{dB}}$ + $(\text{MFCC}_{\text{MPO-APP}}\text{+TV})_{\text{SNR}<15\text{dB}}$] with recognition rates from other referential methods that does not use TVs: $\text{MFCC}_{\text{MPO-APP}}$ (MFCCs after MPO-APP enhancement of speech) and $\text{MFCC}_{\text{LMMSE}}$ (MFCCs after the Log-spectral amplitude Minimum Mean Square Estimator (LMMSE) based speech enhancer (Ephraim & Malah, 1985). The use of articulatory information (in the form of the eight TVs) in addition to MFCCs resulted in superior performance as compared to using speech enhancement alone ($\text{MFCC}_{\text{MPO-APP}}$ and $\text{MFCC}_{\text{LMMSE}}$). This shows the strong potential of the articulatory features for improving ASR noise robustness.

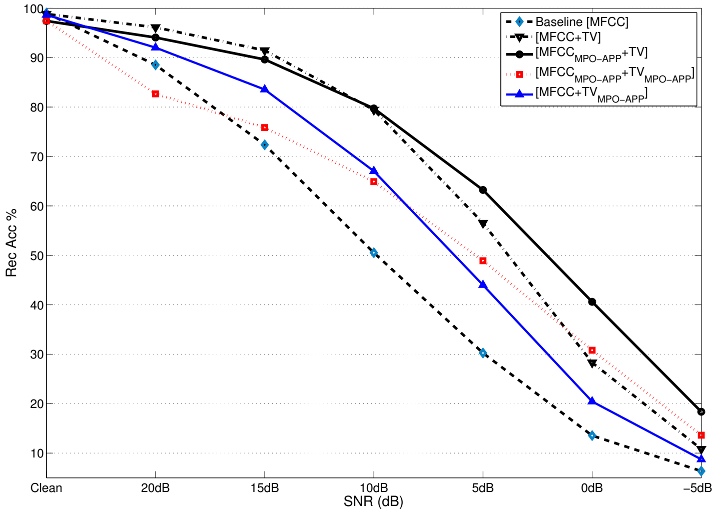


Figure 4.29 Average word recognition accuracy (averaged across all the noise types) for the four different combinations of MFCCs and TVs

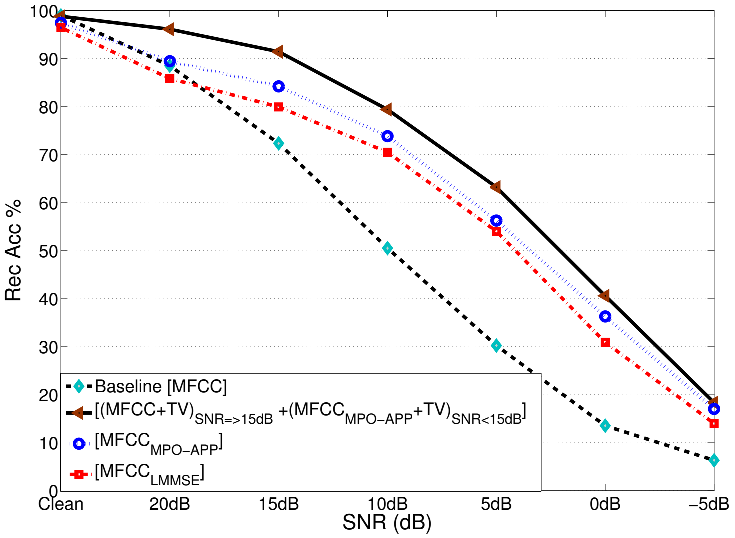


Figure 4.30 Average word recognition accuracy (averaged across all the noise types) for the (a) baseline (MFCC), (b) system using $\{[\text{MFCC+TV}]_{\text{SNR}\geq 15\text{dB}} + [\text{MFCC}_{\text{MPO-APP}}\text{+TV}]_{\text{SNR}<15\text{dB}}\}$, system using the (c) preprocessor based MPO-APP and (d) LMMSE based speech enhancement prior to computing the MFCC features (MFCC) (Mitra *et al.*, to appear)

Use of TVs with different front-end processing and feature sets for ASR

Previously we observed that TVs in word recognition task help to increase the accuracy when they are used in conjunction with the MFCCs. This section examines whether the advantage of using TVs holds for other feature sets (RASTAPLP) and front-end processing (MVA and ESTI).

RelAtive SpecTrA (RASTA) (Hermansky & Morgan, 1994) is a technique that performs low-pass filtering in the log-spectral domain to remove the slowly varying environmental variations and fast varying artifacts. We employed RASTAPLPs as an acoustic feature set instead of MFCCs for the Aurora-2 word recognition task. Similar to our previous observation, we noticed that use of TVs in addition to RASTAPLP exhibited a better accuracy than either TVs or RASTAPLPs alone.

Mean subtraction, Variance normalization and ARMA filtering (MVA) post-processing has been proposed by Chen & Bilmes (2007), which have shown significant error rate reduction for the Aurora-2 noisy word recognition task, when directly applied in the feature domain. We applied MVA to both MFCC and RASTAPLP and used them along with TVs as inputs for the word recognition task.

The ETSI front-ends have been proposed for the Distributed Speech Recognition (DSR). We have considered two versions of the ETSI front-end, the ETSI basic (ETSI ES 201 108 Ver. 1.1.3, 2003) and the ETSI advanced (ETSI ES 202 050 Ver. 1.1.5, 2007). Both the basic and the advanced front-ends use MFCCs, where the speech is sampled at 8 kHz, analyzed in blocks of 200 samples with an overlap of 60% and uses a Hamming window for computing the FFT.

Figure 4.31 compares the overall recognition accuracies from six different front-ends: (1) MFCC, (2) RASTAPLP, (3) MFCC through MVA (MVA-MFCC), (4) RASTAPLP through MVA (MVA-RASTAPLP), (5) ETSI-basic and (6) ETSI-advanced. All these conditions are further separated into cases with and without TVs. The positive effect of using

TVs was consistently observed in most of the noisy scenarios of MFCC, RASTAPLP, MVA-RASTAPLP and ETSI-basic but not in MVA-MFCC and ESTI-advanced. Note, that the TV-estimator being trained with synthetic speech does not generate highly accurate TV estimates when deployed on natural speech. The ETSI-advanced and the MVA-MFCC front-ends show substantial noise robustness by themselves; hence the inaccuracy in the TV estimates factors in more and hence fails to show any further improvement in their performance.

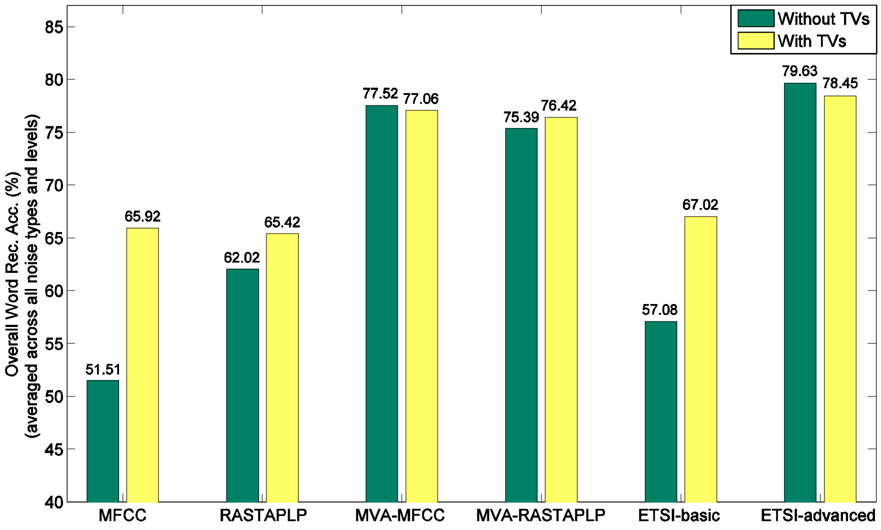


Figure 4.31 Overall word recognition accuracy (averaged across all noise types and levels) for the different feature sets and front-ends with and without TVs

Use of recognized gestures along with the TVs for ASR

In this experiment we used the estimated TVs and recognized gestures along with the acoustic features to perform word recognition on the Aurora-2 corpus. Note that the gesture-recognizer models were retrained using the AUR-SYN database and were then used to recognize the gestural scores for the natural utterances of Aurora-2. As before, training was performed on clean data and testing with noisy utterances. The recognized gestural scores were converted to gestural pattern vectors or GPVs (Zhuang *et al.*, 2009) for use as input to the word recognizer. The acoustic signal in the Aurora-2 was parameterized to feature coefficients (MFCC or RASTAPLP [Hermansky & Morgan, 1994]), using a 25 ms window and a 10ms frame-advance. Since the GPVs had originally been sampled at 5 ms, they had to be resampled for seamless concatenation with the acoustic features. We explored different combinations of GPVs, TVs and acoustic features (MFCC or RASTAPLP), and also each of them singly as possible inputs to the word recognition system. The number of Gaussian mixtures in the HMM whole word states was optimized for each input feature set using the dev-set[18] of Aurora-2 as the development set. It was observed that for the case when input features were concatenations of acoustic features with the TVs and GPVs (i.e., MFCC+TV+GPV and RASTAPLP+TV+GPV) the optimal number was 5 for word mixes, and 8 for 'silence/speech-pause' mixes. For all other input scenarios, the optimal number was 3 for word mixes and 6 for 'silence/speech-pause' mixes.

Table 4.12 presents word recognition accuracies obtained using MFCC and RASTAPLP with and without TVs and GPVs as inputs to the word recognizer. The last two rows show the recognition accuracy when only TVs or GPVs were used as the input to the word recognizer. The estimated TVs and GPVs are found to help improve the noise robustness of the word recognition system when used in conjunction with the acoustic features. However, the estimated TVs and GPVs by themselves were not sufficient for word recognition, which indicate that the acoustic features (MFCC/RASTAPLP) and the articulatory information (TVs & GPVs) are providing complementary information; hence neither of them alone offers results as good when used together. Note also that recognition accuracies of the GPVs were better than that of TVs, implying that the GPVs are better sub-word level representations than TVs. The main factor behind the GPVs' failure to perform as well as the acoustic features for the clean condition is most likely the inaccuracy of the gesture-recognizers and TV estimator. These models were trained with only 960 synthetic

[18] Note that, since the dev-set was used here to optimize the number of states per word, hence the corresponding 200 utterances from the test set were not used to evaluate the performance of the word-recognizer.

utterances (AUR-SYN) which is roughly only 11% of the entire Aurora-2 training set (consisting of 8440 utterances). Moreover as the models were trained on synthetic speech and executed on natural speech, the recognized gestures and the estimated TVs both suffer from acoustic mismatch. However Table 4.12 is encouraging in the sense that even with such inherent inaccuracies, the estimated TVs and the GPVs, when used with the acoustic features, provided improvement in word recognition performance. Figure 4.32 presents the overall word recognition accuracy (averaged across all noise types at all SNRs) when the acoustic features (MFCC & RASTAPLP) are used with and without TVs and the GPVs. Figure 4.33 shows the word recognition accuracy (averaged across all noise types) for 6 different SNRs using MFCCs and RASTAPLPs as the acoustic features with and without the estimated TVs and GPVs. We have added here the word recognition accuracy obtained from using generalized spectral subtraction (GSS) speech enhancement (Virag, 1999), which shows better accuracy over only the MFCCs. Using the estimated TVs and GPVs with the acoustic features (without any speech enhancement) is found to result in higher recognition accuracy than that obtained from using GSS speech enhancement, indicating that the use of articulatory information provided overall better noise-robustness than a traditional speech enhancement architecture.

Table 4.12 Overall Word Recognition accuracy

	Clean	0-20dB	-5dB
MFCC	99.00	51.04	6.35
MFCC+TV	98.82	70.37	10.82
MFCC+TV+GPV	98.56	73.49	16.36
RASTAPLP	99.01	63.03	10.21
RASTAPLP+TV	98.96	68.21	12.56
RASTAPLP+TV+GPV	98.66	75.47	19.88
TV	72.47	42.07	10.06
GPV	82.80	47.50	9.48

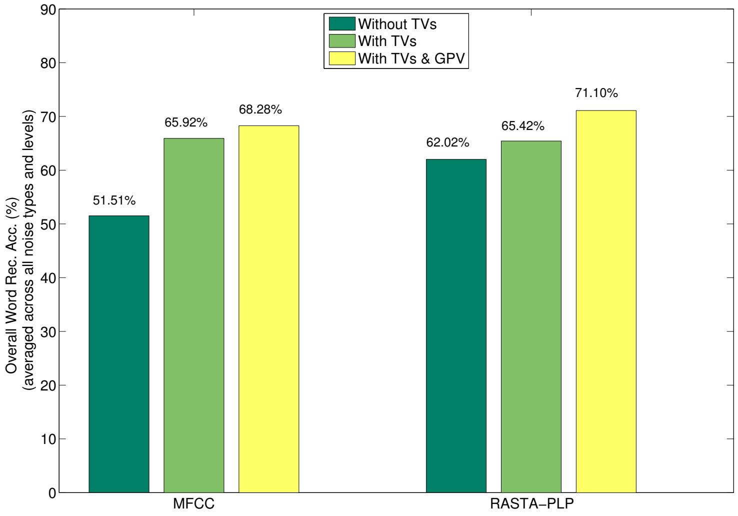


Figure 4.32 Overall word recognition accuracy using MFCC and RASTAPLP with and without the estimated TVs and gestures

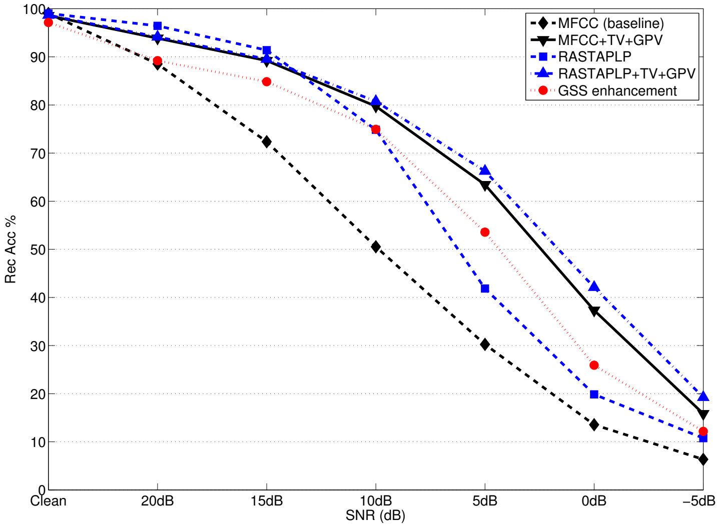


Figure 4.33 Word recognition accuracy (averaged across all noise types) at various SNR in using (a) the baseline MFCC (b) MFCC+TV+GPV, (c) RASTAPLP (b) RASTAPLP+TV+GPV and (d) MFCCs after GSS based speech enhancement of the noisy speech

4.3.3 ASR experiments: Observations

This section investigated the possibility of using TVs and gestures as possible inputs to a speech recognition system in noisy conditions. At the beginning we evaluated how accurately articulatory information (in the form of TVs) can be estimated from noisy speech at different SNRs using a feedforward neural network. The groundtruth TVs and gestural scores at present are only available for a synthetic dataset; hence both the TV-estimator and the gesture recognizer were trained with the synthetic data only. Using the synthetic data trained TV-estimator we evaluated the feasibility to estimate TVs for a natural speech dataset (Aurora-2), consisting of digits. We observed that the TV-estimator can perform reasonably well for natural speech. Secondly, we showed that the estimated TVs and the recognized gestural scores (in the form of GPVs) in conjunction to the baseline MFCC or RASTAPLP features can improve recognition rates appreciably for noisy speech. Such improvements in recognition accuracy, obtained by incorporating articulatory information in the form of TVs and GPVs, indicate that the acoustic features (MFCCs and RASTAPLPs) and the articulatory information (TVs & gestures) are providing partially complementary information about speech. Consequently, neither of them alone can provide accuracy as good as when both are used together, which is in line with the observation made by Kirchhoff (1999, 2002).

It is important to note that the TV-estimator and gesture recognizers presented in this section were not highly accurate, as they were trained with a significantly small number of data (960 utterances) than that available in the Aurora-2 training database (8440 training utterances). Also there exists a strong acoustic mismatch between the training (clean synthetic speech data) and testing (clean and noisy natural speech data) utterances for both the TV-estimator and the gesture recognizer models. Despite these differences, we were able to observe improvement in word recognition accuracies in the noisy cases of the Aurora-2 dataset for acoustic features: MFCCs, RASTAPLPs, which is encouraging. These observations indicate that with better models trained with a larger number of natural speech utterances, further improvement in the word recognition accuracies may be achieved. In order to train such models we require a natural speech database containing utterances with annotated TVs and gestural scores. Unfortunately no such database exists at present. Hence our logical next step was to create such a database on our own and generalize the results presented in this section, which were based primarily on synthetic speech data, to natural speech utterances. In the following section we present an automated approach to annotate a natural speech database with gestural scores and TVs.

Chapter 5: Annotation of Gestural scores and TVs for natural speech

Annotating a large natural speech database with gestural score specifications would not only benefit research in speech technology but also in various speech-related fields such as phonological theories, phonetic sciences, speech pathology, etc. Several efforts have been made to obtain gestural information from the speech signal. Atal (1983) proposed a temporal decomposition method for estimating gestural activation from the acoustic signal. Jung *et al.* (1996) also used the temporal decomposition method to retrieve gestural parameters such as constriction targets, assuming prior knowledge of articulator records. Sun *et al.* (2000) presented a semi-automatic annotation model of gestural scores that required manual gestural annotation to train the model. However, such an approach can potentially suffer from annotation errors due to incongruities among different annotators. Zhuang *et al.* (2008) and Mitra *et al.* (2010) showed that gestural activation intervals and dynamic parameters such as target and stiffness could be estimated from TVs using a TADA-generated synthetic database. Tepperman *et al.* (2009) used an HMM-based iterative bootstrapping method to estimate gestural scores but their approach was limited to a small dataset. Despite all these efforts, the fact remains that no natural speech database exists at present that contains gestural information.

Manually generating gestural annotations for natural speech is a difficult task. Compared to phone annotations, gestural onsets and offsets are not always aligned with acoustic landmarks. Further, articulatory gestures are constricting actions that are defined over finite time intervals and that do not unfold over time in a simple beads-on-a-string pattern; rather, they exhibit a great deal of spacio-temporal overlap, or coarticulation, with one another. While the ability of the gestural framework to naturally handle coarticulation is one of its major theoretical strengths, the task of identifying gestural onsets and offsets from the speech signal is an extremely difficult thing to do, using strictly hands-on manual annotation methods. Consequently, we were led to develop an automated procedure to perform gestural annotation for natural speech.

In this section, we described an iterative analysis-by-synthesis (ABS) landmark-based time-warping architecture (that we developed in collaboration with Haskins Laboratories [Nam *et al.*, 2010]) that can be used to generate gestural score and TV annotations from natural speech acoustic databases for which phone and word boundaries were provided in advance (e.g. Buckeye, TIMIT, Switchboard, etc.). We chose to begin the development of the ABS model using the XRMB database (Westbury, 1994) as it includes the time functions of flesh-point pellets tracked during speech production as well as the corresponding acoustics, which would allow us to cross-validate the articulatory information generated by our approach when applied to acoustics-only databases. The XRMB database includes speech utterances recorded from 47 different American English speakers (25 females and 22 males). Each speaker produced at most 56 types of speech reading tasks, e.g., reading a series of digits, sentences from the TIMIT corpus, or even an entire paragraph from a book. The sampling rate for the acoustic signals is 21.74 kHz. For our study, XRMB utterances were phone-delimited by using the Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008).

5.1 Architecture for Gestural annotation

Given the phone transcript of a natural speech utterance, S_{target}, from the XRMB database, TADA+HLsyn first generates a prototype-gestural score, G_{proto}, TV trajectories, and synthetic speech signal, S_{proto}. The phone content of S_{proto} and S_{target} will be identical because the pronunciation model of S_{target} is used as an input to TADA+HLsyn to create S_{proto}. Since S_{proto} is generated based on the model-driven intergestural timing, it substantially differs from

S_{target} both in rate of speech and individual phone durations. Our ABS procedure uses the mismatch between S_{target} and S_{proto} to iteratively adapt the gestural score, G_{proto}, for S_{proto} in order to make $S_{proto} \approx S_{target}$. For S_{proto}, the phone boundaries are approximated based on its underlying gestural on/offset times. The landmarks, or phone boundaries for S_{proto} are compared to those for S_{target} to measure how different they are in time, i.e. the time-warping scale, $W_{i=1}$, from which the 1st iteration begins. The time warping scale, W_1 is then applied to G_{proto}, generating G_1, which is the time-warped gestural score, and its corresponding acoustic output, S_1, is similar to the target natural speech, S_{target} in terms of pronunciation and individual phone durations. However due to possible errors in estimating phone boundaries from S_{proto}, the time warping might not be optimal. Thus, the phone boundaries for S_{proto} are piecewise modulated in steps of 10ms (to a maximum of ± 20 ms) to find an optimal warping scale. New time-warping scales $W_{i=2,3,4...}$ are obtained from each piecewise modulation and applied to $G_{i=1,2,3...}$, generating $G_{i=2,3,4...}$ and the corresponding speech output, $S_{i=2,3,4...}$. The output signals, $S_{i=2,3,4...}$ are then compared to the natural speech signal, S_{target} to compute the distance measure, $D(S_{target}, S_i)$ at each iteration step i. This procedure (piecewise phone boundary modulation and distance measure) is performed iteratively until $D(S_{target}, S_i)$ is minimized.

Obtaining the W_i at each iteration step i is the analysis part and applying W_i to G_i and consequently synthesizing S_i is the synthesis part in our ABS architecture. At each step i the warping function W_i is obtained by ensuring that the phonetic landmarks (phone onsets-offsets) are similar for $S_{,i}$ compared to S_{target}. Note that at each iteration a number of possible w_i can exist (based on different slope constraints on the warping function [Sakoe & Chiba, 1978]), but the one, W_i, which offers the minimum distance as shown in equation (41) is selected:

$$W_i = \arg\min\left[D(S_{t\arg et}, w_i[S_{i-1}])\right] \tag{41}$$

Hence, W_i (the optimal warping scale at the end of the i^{th} iteration) helps to make S_i more similar to S_{target} compared to $S_{i\text{-}1}$ and the degree of similarity is reflected by the distance measure D. Now, if G_{proto} is the gestural score that was used to generate the initial TADA-synthesized speech signal, S_{proto}, the ABS procedure iteratively creates a series of $W_{i=1\ldots j}$ and corresponding gestural scores, $G_{i=1\ldots j}$ at each step i, which successively minimizes $D(S_{target}, S_{proto})$. The procedure is halted after a given number of steps, N or earlier, when the value of $D(S_{target}, S_{proto})$ ceases to drop any further. If we assume that the procedure continues till j number of steps ($j \leq N$), then an overall warping function can be defined as -

$$W_{opt} = W_j[W_{j\text{-}1}\ [W_{j\text{-}2} \ldots[W_{i\ldots}\ [W_1]\ldots]\ldots]] \tag{42}$$

where W_{opt} is the nonlinear warping function that defines the optimal gestural score, $\hat{G}_{opt}$, i.e., the gestural score that generates the best synthetic estimate, which is defined as -

$$\hat{G}_{OPT}\left[S_{t\arg et}\right] = W_{OPT}\left(G\left[S_{proto}\right]\right) = W_{OPT}\left(G_{proto}\right) \tag{43}$$

The overall architecture of this ABS procedure is shown in Figure 5.1, where the time warping block represents a time-warping procedure different from those typically used in traditional dynamic time warping (DTW) algorithms (Rabiner *et al.*, 1991). We will show in the next section that our iterative ABS warping approach helps to reduce the distance measure $D(S_{target}, S_{proto})$ more effectively than the traditional DTW algorithms.

Figure 5.2 compares the XRMB (top panel), prototype TADA (middle panel), and time-warped TADA (bottom panel) utterances for the word "seven" from task003 of XRMB speaker 11, in which each panel shows the corresponding waveform and spectrogram. Figure 5.2 (middle and bottom panels) also displays the gestural scores for the prototype and time-warped TADA utterances (with lips, tongue tip [TT], and tongue body [TB] gestures as gray blocks overlaid on the spectrogram), showing how gestural timing is modulated by the proposed time-warping procedure. Time warping is performed on a word-by-word basis. The obtained word-level gestural scores are seamlessly concatenated to yield the utterance-level gestural score such that the final phone's offset of one utterance is aligned to the initial phone's onset of the following utterance, which can involve gestural overlap. TADA is executed on the utterance-level gestural scores to generate the corresponding TVs.

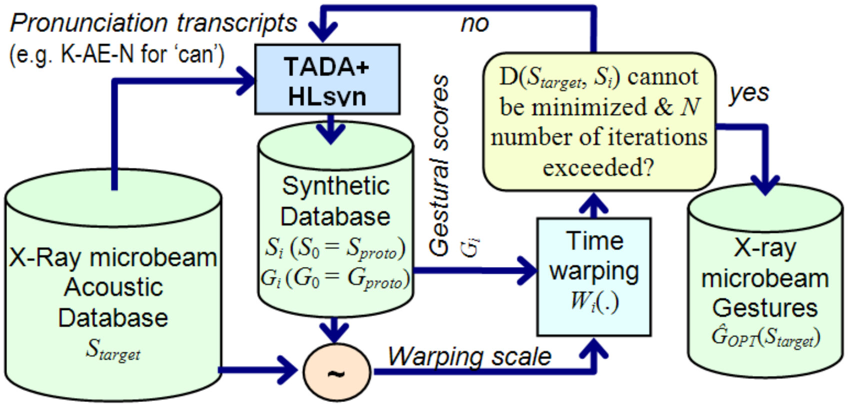

Pronunciation transcripts
(e.g. K-AE-N for 'can')
no
TADA+ HLsyn
D(S_{target}, S_i) cannot be minimized & N number of iterations exceeded?
yes
Gestural scores
G_i
Synthetic Database
S_i ($S_0 = S_{proto}$)
G_i ($G_0 = G_{proto}$)
X-Ray microbeam Acoustic Database
S_{target}
Time warping
$W_i(.)$
X-ray microbeam Gestures
$\hat{G}_{OPT}(S_{target})$
~
Warping scale

Figure 5.1 Block diagram of the overall iterative ABS warping architecture for gesture specification

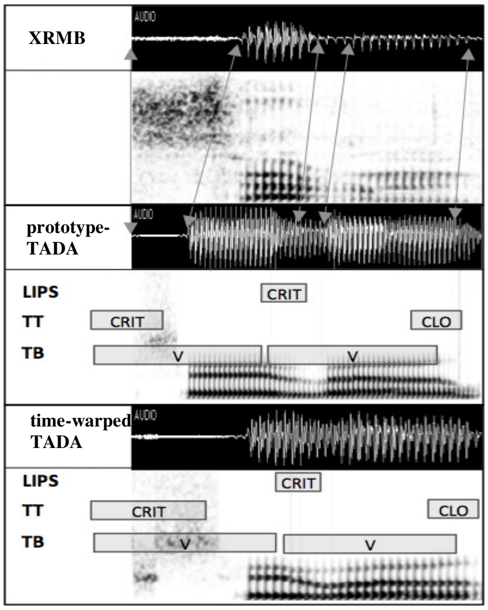


Figure 5.2 Waveform and spectrogram of XRMB, prototype TADA, and time-warped TADA speech for 'seven' (borrowed from Nam *et al.* [2010])

Note that the above approach is independent of any articulatory information from XRMB. Based on word and phone transcriptions, the architecture generates gestural scores and TV trajectories using the default speaker characteristics predefined in TADA. This is ideal for speech recognition as almost all speaker-specific attributes are normalized out of the gestural scores generated by the ABS procedure.

5.2 Analysis of the annotated gestures

We have implemented the proposed landmark-based ABS time-warping architecture for gestural score annotation across all the 56 tasks from 47 speakers of the XRMB database (however, some speakers performed only a subset of the 56 tasks). Figure 5.3 shows the annotated gestures and TVs for a snippet taken from task003 of speaker # 11. The top two panels in Figure 5.3 show the waveform and spectrogram of the utterance "eight four nine five"; the lower eight panels show each gesture's activation time functions (as rectangular blocks) and their corresponding TV trajectories (smooth curves), obtained from our proposed annotation method.

We performed two tasks to evaluate our methodology. First, we compared the proposed time-warping strategy with respect to the standard DTW (Sakoe & Chiba, 1978) method. To compare the effectiveness of those two warping approaches, we used an acoustic distance measure between the XRMB natural speech, S_{target} and the TADA speech (i) after DTW only vs. (ii) our iterative landmark-based ABS time-warping method. We used three distance metrics (a) Log-Spectral Distance (D_{LSD}) as defined in (44), (b) Log-Spectral Distance using the Linear Prediction spectra ($D_{LSD\text{-}LP}$) and the (c) Itakura Distance (D_{ITD}). D_{LSD} is defined as

$$D_{LSD} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{S(\omega)}{\hat{S}(\omega)}\right]^2 d\omega} \tag{44}$$

where $S(\omega)$ and $\hat{S}(\omega)$ are the spectra of the two signals to be compared. For $D_{LSD\text{-}LP}$ the spectra

S(*ω*) and *Ŝ*(*ω*) are replaced with their respective LP spectra that were evaluated using a 25ms window with 15ms overlap. D_{ITD} is defined as

$$D_{ITD} = \ln\left[\frac{1}{2N}\left(\sum_{\omega=-N}^{N}\frac{P(\omega)}{\hat{P}(\omega)}\right)\right] - \frac{1}{2N}\left[\sum_{\omega=-N}^{N}\ln\left(\frac{P(\omega)}{\hat{P}(\omega)}\right)\right] \quad (45)$$

$$where\ 0 \le \omega \le \pi$$

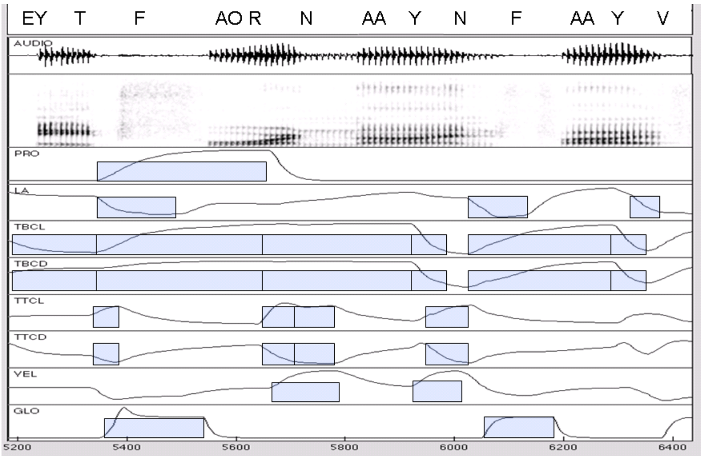


Figure 5.3 Annotated gestures (gestural scores) and TVs for a snippet from an utterance from task003 in XRMB

Twelve different tasks (available from all speakers) were selected randomly from the XRMB database to obtain the distance measure between the natural and synthetic speech. Table 5.1 presents the average distances obtained from using DTW and our proposed iterative time-warping approach.

Table 5.1 Distance measures between the warped signal and the XRMB signal from using (i) DTW and (ii) proposed landmark-based iterative ABS time-warping strategy

	D_{LSD}	D_{LSD-LP}	D_{ITD}
DTW	3.112	2.797	4.213
Iterative warping	2.281	2.003	3.834

Secondly, we evaluated how similar the TV trajectories generated from our proposed approach are compared to those derived from the recorded flesh-point measurements available in the XRMB database. We describe below how the TVs are estimated from the pellet information. LA can be readily estimated as a vertical distance between upper (UL_y) and lower lip (LL_y) pellets in XRMB, as shown by (46)

$$LA = UL_y \sim LL_y \qquad (46)$$

The tongue-associated TVs (TBCL, TBCD, TTCL, TTCD) however involve more complex procedures to be estimated from pellets. They are measures based on a polar coordinate with reference to its origin (F in Figure 1.4). For the polar coordinate, we translated the XRMB coordinate system[19] so that the origin is moved -32 mm on the x-axis and -22 mm on the y-axis. TTCL is an angular measure of T1 with respect to the coordinate origin, F, and TTCD is the minimal distance from T1 to the palate trace. For TBCL and TBCD, a circle was estimated for the tongue body such that it passes through T3 and T4 with a fixed radius[20]. TBCL was estimated as an angle of a line connecting the tongue body circle's center (C in Figure 1.4) and the coordinate origin. To measure TBCD, it is necessary to recover the missing information between the palate trace and the pharyngeal wall. The palate trace was extended backward by obtaining the convex hull of the tongue pellet data cloud and the remaining gap to the pharyngeal wall was linearly interpolated. TBCD was estimated as the shortest distance from the tongue body circle to the hard structure. Note that GLO and VEL were excluded from the evaluation because XRMB does not contain any corresponding flesh-point data.

Once the TV trajectories are derived from the recorded flesh-point data of XRMB, their correlation with the TVs generated from the annotated gestures are computed. For obtaining the correlation measure, we have used the PPMC score (defined in (4)) between the

[19] The XRMB coordinate system is defined at the tip of the maxillary incisors on the x-axis as the maxillary occlusal plane.

[20] We used a tongue body circle of 20 mm radius, which is for a default speaker in TADA.

annotated and the XRMB derived TV trajectories. The correlation analysis was limited to consonants because they exhibit more critical constriction than vowels in the vocal tract. Each phone is associated with a set of gestures, which are activated at the corresponding TVs. The correlation measure was performed during the activation interval of each phone's primary gesture(s) (e.g. tongue tip gesture for /t/ and /s/, lip gesture for /p/, /f/). Table 5.2 shows the correlations obtained between the annotated TVs and those derived from XRMB flesh-point data. It can be seen in Table 5.2, that the correlation scores are encouraging despite the errors and differences we can expect from (a) the gestural score and TV annotation procedure, (b) speaker differences[21], (c) lack of prosodic information[22] and finally (d) inaccuracies in the phone labeling of the forced aligner.

Table 5.2 Correlation[23] between the annotated TVs and the TVs derived from the measured flesh-point information of XRMB database

TVs	Correlation (*r*)
LA	0.715
TTCL	0.291
TTCD	0.596
TBCL	0.510
TBCD	0.579
Avg	0.538

The next task we performed is to evaluate how effective the obtained gestures are for speech recognition. We selected 1692 utterances from the XRMB dataset for training and 801 utterances for testing. The training set consisted of speaker 11 to 46 whereas the testing set consisted of speakers 48 to 63 (speaker 17, 22, 23 38, 47 and 50 did not exist in the XRMB

[21] The annotated XRMB gestural scores and TVs do not correspond to the actual speakers in the actual XRMB database but represent the default speaker model used in TADA.

[22] Prosodic information from the XRMB database has not been used during the annotation process.

[23] Note that LP is not included in the correlation result as LP is not used as the primary articulation distinguishing consonantal gestures

database that we used in our experiment). Table 5.3 gives detailed information about the training and the testing sets. For the word recognition experiments, we converted the sequence of overlapping gestures into an instantaneous "gestural pattern vector" (GPV) as proposed by Zhuang *et al.* (2009) as schematized in Figure 5.4.

From the XRMB training set we observed that altogether 1580 unique GPVs are possible, which indicates that theoretically 1580×1579 ≈ $2\text{x}10^6$ unique GPV bigram sequences are possible. However from the training set we observed that only 5876 unique GPV bigram sequences are observed in our dataset. Hence for the training and test set we created a 5876-dimensional GPV-bigram histogram for each word. Given a word, only a few GPV bigrams will be observed; hence the word dependent GPV-bigram histogram will be a predominantly sparse vector. To address that we interpolated the word-dependent GPV bigrams with the word-independent GPV bigrams (similar to [Zhuang *et al.*, 2009]) using a ratio 5000:1 and observed this ratio be optimal[24] in terms of the word error rates (WER).

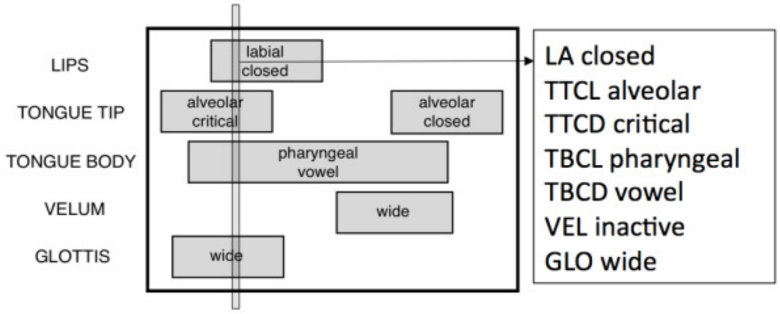


Figure 5.4 Gestural score for the word "span". Constriction organs are denoted on the left and the gray boxes at the center represent corresponding gestural activation intervals. A GPV is sliced at a given time point of the gestural score

To compare the performance of the gesture-based word recognizer with that of a phone-based one, we created phone bigram histograms for each word. We observed that

[24] The optimal ratio was obtained by using 90% of the training data to create word models and using the remaining 10% to obtain the word error rates. The ratio that generated the best WER was considered as optimal.

XRMB database contains 64 phones and 464 possible bigram sequences; hence each word in the training and test set was represented using a 464-dimensional phone-bigram histogram. The 464-dimensional phone bigram histogram can be expected to be sparse; however the sparsity should be less compared to the 5876-dimensional GPV bigram histogram. Like before we interpolated the word-dependent phone bigrams with the word-independent phone bigrams and observed (from using 10% of our training data as development set)[25] that the interpolation does not help in this case. Hence we did not perform any interpolation of the word-dependent phone bigrams.

We realized two different versions of the word recognizer using (1) Kullback-Leibler divergence (KLD) and (2) a three hidden layer neural network (ANN). For the KLD based approach, word level probability mass function (pmf_{word_train}, for word *word_train*, where *word_train* = 1:468, refer to Table 5.3) was created. For each word in the test set, the KLD between the *pmf*s, pmf_{word_train} and pmf_{word_test} was evaluated. The word model *word_train* that gave the least KLD was identified as the recognized word for *word_test*. KLD is defined as

$$KLD\left[pmf_{word\_test} \,||\, pmf_{word\_train}\right] = \sum_{i \in N} pmf_{word\_test,i} \log\left[\frac{pmf_{word\_test,i}}{pmf_{word\_train,i}}\right] \tag{47}$$

as $N \rightarrow \infty$ a link between the likelihood ratio (L) and KLD can be established (Cover & Thomas, 1991) as

$$KLD\left[pmf_{word\_test} \,||\, pmf_{word\_train}\right] = -\log_2(L) \tag{48}$$

which indicates that if pmf_{word_train} and pmf_{word_test} are identical, then L = 1 and $D_{KL} = 0$. Hence word recognition using KLD can be formulated as

$$Word = \underset{word\_train}{\arg\min}\, KLD\left[pmf_{word\_test} \,||\, pmf_{word\_train}\right] \tag{49}$$

[25] The optimal ratio was obtained by using 90% of the training data to create word models and using the remaining 10% to obtain the word error rates. The ratio that generated the best WER was considered as optimal.

Table 5.3 Details of the train & test data of XRMB

	Train	Test
Number of utterances	1692	801
Number of speakers	32	15
Total number of words	49672	23576
Number of unique words	468	388

For the 3-hidden layer NN approach, we used a simple feedforward network with tan-sigmoid activation function, having 400-600-400 neurons in the three hidden layers, trained with scaled-conjugate gradient. The WER obtained from the KLD and ANN based recognizers are shown in Table 5.4, where GPV-bigram histogram provides lower WER than phone-bigram histogram. Note that neither the phones nor the gestures are recognized or estimated from the speech signal; we have used the annotated information in both the cases. Hence the difference in their recognition accuracy reflects the strength of one representation over the other. Thus the results here indicate that GPV-bigrams provide more discreteness than the corresponding phone-bigrams, which was confirmed by examining the recognition error patterns. It is shown that phone representation suffered from pronunciation variability (for example it got confused with the 4 different pronunciations of 'when' [W-EH1-N, HH-W-EH1-N, HH-W-IH1-N, W-IH1-N] and wrongly recognized it as 'an' [AE1-N]), which was not observed for the GPV-bigrams.

Once we have realized a corpus with transcribed gestures we can obtain gestural score automatically (Mitra *et al.*, 2010b) from a given speech in way that preserves lexical information more robustly than does a derived phone string from the audio.

Table 5.4 WER (%) obtained for XRMB

	KLD	NN
GPV-bigram	2.48	8.31
Phone-bigram	6.48	9.36

5.3 Gestural annotation: Observations

In this section we presented a landmark based iterative ABS time-warping architecture that can annotate speech articulatory gestures potentially for any speech database containing word and phone transcriptions and their time alignment. The strength of this approach is that the articulatory information it generates is speaker independent, hence ideal for ASR applications. Word recognition experiments indicate that the gestures are a suitable unit-representation for speech recognition and can offer WER as low as 2.48% for a multi-speaker word recognition task. Given that we can now annotate gestural scores for natural speech, the next logical step is to realize a speech recognition architecture using such annotated natural speech database as the training corpus.

Chapter 6: Building a Gesture-based ASR using natural speech

In chapter 4 we showed that use of articulatory information in the form of TVs and gestures can potentially improve the performance of ASR systems under noisy conditions. Note that in those experiments the articulatory information was not provided as an additional modality, but estimated from the speech signal using models that were trained with synthetic speech. In chapter 5 we presented an approach that can be deployed on the utterances of any natural speech database to obtain its corresponding gestural score and TV annotation. Thus, chapter 5 paves the way to realize natural speech trained models for estimating articulatory information from speech. In this chapter, we present the hidden Gesture-based Dynamic Bayesian Network (G-DBN) framework as the final implementation of our gesture-based ASR for natural speech. In the system, we treat the articulatory gestures as hidden variables in which case no explicit recognition of the gestures is required. We obtained the gestural annotation for Aurora-2 clean training corpus using the methodology outlined in chapter 5. We have demonstrated in chapter 4 using the TADA synthetic speech that (1) the use of TVs in addition to the acoustic features helps to improve the noise-robustness of an ASR system and (2) the knowledge of the TVs help to improve the recognition rate of the articulatory gestures. In this chapter, we present a natural speech trained TV estimator to revalidate our claim made with synthetic speech in chapter 4 that TVs are superior to pellets as articulatory information, describe our gesture-based ASR (G-DBN) and discuss the results.

6.1 Speech Inversion: TVs versus Pellet trajectories

In this section, we aim to (a) present a TV estimation model trained with natural speech, (b) compare the estimation accuracies between TVs and pellet trajectories and (c) compare the TVs and pellet data according to (*i*) a statistical non-uniqueness measure of articulatory-acoustic mappings, and (*ii*) their relative performance in ASR experiments.

6.1.1 Experiments

The speech inversion models presented in this section were trained with the natural utterances of the XRMB database, which were annotated with gestural scores and TV trajectories using the procedure specified in chapter 5. The annotated data contains eight TV trajectories that define the location and degree of different constrictions in the vocal tract (see Table 1.1), where each TV trajectory is sampled at 200Hz. The XRMB data contains pellet trajectory (PT) data (sampled at 145.65Hz) recorded along with the speech waveforms (sampled at 21.74 kHz). The pellets were placed on the upper lip (UL_x & UL_y), lower lip (LL_x & LL_y), tongue tip ($T1_x$ & $T1_y$), mid-tongue ($T2_x$, $T2_y$, $T3_x$ & $T3_y$) and tongue dorsum ($T4_x$ & $T4_y$), where the subscripts x, y represent the horizontal and vertical coordinates of each pellet, resulting in 12 channels of flesh-point data.

Our work presented in this section uses the acoustic data, TVs and PTs for the 56 tasks performed by male speaker 12 from the XRMB database: 76.8% of the data was used for training, 10.7% for validation and the rest for testing. The PTs were upsampled to 200Hz to synchronize with the sampling rate of the TVs. The acoustic signals were downsampled to 16KHz and 8KHz[26] and parameterized as (a) MFCCs, (b) LPCC and (c) PLPCC. For each parameterization, 20 coefficients for 16KHz data and 13 coefficients for 8KHz data were selected that were analyzed at a frame rate of 5ms with analysis window duration of 10ms. The acoustic features and the articulatory data (PT and TV) were z-normalized. The resulting acoustic coefficients were scaled such that their dynamic range was confined within [-0.95, +0.95]. To incorporate dynamic information the acoustic features were temporally contextualized in all the experiments reported here. Specifically the acoustic coefficients were obtained from each of the nine 10ms-windows (middle window centered at the current time with preceding and following windows separated by 20ms intervals), thereby covering

[26] Sampling rate of 16KHz and 8KHz are used here as the commonly used ASR databases usually contain utterances sampled at these frequencies.

170ms of speech. This acoustic information was concatenated into a contextualized acoustic feature vector with a dimensionality of 180 (= 9×20) for 16KHz data and 117 (= 9×13) for 8KHz data.

The speech inversion models were trained as separate FF-ANNs one for each acoustic feature (MFCC, LPCC or PLPCC) and articulatory information (PTs or TVs) and sampling rate (16KHz or 8KHz) set, resulting in twelve individual models. The dimension of the output vectors were eight for the TVs and twelve for the PTs. All FF-ANNs were trained with backpropagation using a scaled conjugate gradient algorithm. The raw estimated trajectories from the FF-ANNs were smoothed using a Kalman smoother. The 3 hidden layer FF-ANNs with tan-sigmoid activation functions were implemented for each of the twelve inversion models. The optimal number of nodes in each hidden layer was obtained by maximizing the PPMC between the actual (groundtruth) and the estimated articulatory trajectories for the development set. Note that the groundtruth PTs were simply taken from the XRMB corpus whereas the groundtruth TVs were generated from the annotation process. We refrained from adding any additional hidden layer beyond the three because with increase in the number of hidden layers: (a) the error surface became more complex with a large number of spurious minima; (b) the training time as well as the network complexity increased; and (c) no appreciable improvement was observed. The ANNs were trained with a training epoch of 4000.

Table 6.1 presents the overall PPMC obtained by comparing the groundtruth and the estimated articulatory data averaged across all 12 channels for PT data and across 6 channels for TV data (note: GLO and VEL TVs are excluded for the comparison because there are no counterparts in the pellet data), which were obtained using each of the different acoustic parameterizations at each sampling rate. Overall the PPMC values for the estimated TVs were higher than that for the estimated PTs, indicating that TVs were estimated more accurately by the FF-ANNs. The PPMC of the TV estimates obtained from the three different acoustic parameterizations were quite similar to each other, indicating that accuracy of the TV estimation was somewhat independent of the particular set of acoustic parameters considered; such close similarity, however, was not as evident for the PTs. Table 6.2 compares the obtained PPMC values between individual TV and pellet estimates for 16KHz data. Taken together, the results in Tables 6.1 and 6.2 indicate that TVs can be estimated more accurately than PTs from the speech signal. Figure 6.1 shows the actual and estimated TVs (LA, LP, TBCD & TTCD) for utterance the "across the street" obtained from the 3-hidden layer FF-ANN TV-estimator using 8KHz speech data with MFCC as the signal parameterization.

Table 6.1 PPMC averaged across all trajectories for TV and Pellet data using different acoustic parameterization of 8KHz and 16KHz speech. The numbers in the parentheses denote the number of neurons used in each of the 3 hidden layers

		MFCC	PLPCC	LPCC
16KHz	TV trajectory	0.828 (250-150-225)	0.825 (175-100-125)	0.827 (150-100-225)
	Pellet trajectory	0.780 (250-125-75)	0.774 (200-75-150)	0.734 (150-125-225)
8KHz	TV trajectory	0.832 (225-150-225)	0.821 (250-175-125)	0.820 (200-75-175)
	Pellet trajectory	0.778 (250-150-200)	0.767 (275-150-150)	0.762 (175-75-200)

Table 6.2 Comparison of PPMC between relevant articulatory pellet and TV data using MFCC as the acoustic parameterization

TVs	*PPMC*	Pellets	*PPMC*
LP	0.852	LL_x	0.822
		UL_x	0.773
LA	0.786	LL_y	0.844
		UL_y	0.676
TTCL	0.814	$T1_y$	0.903
		$T1_x$	0.887
TTCD	0.794	$T2_y$	0.918
		$T2_x$	0.883
TBCL	0.838	$T3_y$	0.775
		$T3_x$	0.491
TBCD	0.831	$T4_y$	0.706
		$T4_x$	0.422
Avg	0.819	*Avg*	0.758

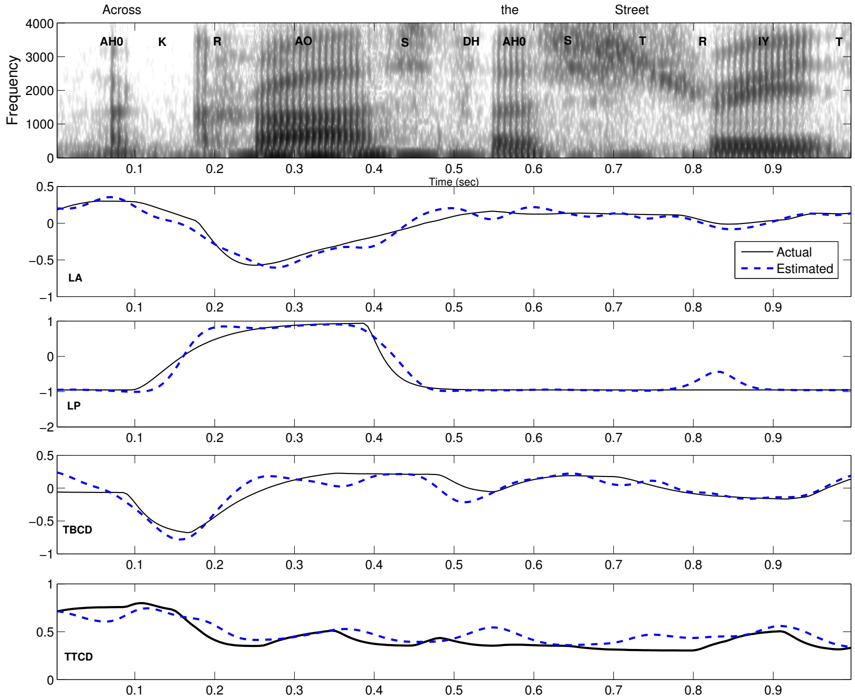


Figure 6.1 Plot of the actual and estimated TVs (LA, LP, TBCD & TTCD) for natural utterance “across the street” taken from the XRMB database

As stated earlier, since TVs are relative measure, they can be expected to suffer less from non-uniqueness than PTs (McGowan, 1994), which may be the reason why the TVs are estimated more accurately than the PTs. To analyze and quantify non-uniqueness in the speech inversion models using TVs and PTs, we performed a statistical analysis motivated by the work performed by Ananthakrishnan *et al.* (2009). In this analysis, the conditional probability function of the inversion, p(a|s) is first estimated, where a is the articulatory configuration and s is the acoustic vector at any given time instant. We used a Mixture Density Network (MDN) (instead of the Gaussian Mixture Model (GMM) used by Ananthakrishnan *et al.* (2009)) to estimate p(a|s) from acoustic and articulatory data in each phone context (see section 4.1.2.5 for a brief overview on MDN).

According to Ananthakrishnan *et al.* (2009), non-uniqueness in speech inversion exists when the conditional probability function $p(a|s)$ exhibits more than one probable articulatory configuration for a given acoustic observation. In such a case, the degree of non-uniqueness in the inverse mapping can be quantified using the deviations of the peaks of the conditional probability function *p*(*a*|*s*) from the mean peak location. We have used the unit-less measure proposed in (Ananthakrishnan *et al.*, 2009), the Normalized Non-Uniqueness (*NNU*$_t$) measure defined as

$$NNU_t = \sqrt{\sum_{q=1}^{Q} P_q (M_q - \mu_t)^T (\Sigma_t)^{-1} (M_q - \mu_t)}$$
$$P_q = \frac{p_{a|s}(a = M_q \mid s_t)}{\sum_{q=1}^{Q} p_{a|s}(a = M_q \mid s_t)} \qquad (50)$$

where Q is the number of local maxima (or the peaks) at locations M_q ($1 \leq q \leq Q$), P_q is the normalized probability, μ_t is the mean location of the peaks and Σ_t is the variance of the conditional probability function. Since *NNU* provides a measure of the spread of the local peaks in the conditional pdf, *p*(*a*|*s*), a higher *NNU* indicates a higher degree of non-uniqueness in the mapping.

Since (Ananthakrishnan *et al.*, 2009; Neiberg *et al.*, 2008; Qin *et al.*, 2007) showed that non-uniqueness is commonly observed for consonants, we have selected six consonants (/r/, /l/, /p/, /k/, /g/ and /t/) that these studies have shown to be mostly affected by non-uniqueness. A single MDN with 100 hidden layers and 16 mixture components with spherical Gaussian mixtures was trained for 2500 iterations for each articulatory channel in each phone context, where the acoustic observations were parameterized as contextualized MFCCs. We computed the Normalized Non-uniqueness (*NNU*) measure for the data in the testing set. As shown in Figure 6.2, the *NNU* score of TVs is almost always lower than that of the PTs, indicating that the inverse mapping between acoustics and TVs is less non-unique compared to that between acoustics and PTs. Please note here that the result shown in Figure 6.2 is for 16KHz data. We have also obtained the *NNU* scores for 8KHz data, where the overall *NNU* scores are found to be slightly higher than those for 16KHz data, indicating that lowering the sampling rate increases non-uniqueness in the inverse mapping.

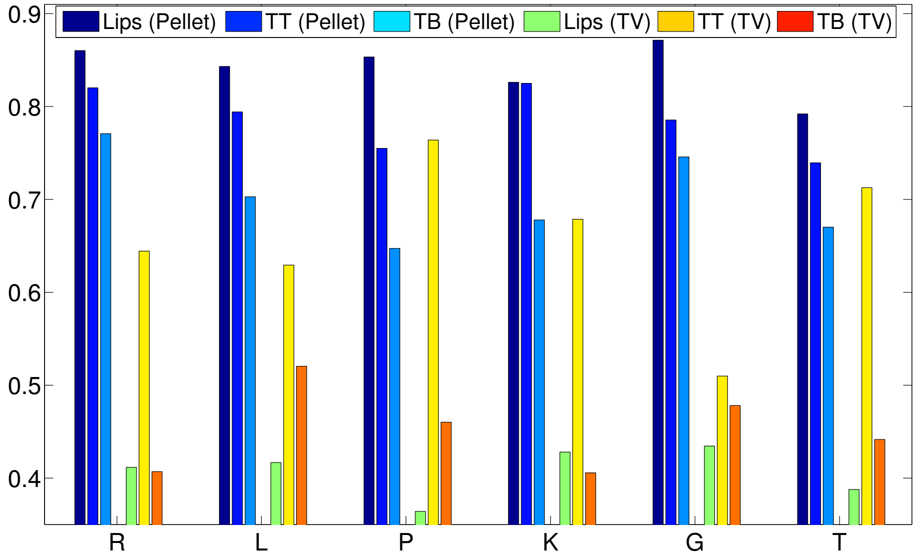


Figure 6.2 Graph comparing the Normalized Non-uniqueness measure (NNU) for speaker 12 in XRMB database across 6 different phonemes (/r/, /l/, /p/, /k/, /g/ & /t/) for Lips, Tongue-Tip (TT) and Tongue-Body (TB) pellet-trajectories and TVs

Finally, we evaluated the relative utility of TVs and PTs in a simple word recognition task using the Aurora-2 corpus. This recognizer incorporates a hidden Markov model (HMM) backend that uses eleven whole word HMMs, each with 16 states (in addition to 2 dummy states) with each state having three Gaussian mixture components. Two pause models, one for silence (‘sil’) and another for speech-pause (‘sp’) were used; the ‘sil’ model has three states and each having six mixtures, while the ‘sp’ model has only a single state with three mixtures. Training in the clean condition and testing in the noisy scenario is used for this experiment. The HMMs were trained with three different observation sets (a) MFCCs, (b) MFCCs + estimated TVs, (c) MFCCs + estimated PTs. Note that the sampling rate for the Aurora-2 database is 8KHz; hence, the 8KHz version of the TV estimator and the PT estimator had to be used.

Figure 6.3 compares the word recognition accuracy obtained from the word recognition experiments using the Aurora-2 database, where the accuracies at a given SNR are averaged across all the noise types. Adding the estimated TVs or the PTs to the MFCCs improved the word recognition accuracy compared to the system using MFCCs only. However, the improvement is higher for TVs compared to the PTs, which further highlights the strength of TVs.

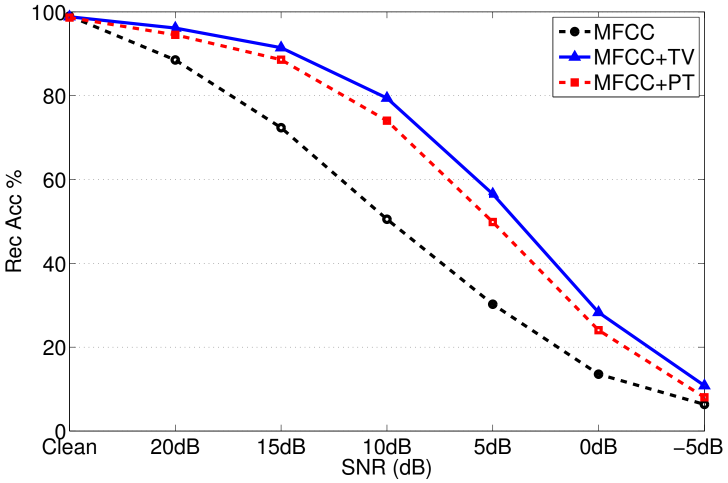


Figure 6.3 Average word recognition accuracy (averaged across all the noise types) for MFCC only, MFCC+TV and MFCC+PT

6.1.2 Observations

In the previous section we have demonstrated that TVs can be estimated more accurately than pellet-trajectories (PTs) using three different speech parameterizations. While the TV-based inverse model was relatively independent of the differences in speech parameterization, the pellet-based model was not. Further, using a model-based statistical paradigm, we found that non-uniqueness in the TV-based inverse model was comparatively lower than the pellet-based model for six consonants. Finally, in a word recognition experiment we observed that TVs perform better than PTs when used along with MFCCs, indicating that estimated TVs are better than the PTs in terms of improving the robustness of word recognition system.

6.2 Gesture-based Dynamic Bayesian Network for word recognition

In section 4 we presented different models for estimating articulatory gestures and vocal tract variable (TV) trajectories from synthetic speech. We showed that when deployed on natural speech, the TVs and the gestures generated by such models helped to improve the noise robustness of a HMM based speech recognition system. Note that such architecture requires explicit recognition of the gestures. In this section, we propose a Gesture based Dynamic Bayesian Network (G-DBN) architecture that uses the gestural activations as hidden random variables, eliminating the necessity for explicit gesture recognition. In G-DBN the gestural activation random variables are treated as observed during the training[27] but as hidden during the testing. The proposed G-DBN uses MFCCs and estimated TVs as observations, where the estimated TVs are obtained from the FF-ANN based TV-estimator presented in section 6.1. Using the proposed architecture we performed a word recognition task for the noisy utterances of Aurora-2 and present the results in this section.

6.2.1 The G-DBN architecture

In section 4.2.3 we showed that articulatory gestures can be recognized with a higher accuracy if the knowledge of the TV trajectory is used in addition to the acoustic parameters (MFCCs) as opposed to using either the acoustic parameters or TVs alone. In a typical ASR setup, the only available observable is the acoustic signal, which is parameterized as acoustic features. Thus the TV-estimator presented in section 6.1 can be used to estimate the TVs from the acoustic parameters. We noted that a 4-hidden layer FF-ANN based TV-estimator with MFCC as acoustic feature gives a slight improvement in performance over the 3-hidden layer architecture used in section 6.1. Also note that the TV-estimator is trained with 8KHz data because Aurora-2 contains utterances sampled at 8KHz sampling rate. The ANN was trained

[27] The training corpus must contain annotated gestural activation functions.

using the same way as mentioned in section 6.1. The optimal number of neurons in each of hidden layer of the ANN was found to be 225, 150, 225 and 25. The raw ANN outputs were processed with a Kalman smoother to retain the intrinsic smoothness characteristic of the TVs. Table 6.3 presents the PPMC and RMSE between the actual (groundtruth) and the estimated TVs obtained from the 4-hidden layer FF-ANN after Kalman smoothing. Note that the average PPMC in Table 6.3 is slightly better than that shown for the MFCCs in Table 6.1. Also the average RMSE is not shown in Table 6.3 as RMSE for each TV has a different unit of measure and hence taking their average may not be meaningful. The input to the FF-ANN was contextualized MFCC coefficients (contextualized in the same way as performed in section 6.1) and the outputs were the eight TVs. The ANN outputs were used as an observation set by the DBN.

Table 6.3 RMSE and PPMC of the estimated TVs obtained from the 4-hidden layer FF-ANN

	RMSE	PPMC
GLO	0.080	0.853
VEL	0.036	0.854
LA	1.871	0.801
LP	0.593	0.834
TBCL	12.891	0.860
TBCD	2.070	0.851
TTCL	8.756	0.807
TTCD	4.448	0.801
Avg		0.833

A DBN (Ghahramani, 1998) is essentially a Bayesian Network (BN) that contains temporal dependency. A BN is a form of graphical model where a set of random variables (RVs) and their inter-dependencies are modeled using nodes and edges of a directed acyclic graph (DAG). The nodes represent the RVs and the edges represent their functional dependency. BNs help to exploit the conditional independence properties between a set of RVs, where dependence is reflected by a connecting edge between a pair of RVs and independence is reflected by its absence. For *N* RVs, X_1, X_2, … X_n, the joint distribution is given by

$$p\left(x_1, x_2, \ldots. x_N\right) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1 x_2) \ldots. p(x_N \mid x_1 \ldots x_N) \qquad (51)$$

Given the knowledge of conditional independence, a BN simplifies equation (51) into

$$p\left(x_1, x_2, \ldots. x_N\right) = \prod_{i=1}^{N} p(x_i \mid x_{\pi_i}) \qquad (52)$$

where x_{π_i} are the conditional parents of x_i.

Figure 6.4 shows a DBN with three discrete hidden RVs and two continuous observable RVs. The ‘prologue’ and the ‘epilogue’ in Figure 6.4 represent the initial and the final frames and the ‘center’ represents the intermediate frames, which are unrolled in time to match the duration of a specific utterance (more details about them can be obtained from Bilmes [2002]). Unlike HMMs, DBNs offer the flexibility to realize multiple hidden state variables at a time, which makes DBNs appropriate for realizing the gestural framework that involves multiple variables (gestures in our case, e.g. LA, TBCD, TTCD, etc). Hence, DBNs can explicitly model the interdependencies amongst the gestures and simultaneously perform gesture recognition and word recognition, eliminating the necessity of performing explicit gesture recognition as a prior separate step. In this work we have used the GMTK (Bilmes, 2002) to implement our DBN models, in which conditional probability tables (CPTs) are used to describe the probability distributions of the discrete RVs given their parents, and Gaussian mixture models (GMMs) are used to define the probability distributions of the continuous RVs.

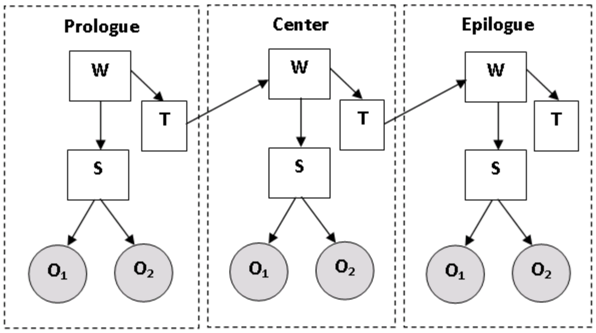


Figure 6.4 A sample DBN showing dependencies between discrete RVs (W, S, T) and continuous observations (O_1 & O_2). Round/square nodes represent continuous/discrete RV and shaded/unshaded nodes represent observed/hidden RVs

In a typical HMM based ASR setup, word recognition is performed using Maximum a Posteriori probability

$$\begin{aligned} w &= \arg\max_i P(w_i \mid o) \\ &= \arg\max_i \frac{P(w_i)P(o \mid w_i)}{P(o)} \\ &\approx \arg\max_i P(w_i)P(o \mid w_i) \end{aligned} \tag{53}$$

where o is the observation variable and $P(w_i)$ is the language model, which can be ignored for an isolated word recognition problem where all the words w are equally probable. Hence we can only focus on $P(o|w_i)$ which can be simplified further as

$$\begin{aligned} P(o \mid w) &= \sum_q P(q, o \mid w) \\ &= \sum_q P(q \mid w) P(o \mid q, w) \\ &\approx \sum_q P(q_1 \mid w) P(o_1 \mid q_1, w) \prod_{i=2}^{n} P(q_i \mid q_{i-1}, w) P(o_i \mid q_i, w) \end{aligned} \tag{54}$$

where q is the hidden state in the model. Thus in this setup the likelihood of the acoustic observation given the model is calculated in terms of the emission probabilities $P(o_i|q_i)$ and the transition probabilities $P(q_i|q_{i-1})$. Use of articulatory information introduces another RV, a and then (54) can be reformulated as

$$P(o\mid w)\approx\sum_{q}P(q_1\mid w)P(o_1\mid q_1,a_1,w)\times \prod_{i=2}^{n}P(q_i\mid q_{i-1},w)P(a_i\mid a_{i-1},q_i)P(o_i\mid q_i,a_i,w) \tag{55}$$

A DBN can realize the causal relationship between the articulators and the acoustic observations $P(o|q,a,w)$ and also model the dependency of the articulators on the current phonetic state and previous articulators $P(a_i|a_{i-1},q_i)$. Based on this formulation, the G-DBN shown in Figure 6.5 can be constructed, where the discrete hidden RVs, W, P, T and S represent the word, word-position, word-transition and word-state. The continuous observed RV, O_1 is the acoustic observation in the form of MFCCs, and O_2 is the articulatory observation in the form of the estimated TVs. The partially shaded discrete RVs, A_1, …A_N represent the discrete hidden gestures. They are partially shaded as they are observed at the training stage and then made hidden during the testing stage. The overall hybrid ANN-DBN architecture is shown in Figure 6.6.

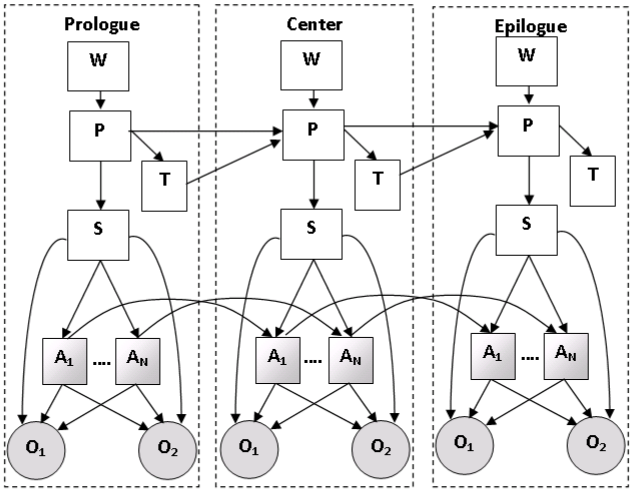


Figure 6.5 The G-DBN graphical model

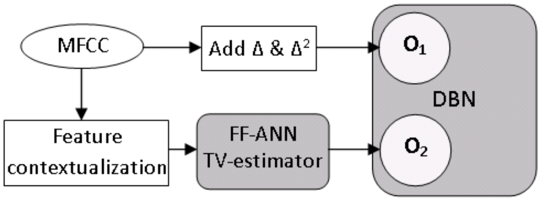


Figure 6.6 The hybrid ANN-DBN architecture

In the hybrid ANN-DBN architecture, there are two sets of observation fed to the DBN, (1) O_1: the 39 dimensional MFCCs (13 cepstral coefficients along with their Δ and Δ^2), (2) O_2: the estimated TVs obtained from the FF-ANN based TV-estimator.

6.2.2 Word Recognition Experiments

We implemented 3 different versions of the DBN, in the first version we used just the 39 dimensional MFCCs as the acoustic observation and no articulatory gesture RV was used. We name this model as the DBN-MFCC-baseline system. In this setup the word models consisted of 18 states (16 states per word and 2 dummy states). There were 11 whole word models (zero to nine and oh) and 2 models for 'sil' and 'sp', with 3 and 1 state(s) respectively. The maximum number of Gaussian mixtures allowed per state was four with vanishing of mixture-coefficients allowed for weak mixtures. The second version is identical to the first version, except that there was an additional observation RV corresponding to the estimated TVs. We name this model as the DBN-MFCC-TV system. Finally the third version was the G-DBN architecture (shown in Figure 6.5, with MFCC and the estimated TVs as two sets of observation) where we used 6 articulatory gestures as hidden RV, so *N* in Figure 6.5 was 6. Note that the articulatory gesture RVs modeled only the gestural activations, i.e., they were only binary RVs reflecting whether the gesture is active or not and do not have any target information (i.e., degree and location of the constriction information). This was done deliberately to keep the system tractable, otherwise the multi-dimensional conditional CPT linking the word state RVs and the gesture state RVs became extremely large making the DBN overly complex. Hence our current implementation of G-DBN uses 6 gesture RVs: GLO, VEL, LA, LP, TT and TB. Since the gestural activations for TTCL and TTCD are identical they were replaced by a single RV, TT (tongue tip) and the same is true for TBCL and TBCD, which were replaced by TB (tongue body). Since the TVs were used as a set of observation and the TVs by themselves contain coarse target specific information about the gestures, it can be expected that the system has gestural target information to some extent. The word models in the G-DBN architecture uses lesser number of states per word[28] (eight with two additional dummy states) compared to that (16 states per word and 2 dummy states) of the DBN-MFCC-baseline and DBN-MFCC-TV systems. The number of states for 'sil' and 'sp' were kept the same as before. In this setup the discrete gesture RVs are treated as observable during the training session and then converted to a hidden RV during the testing. Figure 6.7 shows the overall word recognition accuracy obtained from the three DBN versions implemented.

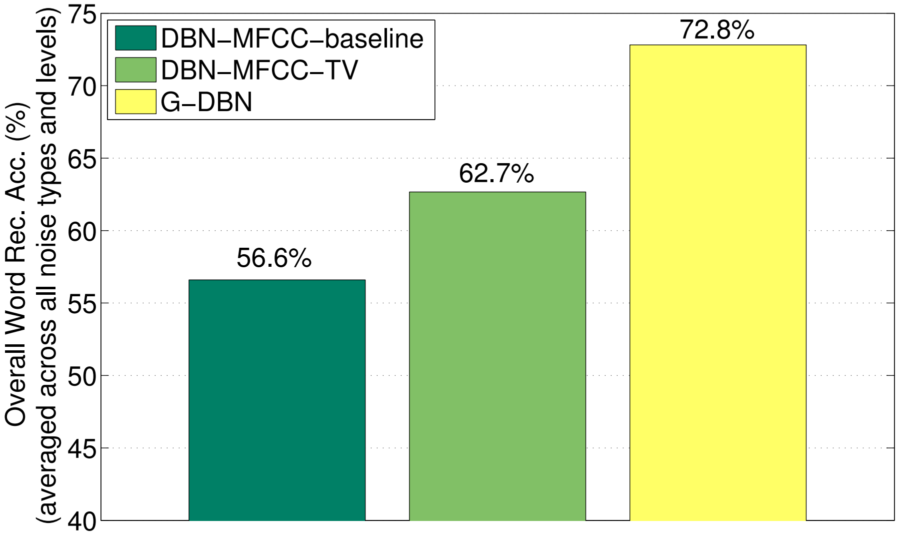


Figure 6.7 Overall word recognition accuracy obtained from the three DBN versions

[28] This reduction was done to reduce the DBN complexity

Figure 6.7 show that the G-DBN provided the best overall word recognition accuracy despite having a lower number of states per word model. Use of estimated TVs in addition to the MFCCs offered higher recognition accuracy than the MFCC-baseline which is in line with our previous observation (Mitra *et al.*, to appear). In Figure 6.8 we compare our results (for SNR 0dB to 20dB) with some of the previously obtained HMM-based results. Note that all the HMM based systems use a 16 state/word model, whereas the G-DBN uses an 8-state/word model.

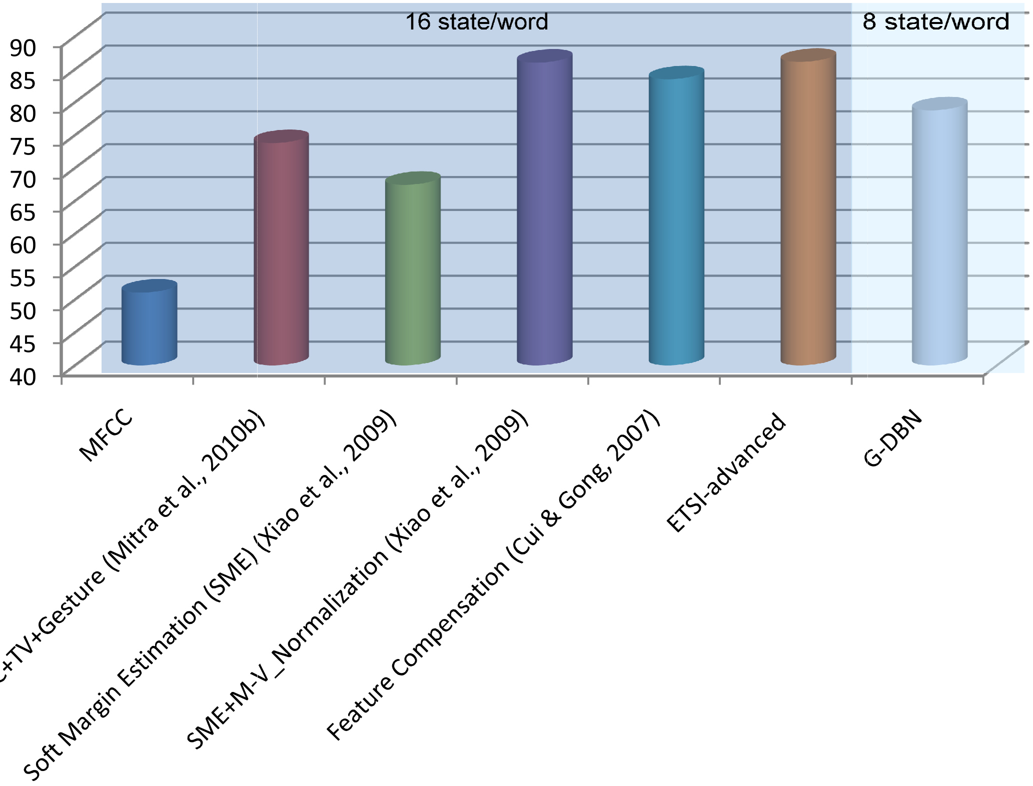


Figure 6.8 Averaged Recognition accuracies (0 to 20dB) obtained from using the G-DBN architectures presented in this section, our prior HMM based articulatory gestures based system (Mitra *et al.*, 2010b) and some state-of-the-art word recognition systems that has been reported so far

For a fair comparison, we created an 8 state/word model for the ETSI-advanced[29] and the ETSI-advanced with the G-DBN back-end. The recognition results obtained are compared to that of the G-DBN in Table 6.4.

Table 6.4 Word recognition accuracy at clean, 0-20dB and -5dB for the whole Aurora-2 database, using G-DBN[30], ETSI-advanced front-end and ETSI-advanced front-end with G-DBN. The numbers in bold denote the highest recognition accuracies obtained at that SNR range.

	Clean	0-20dB	-5dB
G-DBN	98.52	78.77	17.42
ETSI-advanced	98.14	**82.01**	23.71
ETSI-advanced+G-DBN	**98.62**	81.48	**23.89**

Table 6.4 shows that both the G-DBNs showed better word recognition accuracies than the ETSI-advanced front-end at clean. Also when the ETSI-advanced frontend is used with the G-DBN back-end, it offered higher word recognition accuracy at -5dB than only the ETSI-advanced front-end.

To compare the performance of the G-DBN system (which uses whole word models) with a phone-based model, we built a DBN (using MFCC as acoustic features) where the total number of phones was 60 and the maximum number of phones per word was 30. We performed a word recognition experiment, which compares the mono-phone based DBN to the G-DBN for a clean test set in Aurora-2 and the results are shown in Table 6.5

[29] In this case we used a DBN back end, without any hidden gesture variables. Hence essentially the backend is an HMM system.

[30] In case of G-DBN only, the acoustic feature consists of MFCC coefficients with their Δs and Δ^2s.

Table 6.5 Word recognition accuracy at clean condition: G-DBN versus mono-phone DBN

Mono-phone DBN	98.31
G-DBN	98.93

Table 6.5 shows that the G-DBN architecture provides improved recognition accuracy at clean condition than the mono-phone based model, indicating that gestural representation potentially can improve word recognition rates over the mono-phone based representation.

6.2.3 Discussion

In this section we proposed and presented an articulatory gesture based DBN architecture that uses acoustic observations in the form of MFCC and estimated TV trajectories as input. Using an eight state per word model we have shown that the G-DBN architecture can significantly improve the word recognition accuracy over the DBN architectures using MFCCs only or MFCCs along with TVs as input. Our results also show that the proposed G-DBN significantly improves the performance over a gesture based HMM architecture we previously proposed in (Mitra *et al.*, 2010b), indicating the capability of DBNs to properly model parallel streams of information (in our case the gestures). Note that the current system has several limitations as follows. First, the TV estimator is trained with only a single speaker, and a multi-speaker trained TV estimator can potentially increase the TV estimates for the Aurora-2 database, which in turn can further increase the word recognition accuracy. Second, we only modeled the gestural activations as hidden binary RVs. Future research should include gestural target information as well. Finally, we have seen (Mitra *et al.*, 2010b) that contextualized acoustic observation can potentially increase the performance of gesture recognition. However, in our current implementation the acoustic observation had no contextual information. Contextual information should be pursued in future research.

Chapter 7: Summary and future work

7.1 Summary

This dissertation presents an alternative approach to automatic speech recognition, where articulatory gestures are used as speech sub-units instead of phones. The new architecture not only introduces robustness against variability in speech due to contextual variation but also against ambient noise contamination. In order to use articulatory gestures as sub-word speech units, these gestures need to be extracted / recognized from the speech signal, so the first logical step is to see if appropriate models can be build that can generate / recognize the corresponding gestures from a given speech input. However to build / train such models we require a database containing speech utterances and their corresponding groundtruth gestural scores, but unfortunately no such natural speech database existed during the time we begin our experiments. Hence, we had to use synthetic speech data that contain acoustic waveforms and their corresponding gestural scores and TV trajectories. In chapter 3 we introduced Haskins Laboratories TAsk Dynamics Application (TADA) model, which given a word or its arpabet, generates synthetic speech acoustics, its corresponding gestural scores and TV trajectories. In that chapter we also specified the synthetic databases that we created for the initial studies performed in this research.

Chapter 4 presented a set of initial studies performed on the synthetic databases presented in chapter 3. In the initial study we presented different machine learning strategies to recover TVs from a speech signal. We observed that using contextual information in the acoustic space helps to better estimate the TV than without using any contextual information. We showed that TVs can be estimated with an overall better accuracy than articulatory pellet trajectories, from the speech signal. Our study used different machine learning approaches for TV estimation, and the approach (3-hidden layer FF-ANN architecture) that modeled the non-linearity well was found to offer the overall best result; which may indicate that the non-linearity may be the critical factor rather than non-uniqueness for speech inversion using TVs. Also we observed that the raw TV estimates from the TV estimators were almost always corrupted with an estimation noise, hence we used a Kalman smoother post-processor to smooth the raw TV-estimates, which helped to improve the overall TV estimation performance. For gesture recognition, we proposed a cascaded neural network architecture that generates the gestural scores as the output. We observed that when acoustic parameters (derived from the acoustic signals) are used with the estimated TVs as input, the architecture offers greater recognition accuracy over that using the acoustic parameters or the TVs alone. Which indicates that the use of estimated TVs as tandem-features with acoustic parameters ensure higher accuracy for gesture recognition. Finally, we investigated the possibility of using the estimated TVs and recognized gestural scores as a possible input to a word recognizer both at clean and noisy conditions. The word recognition results indicate that using articulatory information in the form of TVs and gestural scores (represented as GPVs) in addition to acoustic features can improve the recognition rates appreciably for noisy speech. Clearly showing that use of articulatory information can potentially improve noise robustness of ASR systems. Note that all of our initial exploration used models trained with synthetic speech corpus, which might have limited the capability of these models in predicting the TVs and gestural scores when deployed on natural speech. To account this, we wanted to create a natural speech corpus containing TVs and gestural score specifications, so that such information could be used to train the TV-estimator and gestural score recognizer models.

In Chapter 5 we presented a landmark based iterative analysis-by-synthesis time-warping architecture that can annotate speech articulatory gestures and TV trajectories, potentially to any speech database containing word and phone transcriptions and their time alignment. This approach generates speaker independent articulatory information making them ideal for ASR applications. Using that architecture we annotated the TVs and gestural scores for the whole of XRMB database. Since XRMB contain recorded articulatory pellet trajectories and some of whom can be used to coarsely predict the TV trajectories, we performed a comparison between the annotated TVs and the TVs deciphered from the flesh-point data and show that the two correlate well. This indicated that the annotation procedure is indeed generating meaningful articulatory information. In a different study we used the annotated gestural scores from a part of the XRMB database to train gestural-score bigram word models which were used to perform word recognition on the remainder of the database. An error rate as low as 2.5% was obtained, demonstrating that the gestural scores are indeed a viable representation for speech recognition tasks.

In chapter 6, we re-evaluated our observations made in the initial study with natural speech data. In the first experiment we observed that TVs can be estimated more accurately than PTs which confirms our observation with synthetic speech presented in chapter 4. We also observed that the MFCCs are a better acoustic parameterization for TV-based speech inversion task. Here we also performed a model-based statistical non-uniqueness analysis of the TV-based and pellet-based inverse model and quantitatively demonstrated that the former has comparatively lower non-uniqueness than the latter for six consonants. Using a word recognition experiment we showed that the TVs perform better than pellets when used along with MFCCs; indicating that the TVs are a better representation for ASR. In the final experiment in chapter 6, we presented a DBN architecture that performs word recognition using articulatory gestures as a hidden random variable, eliminating the necessity for explicit gesture recognition as performed in chapter 4. The proposed articulatory gesture based DBN architecture uses acoustic observations in the form of MFCC and estimated TV trajectories as input. The proposed hidden gesture based DBN architecture showed significant improvement in word recognition accuracies over the DBN architectures using MFCCs only or MFCCs along with the TVs as input.

7.2 Future Direction

There are several directions that the research presented in this dissertation could be pursued in the future:

(1) Large Vocabulary Continuous Speech Recognition (LVSCR): As this dissertation for the first time realized a full-blown running speech recognition system that uses articulatory gestures as hidden variables, we have tried to keep the recognition experiments simple to confirm the fact

that articulatory gestures indeed offers promise for speech recognition. Future research should extend the experiments reported in this dissertation to medium and large vocabulary continuous speech recognition tasks and, in such cases, the strength of the articulatory gestures to model coarticulation well, should be more apparent. To be able to train acoustic models for large vocabulary, we need to annotate the training data with TVs and gestural scores, which is certainly doable given the training data's word and phone transcripts with their time alignment information.

(2) Speaker Recognition: The gestural score annotation procedure laid out in this dissertation to decipher the gestural scores and TV trajectories of a natural speech utterance is based on a canonical gestural model in TADA. Hence such information can be expected to be relatively speaker independent and suitable for primarily speech recognition tasks. The gestural annotation procedure can be modified in a way that it can learn speaker specific attributes such as (a) structural differences (due to vocal tract length, gender etc.) and (b) stylistic differences (due to speech dynamics, intergestural timing differences, prosody, speaker idiosyncrasy, etc.). Hence creating a set of parameters that are speaker specific in nature can be used as input cues in a speaker identification (SID) task.

(3) Speech Enhancement: The idea of estimating the TVs from the acoustic signal can also find its application in speech enhancement (noise suppression) applications. Usually in a speech enhancement application the voiced regions are extracted and retained very well due to their

inherent periodic structure which helps them to be identified relatively easily from the background aperiodic noise. However unvoiced consonantal regions being aperiodic regions get blended well with the background aperiodic noise, making them increasingly difficult to detect and extract. The estimated TV information can be used to robustly separate consonant speech sounds from the background noise. Remember that TVs specify the location and degree of constriction at different constriction sites in the human vocal tract. If the consonant regions can be recognized from the estimated TVs and their place and manner of articulation identified, then such consonantal information can be pulled out of the background noise robustly.

(4) Assistive Devices: The articulatory information presented in this dissertation can have its application in different assistive devices.

   a. *Visual Speech*: the TV and the gestural score information obtained from the acoustic signal can be used to create a 3-dimensional dynamic vocal tract model. Such a 3-D model can be used to develop a talking head with the help of computer graphics. Such talking head may find its application in creating visual speech for the hearing impaired, visual aids for subjects suffering from speech disfluencies etc.

   b. *Second Language acquisition*: 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 properly reach the target articulation or may use a wrong articulation pattern that results in failure to produce such sounds properly. Given that the TVs can be estimated form the subjects speech signal, the subjects' articulatory dynamics can be studied and compared with that of its canonical pronunciation to obtain information regarding what the subject is doing wrong in terms of the articulation and how he/she can correct it.

(5) Multi-language ASR: Speech Gestures, being the action units responsible for articulatory motions can potentially be language-independent recognition units, unlike phonemes; hence should allow for portability of ASR systems from one language to another. Such a task if achieved would indicate the economy and versatility of using gestures as subword units as opposed to the conventional phones. Future experiments need to be performed to see how a gesture-based ASR architecture trained on speech from one language can be ported to another language and hence perform cross-lingual speech recognition tasks across languages such as English, Spanish, French etc.

Appendices

Appendix A: List of APs

Table A-A.1 List of APs

	APs	**Description**
1	E0_lessF3_SF	Ratio of the Energy in BW [0 - F3_avg-1000] Hz to Energy in BW [F3_avg-1000 - Fs/2] (BW: bandwidth; F3 = 3rd formant frequency; Fs = sampling rate)
2	k_1	The first Reflection coefficient
3	E200_3000	Energy in BW [200Hz - F3_avg Hz], previously was E[0,F3-1000], the -1000 was dropped later (E: Energy)
4	E3000_6000	Energy in BW [F3_avg - Fs/2] Hz
5	E_total	Total Energy
6	voice_bars	ratio of the (Peak Energy in 0-400Hz) w.r.t (Peak Energy in 1000-fs/2) measured in dB
7	paf	Energy in in the band (F3_avg-187)Hz to 781Hz
8	Av_maxA23	Amplitude of the low frequency peak of the vowel spectrum - Amplitude of the max frequency in F2 - F3 range (F2 = 2nd formant frequency)
9	Av_Ahi	Amplitude of the low frequency peak of the vowel spectrum - Amplitude of the max frequency peak at the burst spectrum
10	F0_out	pitch profile
11	AhiArray	Amplitude of the high frequency peak at the burst spectrum

	APs	Description
12	AvLocArray	Location of Av in Hertz [Juneja (2004)]
13	dip640	Juneja (2004)
14	dip2000	Juneja (2004)
15	peak640	Juneja (2004)
16	E640_2800_raw	Energy in BW [640Hz - 2800Hz] not normalized
17	E2000_3000_raw	Energy in BW [2000Hz - 3000Hz] not normalized
18	zcr_vals_sm	zero crossing rate
19	hifreq_zcr_vals_sm	high frequency zero crossing, to capture zc overriding on signal envelope (zcr for hi-pass filtered signal, where the high pass cutoff frequency is F3_avg+1000 Hz) (zcr: zero crossing rate)
20	FB1_B0	Formant 1 - Formant 0 in Bark scale
21	FB2_B1	Formant 2 - Formant 1 in Bark scale
22	FB3_B2	Formant 3 - Formant 2 in Bark scale
23	F1_out	Formant 1 profile
24	F2_out	Formant 2 profile
25	F3_out	Formant 3 profile
26	E5000_6250_0_3000	Ratio of the Energy in BW [5000Hz - 6250Hz] and Energy in BW [0Hz to 3000Hz] in dB
27	E0_320_5360	Ratio of the energy in BW [0 to 320Hz] and energy in BW [320 to 5360Hz] measured in dB
28	mean_hilbert_env	Mean of Hilbert Envelop estimated for each frame
29	std_hilbert_env	standard deviation of Hilbert Envelop estimated for each frame

	APs	Description
30	per_0_1800	Periodic energy from APP detector for BW 0-1800Hz (Aperiodic Periodic and Pitch detector [Deshmukh *et al.*, 2005])
31	per_1800_2600	Periodic energy from APP detector for BW 1800-2600Hz
32	per_2600_3500	Periodic energy from APP detector for BW 2600-3500Hz
33	per_3500_Fs2	Periodic energy from APP detector for BW 3500-Fs/2Hz
34	PER_0_500	Periodic energy from APP detector for BW 0-500Hz
35	aper_0_1800	APeriodic energy from APP detector for BW 0-1800Hz
36	aper_1800_2600	APeriodic energy from APP detector for BW 1800-2600Hz
37	aper_2600_3500	APeriodic energy from APP detector for BW 2600-3500Hz
38	aper_3500_Fs2	APeriodic energy from APP detector for BW 3500-Fs/2Hz
39	aper_1000_Fs2	APeriodic energy from APP detector for BW 1000-Fs/2Hz
40	Per_smmry	Periodic energy summary

Appendix B: Significance Tests

The significance test (using the approach specified by Gillick & Cox [1989]) results showing that the TVs (estimated using synthetically trained TV estimator) and their Δs offered better word recognition accuracy over MFCCs at 0dB and -5dB SNR (as stated in section 4.3.2.3) is presented in Table A-B.1

Table A-B.1 Significance Tests for TV-MFCC, (TV+Δ)-MFCC, (TV+Δ+Δ^2)-MFCC, (TV+Δ+Δ^2+Δ^3)-MFCC pairs for 0dB and -5dB SNR

	TV-MFCC	(TV+Δ)-MFCC	(TV+Δ+Δ^2)-MFCC	(TV+Δ+Δ^2+Δ^3)-MFCC
0dB	1.60E-10	1.65E-08	3.60E-07	2.3E-03
-5dB	4.49E-08	5.63E-08	2.07E-08	6.32E-04

We also performed the significance test to show that the contextualized TVs (i.e., TVs with their Δs) did not show better accuracies than TVs alone (as specified in section 4.3.2.3) and the result is given in Table A-B.2

Table A-B.2 Significance Tests for TV-(TV+Δ), TV-(TV+Δ+Δ^2), TV-(TV+Δ+Δ^2+Δ^3) across all noise type and noise-levels in Aurora-2

TV-(TV+Δ)	TV-(TV+Δ+Δ^2)	TV-(TV+Δ+Δ^2+Δ^3)
4.67E-2	4.89E-04	1.93E-11

Bibliography

Ananthakrishnan, G., Neiberg, D. and Engwall, O. (2009) "In search of Non-uniqueness in the Acoustic-to-Articulatory Mapping", in Proc. of Interspeech, pp. 2799-2802, Brighton, UK.

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", J. Acoust. Soc. of Am., 63, pp. 1535-1555.

Atal, B.S. (1983), "Efficient coding of LPC parameters by temporal decomposition", Proc. of ICASSP, pp. 81-84, Boston, MA, USA.

Barker, J., Josifovski, L., Cooke, M.P. and Green, P.D. (2000) "Soft decisions in missing data techniques for robust automatic speech recognition", Proc. of Int. Conf. Spoken Lang. Processing, pp. 373-376.

Barker, J. and Cooke, M.P. (2007) "Modelling speaker intelligibility in noise," Speech Communication, 49: 402-417.

Bilmes, J. (2002) "GMTK: The Graphical Models Toolkit", SSLI Laboratory, Univ. of Washington. [http://ssli.ee.washington.edu/~bilmes/gmtk/]

Bishop, C. (1994) "Mixture density networks", Tech. Report NCRG/4288, Neural Computing Research Group, Dept. of Comp. Sc., Aston Univ., Birmingham, U.K.

Boyce, S. and Espy-Wilson, C.Y. (1997) "Coarticulatory stability in American English /r/", J. Acoust. Soc. of Am., **101**(**6**), pp. 3741-3753.

Browman, C. and Goldstein, L. (1989) "Articulatory Gestures as Phonological Units", Phonology, 6: 201-251.

Browman, C. and Goldstein, L. (1990) "Gestural specification using dynamically-defined articulatory structures", J. of Phonetics, Vol. 18, pp. 299-320.

Browman, C. and Goldstein, L. (1992) "Articulatory Phonology: An Overview", Phonetica, 49: 155-180.

Byrd, D. (2000) "Articulatory vowel lengthening and coordination at phrasal junctures", Phonetica, 57 (1), pp. 3-16.

Byrd, D. and Saltzman, E. (2003) "The elastic phrase: modeling the dynamics of boundary-adjacent lengthening", J. of Phonetics, **31**(**2**), pp. 149-180, Elsevier Science Ltd.

Cetin, O., 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", Proc. ICASSP, Vol. 4, pp. 645-648.

Chang, J. and Glass, J. (1997) “Segmentation and modeling in segment-based recognition”, Proc. of Eurospeech, pp. 1199-1202, Rhodes, Greece.

Chang, S., Wester, M. and Greenberg, S. (2005) “An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language”, Speech Comm., **47**(**3**), pp. 290-311.

Chen, C. and Bilmes, J. (2007) “MVA Processing of Speech Features”, IEEE Trans. Audio, Speech & Lang. Processing, 15(1), pp. 257-270.

Chen, S. and Alwan, A. (2000) “Place of articulation cues for voiced and voiceless plosives and fricatives in syllable-initial position”, Proc. of ICSLP, Vol. 4, pp. 113-116.

Cho, T. (2005) “Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /A, i/ in English”, J. of the Acoust. Soc. of Am., 117 (**6**), pp. 3867-3878.

Choi, J.Y. (1999) “Detection of Consonant Voicing: A Module for a Hierarchical Speech Recognition System”, Ph.D. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

Chomsky, N. and Halle, M. (1968) “The Sound Pattern of English”, Harper & Row, New York.

Clark, J. and Yallop, C. (1995) “An introduction to Phonetics and Phonology”, 2nd ed., Blackwell Publishers Ltd., Oxford, UK.

Clements, G.N. and Hume, E.V. (1995) “The Internal Organization of Speech Sounds”, Handbook of Phonological Theory, Goldsmith J.A. eds., Blackwell Publishers, Cambrige.

Cole, R., Stern, R.M. and Lasry, M.J. (1986) “Performing Fine Phonetic Distinctions: Templates versus Features”, in Invariance and Variability of Speech Processes, J.S. Perkell and D. Klatt eds., Lawrence Erlbaum Assoc., Hillsdale, NJ, chap. 15, pp. 325–345.

Cole, R., Noel, M., Lander, T. and Durham, T. (1995) “New telephone speech corpora at CSLU”, Proc. of 4th European Conference on Speech Communication and Technology, Vol. 1, pp. 821-824.

Cooke, M., Green, P., Josifovski, L. and Vizinho, A. (2001) “Robust automatic speech recognition with missing and uncertain acoustic data”, Speech Comm., Vol. 34, pp. 267-285.

Cooke, M., Barker, J., Cunningham, S. and Shao, X. (2006) “An audio-visual corpus for speech perception and automatic speech recognition”, Journal of Acoustical Society of America, 120: 2421-2424.

Cover, T. M. and Thomas, J.A. (1991), “Elements of Information Theory”, Wiley, New York, USA.

Cui, X. and Gong, Y. (2007) “A Study of Variable-Parameter Gaussian Mixture Hidden Markov Modeling for Noisy Speech Recognition”, IEEE Trans. Audio, Speech & Lang. Process., Vol. 15, Iss. 4, pp. 1366-1376.

Dalsgaard, P. (1992) “Phoneme Label alignment using acoustic-phonetic features and Gaussian probability density functions”, Computer, Speech & Language, Vol. 6, pp. 303-329.

De Mori, R., Laface, P. and Piccolo, E. (1976) “Automatic detection and description of syllabic features in continuous speech”, IEEE Trans. on Acoust., Speech & Sig. Processing, **24**(**5**), pp. 365–379.

Demuth, H., Beale, M. and Hagan, M. (2008), “Neural Network ToolboxTM 6, User’s Guide”, The MathWorks Inc., Natick, MA. [www.mathworks.com/access/helpdesk/help/pdf_doc/nnet/nnet.pdf].

Deng, L. and Erler, K. (1991) “Microstructural speech units and their HMM representations for discrete utterance speech recognition”, Proc. of ICASSP, pp. 193-196.

Deng, L. (1992) “A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal”, Signal Processing, **27**(**1**), pp. 65-78.

Deng, L. and Sun, D (1994a) “A statistical approach to ASR using atomic units constructed from overlapping articulatory features”, J. Acoust. Soc. of Am., Vol. 95, pp. 2702-2719.

Deng, L. and Sun, D. (1994b) “Phonetic classification and recognition using HMM representation of overlapping articulator features for all classes of English sounds”, Proc. of ICASSP, pp. 45-47.

Deng, L. (1997) “Autosegmental representation of phonological units of speech and its phonetic interface”, Speech Comm., **23**(**3**), pp. 211-222.

Deng, L., Ramsay, G. and Sun, D. (1997) “Production models as a structural basis for automatic speech recognition”, Spec. Iss. on Speech Prod. Modeling, Speech Comm., **22**(**2**), pp. 93-112.

Deng, L. (1998) “A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition”, Speech Comm., **24**(**4**), pp. 299-323.

Deng, L. and Ma, J. (2000) “Spontaneous speech recognition using a statistical coarticulatory model for the hidden vocal-tract-resonance dynamics”, J. of Acoust. Soc. Am., **108**(**6**), pp. 3036-3048.

Deng, L., Lee, L., Attias, H. and Acero, A. (2004) “A Structured Speech Model with Continuous Hidden Dynamics and Prediction-Residual Training for Tracking Vocal Tract Resonances”, Proc. of ICASSP, pp. I557-I560.

Deng, L., Yu, D. and Acero, A. (2006) “Structured Speech Modeling”, IEEE Trans. Audio, Speech & Lang. Processing, **14**(**5**), pp. 1492-1504.

Deshmukh, O., Espy-Wilson, C., Salomon, A. and Singh, J. (2005) “Use of Temporal Information: Detection of the Periodicity and Aperiodicity Profile of Speech”, IEEE Trans. Speech & Audio Processing, 13(5), pp.776-786.

Deshmukh, O., Espy-Wilson, C. and Carney, L.H. (2007) “Speech Enhancement Using The Modified Phase Opponency Model”, J. Acoust. Soc. of Am., 121(6), pp. 3886-3898.

Dharanipragada, S., Yapanel, U. and Rao, B. (2007), “Robust Feature Extraction for Continuous Speech Recognition Using the MVDR Spectrum Estimation Method”, IEEE Trans. on Audio, Speech, & Language Processing, 15(1), pp. 224-234.

Dusan, S. (2000) “Statistical Estimation of Articulatory Trajectories from the Speech Signal Using Dynamical and Phonological Constraints”, PhD Thesis, Dept. of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada.

Dusan, S. (2001) “Methods for Integrating Phonetic and Phonological Knowledge in Speech Inversion”, Proc. of the International Conference on Speech, Signal and Image Processing, SSIP, pp. 194-200, Malta.

Eide, E., Rohlicek, J.R., Gish, H. and Mitter, S. (1993) “A linguistic feature representation of the speech waveform”, Proc. of ICASSP, pp. 483-486.

Elenius, K. and Tacacs, G. (1991) “Phoneme recognition with an artificial neural network”, Proc. of Eurospeech, pp. 121-124.

Elenius, K. and Blomberg, M. (1992) “Comparing phoneme and feature based speech recognition using artificial neural networks”, Proc. of ICSLP, pp. 1279-1282.

Ephraim, Y. and Malah, D. (1985) “Speech enhancement using a minimum mean square log-spectral amplitude estimator”, IEEE Trans. Acoust., Speech & Sig. Processing, Vol. ASSP-33(2), pp. 443-445.

Erler, K. and Deng, L. (1993) “Hidden Markov model representation of quantized articulatory features for speech recognition”, Computer, Speech & Language, Vol. 7, pp. 265–282.

Espy-Wilson, C.Y. and Boyce, S.E., (1999) “The relevance of F4 in distinguishing between different articulatory configurations of American English /r/”, J. Acoust. Soc. of Am., **105**(**2**), 1400.

Espy-Wilson, C.Y., Boyce, S.E., Jackson, M., Narayanan, S. and Alwan, A. (2000) “Acoustic modeling of American English /r/”, J. Acoust. Soc. of Am. **108**(**1**), pp. 343-356.

ETSI ES 201 108 Ver. 1.1.3, (2003). Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms.

ETSI ES 202 050 Ver. 1.1.5, (2007). Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced frontend feature extraction algorithm; compression algorithms.

Flynn, R. and Jones, E. (2008) “Combined speech enhancement and auditory modelling for robust distributed speech recognition”, Speech Comm., Vol.50, pp. 797-809.

Fowler, C. and Saltzman, E. (1993) “Coordination and coarticulation in speech production”, Language & Speech, Vol. 36, pp. 171-195.

Fowler, C.A. and Brancazio, L. (2000) “Coarticulation resistance of American English consonants and its effects on transconsonantal vowel-to-vowel coarticulation”, Language & Speech, Vol. 43, pp. 1-42.

Fowler, C.A. (2003) “Speech production and perception”, Handbook of psychology, A. Healy and R. Proctor eds., Vol. 4, Experimental Psychology, pp. 237-266, New York, John Wiley & Sons.

Frankel, J., Richmond, K., King, S. and Taylor, P. (2000) “An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces”, Proc. of ICSLP, Vol. 4, pp. 254-257.

Frankel, J. and King, S. (2001) “ASR - Articulatory Speech Recognition”, Proc. of Eurospeech, pp. 599-602, Aalborg, Denmark.

Frankel, J., Wester, M. and King, S. (2004) “Articulatory feature recognition using dynamic Bayesian networks”, Proc. of ICASSP, pp. 1202-1205, Jeju, Korea.

Frankel, J. and King, S. (2005) “A Hybrid ANN/DBN Approach to Articulatory Feature Recognition”, Proc. of Eurospeech, Interspeech, pp. 3045-3048, Lisbon, Portugal.

Frankel, J., Çetin, Ö. and Morgan, N. (2008), “Transfer Learning for Tandem ASR Feature Extractionn”, Machine Learning for Multimodal Interaction, Lecture Notes in Computer Science, 4892/2008, pp. 227-236, Springer-Verlag, Germany

Fujimura, O. (1986) “Relative Invariance of Articulatory Movements: An Iceberg Model”, in Invariance and Variability of Speech Processes, J. S. Perkell and D. Klatt eds., Lawrence Erlbaum Assoc., Hillsdale, NJ, chap. 11, pp. 226–242.

Ghahramani, Z. (1998) “Learning dynamic Bayesian networks”, in, Adaptive Processing of Temporal Information, C. L. Giles and M. Gori, eds, pp. 168–197. Springer-Verlag.

Gales, M.J.F. and Young, S.J. (1996) “Robust continuous speech recognition using parallel model combination”, IEEE Trans. Speech & Audio Processing, 4(5), pp. 352–359.

Gillick, L. and Cox, S. (1989) “Some statistical issues in the compairson of speech recognition algorithms”, Proc. of ICASSP, Vol. 1, pp. 532-535.

Glass, J. (1988) “Finding acoustic regularities in speech : applications to phonetic recognition”, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1988.

Glass, J. (2003) “A Probabilistic Framework for Segment-Based Speech Recognition”, Computer Speech & Language, Vol. 17, pp. 137-152.

Goldberg, H.G. and Reddy, R. (1976) “Feature extraction, segmentation and labelling in the Harpy and Hearsay-II systems”, J. Acoust. Soc. of Am. **60**(S**1**), pp.S11.

Goldsmith, J.A. (1990) “Autosegmental and Metrical Phonology”, Blackwell Pubs.

Hanson, H.M. and Stevens, K.N. (2002) “A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn”, J. Acoust. Soc. of Am., **112**(**3**), pp. 1158-1182.

Halberstadt, A. and Glass, J. (1998) “Heterogeneous measurements and multiple classifiers for speech recognition”, Proc. of ICSLP, pp. 995-998.

Harrington, J. (1987) “Acoustic cues for automatic recognition of English consonants”, in Speech Technology: a survey, M.A. Jack and J. Laver eds., Edinburgh University Press, Edinburgh, pp. 19-74.

Harris, J. (1994) “English Sound Structure”, Wiley-Blackwell, Oxford, UK & Cambridge, MA, USA.

Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., Juneja, A., Kirchhoff, K., Livescu, K., Mohan, S., Muller, J., Sonmez, K. and Wang, T. (2005) “Landmark-based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop”, Technical Report, Johns Hopkins University, 2005.

Hasegawa-Johnson, M., Livescu, K., Lal, P. and Saenko, K. (2007) “Audiovisual speech recognition with articulator positions as hidden variables”, Proc. of ICPhS, pp. 297-302, Saarbrucken.

He, X. and Deng, L. (2008) “Discriminative Learning for Speech Processing”, Morgan & Claypool Pub., G.H. Juang eds.

Hermansky, H. and Morgan, N. (1994) “RASTA processing of speech”, IEEE Trans. Speech & Audio Processing, 2(4), pp. 578-589.

Hinton, G.E., Osindero, S. and Teh, Y. (2006) “A fast learning algorithm for deep belief nets”, Neural Comp., Vol. 18, pp. 1527-1554.

Hirsch, H.G. and Pearce, D. (2000) “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions”, Proc. of ISCA ITRW, ASR2000, pp. 181-188, Paris, France.

Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P. and Saltzman, E. (1996) “Accurate recovery of articulator positions from acoustics: new conclusions based on human data”, J. Acoust. Soc. of Am., **100**(**3**), pp. 1819–1834.

Hogden, J., Nix, D. and Valdez, P. (1998) “An Articulatorily Constrained, Maximum Likelihood Approach to Speech Recognition”, Tech. Report, LA-UR--96-3945, Los Alamos National Laboratory, NM.

Howitt, A.W. (1999) “Vowel Landmark Detection”, Proc. of Eurospeech, Vol. 6, pp. 2777-2780, Budapest, Hungary.

Huang, F.J., Cosatto, E. and Graf, H.P. (2002) “Triphone based unit selection for concatenative visual speech synthesis”, proc. of ICASSP, Vol. 2, pp. 2037-2040, Orlando, FL.

Jordan, M.I. and Rumelhart, D.E. (1992) "Forward models-Supervised learning with a distal teacher", Cogn. Sci., 16, pp. 307-354..

Josifovski, L., Cooke, M., Green, P. and Vizinho, A. (1999) "State based imputation of missing data for robust speech recognition and speech enhancement", Proc. of Eurospeech, Vol. 6, pp. 2833–2836.

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

Juneja, A. and Espy-Wilson, C. (2003a) "An event-based acoustic-phonetic approach to speech segmentation and E-set recognition", Proc. of ICPhS, pp. 1333-1336, Barcelona, Spain..

Juneja, A. and Espy-Wilson, C., (2003b) "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines", Proc. of IJCNN, Vol. 1, pp. 675-679, Portland, Oregon.

Juneja, A. and Espy-Wilson, C. (2008) "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition", J. Acoust. Soc. of Am., **123**(**2**), pp. 1154-1168.

Jung, T.P., Krishnamurthy, A.K., Ahalt, S.C., Beckman, M.E. and Lee, S.H. (1996), "Deriving gestural scores from articulator-movement records using weighted temporal decomposition", IEEE Trans. on Speech & Audio Processing, **4**(**1**), pp. 2-18.

Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Xiuyang, Y. and Sen, Z. (2001) "What kind of pronunciation variation is hard for triphones to model?", Proc. of ICASSP, Vol.1. pp.577-580.

Katsamanis, A., Papandreou, G. and Maragos, P. (2009) "Face Active Appearance Modeling and Speech Acoustic Information to Recover Articulation", IEEE Trans. Audio, Speech & Lang. Processing, 17(3), pp. 411-422.

King, S. and Taylor, P. (2000) "Detection of Phonological Features in Continuous Speech using Neural Networks", Computer Speech & Language, 14(4), pp. 333-353.

King, S. Bartels, C. and Bilmes, J. (2005) "SVitchboard 1: Small vocabulary tasks from Switchboard 1", Proc. of Interspeech, pp. 3385-3388, Lisbon, Portugal.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K. and Wester, M. (2007) "Speech production knowledge in automatic speech recognition", J. of Acoust. Soc. of Am., 121(**2**), pp. 723-742.

Kingsbury, B., Saon, G., Mangu, L., Padmanabhan, M. and Sarikaya, R. (2002) "Robust speech recognition in noisy environments: the 2001 IBM SPIN Eevaluation system", Proc. of ICASSP, Vol. 1, pp. I-53–I-56, FL.

Kirchhoff, K. (1999), "Robust Speech Recognition Using Articulatory Information", PhD Thesis, Univ. of Bielefeld, Germany.

Kirchhoff, K., Fink, G.A. and Sagerer, G. (2002) “Combining acoustic and articulatory feature information for robust speech recognition”, Speech Communication, vol. 37, pp. 303-319.

Kobayashi, T., Yagyu, M. and Shirai, K. (1985) “Application of Neural networks to articulatory motion estimation”, Proc. of ICASSP, pp. 1001-1104, Tampa, FL.

Ladefoged, P. (1975) “A Course in Phonetics”, Harcourt College Pub, New York, US.

Ladefoged, P., Harshman, R., Goldstein, L. and Rice, L. (1978) “Generating vocal tract shapes from formant frequencies”, J. Acous. Soc. of Am., 64, pp. 1027-1035.

Lapedes, A. and Farber, R. (1988), “How Neural networks work”, Technical Report, Los Alamos: NM: Los Alamos National Library. Tech, Rep. LA-UR-88-418, NM, USA.

Laver, J. (1994) “Principles of phonetics”, Oxford University Press., Oxford, UK.

Leggetter, C. and Woodland, P. (1995) “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, Computer, Speech & Lang., Vol. 9, pp. 171-185.

Lindau, M. (1978) “Vowel features”, Language, Vol. 54, pp.541-563.

Lippmann, R. (1997) “Speech recognition by machines and humans”, Speech Communication, Vol. 22, 1–15.

Livescu, K., Bezman, A., Borges, M., Yung, L., Cetin, O., Frankel, J., King, S., Magimai-Doss, M., Chi, X. and Lavoie, L. (2007a) “Manual Transcription of Conversational Speech at the Articulatory Feature Level”, Proc. of ICASSP, Vol. 4, pp. 953-956.

Livescu, K. Cetin, O., Hasegawa-Johnson, M., King, S., Bartels, C., Borges, N., Kantor, A., Lal, P., Yung, L., Bezman, A., Dawson-Haggerty, S., Woods, B., Frankel, J., Magimai-Doss, M. and Saenko, K. (2007b) “Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop”, Proc. of ICASSP, Vol. 4, pp. 621-624.

Lochschmidt, B. (1982) “Acoustic-phonetic analysis based on an articulatory model”, in Automatic Speech Analysis & Recognition, J.P. Hayton eds., (D. Reidel, Dordrecht), pp. 139–152.

Lockwood, P. and Boudy, J. (1991) “Experiments with a non linear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars”, in Proc. of Eurospeech, pp. 79-82.

Ma, J. and Deng, L. (2000) “A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamic model of speech”, Comp., Speech & Language, Vol. 14, pp. 101-104.

Manuel, S.Y. and Krakow, R.A. (1984) “Universal and language particular aspects of vowel-to-vowel coarticulation”, Haskins Lab. Star. Rep, Speech Res. SR-77/78, pp.69-78.

Manuel, S.Y. (1990) "The role of contrast in limiting vowel-to-vowel coarticulation in different languages", J. of Acoust. Soc. Am., Vol.88, pp.1286-1298.

Martinet, A. (1957) "Phonetics and linguistic evolution", Manual of Phonetics, B. Malmberg eds., North-Holland, Amsterdam, pp. 252-272.

McGowan, R.S. (1994) "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: preliminary model tests", Speech Comm., Vol.14, Iss.1, pp. 19-48.

Mermelstein, P. (1973) "Articulatory model for the study of speech production", J. Acoust. Soc. of Am., **53**(**4**), pp.1070–1082.

Metze, F. and Waibel, A. (2002) "A Flexible Stream Architecture for ASR Using Articulatory Features", Proc. of ICSLP, pp. 2133-2136, Denver, CO.

Ming, J. and Smith, F.J. (1998) "Improved phone recognition using Bayesian Triphone Models" Proc. of ICASSP, pp. 409-412.

Mitra, V., Özbek, I., Nam, H., Zhou, X. and Espy-Wilson, C. (2009a) "From Acoustics to Vocal Tract Time Functions", Proc. of ICASSP, pp.4497-4500, Taiwan.

Mitra, V., Nam, H. and Espy-Wilson, C. (2009b) "A step in the realization of a speech recognition system based on gestural phonology and landmarks", 157th meeting of the ASA, J. Acoust. Soc. of Am., Vol.125, pp.2530, 2009.

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E. and Goldstein, L. (2009c) "Noise robustness of Tract Variables and their application to Speech Recognition", Proc. of Interspeech, pp. 2759-2762, UK.

Mitra, V., Borgstrom, B.J., Espy-Wilson, C. and Alwan, A. (2009d) "A Noise-type and Level-dependent MPO-based Speech Enhancement Architecture with Variable Frame Analysis for Noise-robust Speech Recognition", Proc. of Interspeech, pp. 2751-2754, Brighton, UK.

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E. and Goldstein, L. (2010a), "Retrieving Tract Variables from Acoustics: a comparison of different Machine Learning strategies", In press, IEEE J. of Selected Topics on Signal Processing, Special Issue on Statistical Learning Methods for Speech and Language Processing, Vol. 4(6).

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E. and Goldstein, L. (2010b), "Robust Word Recognition using articulatory trajectories and Gestures", Proc. of Interspeech, pp. 2038-2041, Makuhari, Japan.

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E. and Goldstein, L. (to appear) "Tract variables for noise robust speech recognition", accepted for publication in IEEE Trans. on Audio, Speech & Signal Processing.

Mohamed, A., Dahl, G. and Hinton, G. (2009) "Deep Belief Networks for phone recognition", NIPS-22 workshop on Deep Learning for Speech Recognition and Related Applications, Canada.

Moller, M.F. (1993) “A scaled conjugate gradient algorithm for fast supervised learning”, Neural Networks, Vol.6, pp. 525-533.

Muzumdar, M. (1996) “Automatic Acoustic Measurement Optimization for Segmental Speech Recognition”, MS thesis, MIT, Department of Electrical Engineering and Computer Science, Cambridge, MA.

Nam, H., Goldstein, L., Saltzman, E. and Byrd, D. (2004) “Tada: An enhanced, portable task dynamics model in matlab”, J. Acoust. Soc. of Am., **115**(**5**), 2, pp. 2430.

Nam, H., Mitra, V., Tiede, M., Saltzman, E., Goldstein, L., Espy-Wilson, C. and Hasegawa-Johnson, M. (2010) “A procedure for estimating gestural scores from natural speech”, Proc. of Interspeech, pp. 30-33, Japan.

Neiberg, D., Ananthakrishnan, G. and Engwall, O. (2008) “The Acoustic to Articulation Mapping: Non-linear or Non-unique?”, Proc. of Interspeech, pp. 1485-1488, Brisbane, Australia.

Ohman, S.E.G. (1966) “Coarticulation in VCV utterances: Spectrographic measurements”, J. Acoust. Soc. of Am., **39**, pp.151-168.

Okadome, T., Suzuki, S. and Honda, M. (2000) “Recovery of articulatory movements from acoustics with phonemic information”, Proc. of the 5th Seminar on Speech Production, Kloster Seeon, Bavaria, pp. 229–232.

Omar, M.K. and Hasegawa-Johnson, M. (2002) “Maximum Mutual Information Based Acoustic Features Representation of Phonological Features for Speech Recognition”, Proc. of ICASSP, Vol. 1, pp. 81-84.

Ostendorf, M. (1999) “Moving beyond the ‘beads-on-a-string’ model of speech”, Proc. of IEEE Auto. Speech Recog. & Understanding Workshop. Vol.1, pp.79-83, CO.

Ouni, S. and Laprie, Y. (1999) “Design of hypercube codebooks for the acoustic-to-articulatory inversion respecting the non-linearities of the articulatory-to-acoustic mapping”, Proc. of Eurospeech, Vol. 1, pp.141-144, Budapest, Hungary.

Papcun, G., Hochberg, J., Thomas, T.R., Laroche, F., Zachs, J. and Levy, S. (1992) “Inferring articulation and recognizing gestures from acoustics with a neural network trained on X.ray microbeam data”, J. Acoust. Soc. of Am., **92**(**2**), pp. 688-700.

Pearce, D. and Hirsch, H.G. (2000) “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions”, Proc. of Automatic Speech Recognition: Challenges for the new Millenium, ASR-2000, pp. 181-188, Paris, France.

Pols, L.C.W. (1982) “How humans perform on a connected-digits data base”, Proc. of ICASSP, Vol. 2, pp. 867–870.

Pruthi, T. (2007), “Analysis, Vocal-Tract Modeling and Automatic Detection of Vowel Nasalization”, PhD Thesis, Univ. of Maryland, College Park, MD, USA.

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

Rabiner, L., Rosenberg, A., and Levinson, S. (1991) "Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition", IEEE Trans. on ASSP, **26**(**6**), pp.575-582, 1978.

Rahim, M.G., Kleijn, W.B., Schroeter, J. and Goodyear, C.C. (1991) "Acoustic-to-articulatory parameter mapping using an assembly of neural networks", Proc. of ICASSP, pp.485-488.

Rahim, M.G., Goodyear, C.C., Kleijn, W.B., Schroeter, J. and Sondhi, M. (1993) "On the use of neural networks in articulatory speech synthesis", J. Acous. Soc. of Am., 93 (2), 1109–1121.

Recasens, D. (1984) "Timing constraints and coarticulation: Alveolo-palatals and sequences of alveolar + [j] in Catalan", Phonetica, Vol. 41, pp. 125-139.

Richardson, M., Bilmes, J. and Diorio, C. (2003) "Hidden-articulator Markov models for speech recognition", Speech Comm., **41**(**2-3**), pp. 511-529.

Richmond, K. (2001) "Estimating Articulatory parameters from the Speech Signal", PhD Thesis, University of Edinburgh.

Richmond, K. (2007) "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion", Lecture Notes in Comp. Sc., Vol.4885/2007, pp. 263-272.

Ryalls, J. and Behrens, S. J. (2000) "Introduction to Speech Science: From Basic Theories to Clinical Applications", Allyn & Bacon.

Sakoe, H., and Chiba, S. (1978) "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. on Acoust., Speech & Signal Process., **26**(**1**), pp. 43-49.

Saltzman, E. and Munhall, K. (1989) "A Dynamical Approach to Gestural Patterning in Speech Production", Ecological Psychology, **1**(**4**), pp. 332-382.

Schmidbauer, O. (1989) "Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations", Proc. of ICASSP, pp. 616-619.

Schrauwen, B. and Buesing, L. (2009) "A Hierarchy of Recurrent Networks for Speech Recognition", NIPS-22 workshop on Deep Learning for Speech Recognition and Related Applications, Canada.

Shinozaki, T. and Furui, S. (2003) "An Assessment of Automatic Recognition Techniques for Spontaneous Speech in Comparison With Human Performance", ISCA and IEEE workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, pp.95-98

Shirai, K. and Kobayashi, T. (1986) "Estimating articulatory motion from speech wave", Speech Commun. 5, pp.159–170.

Smola, A. and Scholkhopf, B. (2004) “A tutorial on support vector regression”, Statistics & Computing, 14(3), pp.199–222.

Srinivasan, S. and Wang, D.L. (2007) “Transforming Binary Uncertainties for Robust Speech Recognition”, IEEE Trans. Audio, Speech & Lang. Processing, 15(7), pp. 2130-2140.

Steingrimsson, P., Markussen, B., Andersen, O., Dalsgaard, P. and Barry, W. (1995) “From acoustic signal to phonetic features: a dynamically constrained self-organising neural network”, Proc. of International Congress of Phonetic Sciences, Vol. 4, pp.316-319, Stockholm, Sweden.

Stevens, K.N. (1960), “Toward a model for speech recognition”, J. of Acoust. Soc. Am., Vol.32, pp. 47-55.

Stevens, K.N., Manuel, S. and Matthies, M. (1999) “Revisiting place of articulation measures for stop consonants: Implications for models of consonant production”, Proc. of ICPhS, Vol-2, pp. 1117-1120.

Stevens, K.N. (2000a) “Acoustic Phonetics (Current Studies in Linguistics)”, MIT Press.

Stevens, K.N. (2000b) “From acoustic cues to segments, features and words”, Proc. of the International Conference on Spoken Language Processing, Vol. 1, pp. A1–A8, Beijing, China.

Stevens, K.N. (2002) “Toward a model for lexical access based on acoustic landmarks and distinctive features”, J. Acous. Soc. of Am., **111**(**4**), pp. 1872-1891.

Strope, B. and Alwan, A. (1997) “A model of dynamic auditory perception and its application to robust word recognition”, IEEE Trans. Speech & Audio Processing, 5(5), pp. 451-464.

Sullivan, T.M. (1996) “Multi-Microphone Correlation-Based Processing for Robust Automatic Speech Recognition”, Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1996.

Sun, J. and Deng, L. (1998) “Use of High-level Linguistic Constraints for Constructing Feature-based Phonological Model in Speech Recognition”, Journal of Intelligent Information Processing Systems, Vol. 5, No. 4, pp. 269-276.

Sun, J. and Deng, L. (2000a) “Annotation and use of speech production corpus for building language-universal speech recognizers”, Proc. of the 2nd International Symposium on Chinese Spoken Language Processing, ISCSLP, Beijing, Vol. 3, pp. 31-34.

Sun, J., Jing, X. and Deng, L. (2000b) “Data-Driven Model Construction for Continuous Speech Recognition Using Overlapping Articulatory Features”, Proc. of ICSLP, Vol. 1, pp. 437-440.

Sun, J. and Deng, L. (2002) “An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition”, J. Acoust. Soc. of Am., **111**(**2**), pp.1086-1101.

Tang, M., Seneff, S. and Zue, V. (2003) “Two-Stage Continuous Speech Recognition Using Feature-Based Models: A Preliminary Study”, Proc. of IEEE ASRU Workshop, pp. 49-54, U.S. Virgin Islands.

Tepperman, J., Goldstein, L., Lee, S. and Narayanan, S. (2009) “Automatically Rating Pronunciation Through Articulatory Phonology”, Proc. of Interspeech, pp. 2771-2774.

Tiede, M., Perkell, J., Zandipour, M., & Matthies, M. (2001) “Gestural timing effects in the 'perfect memory' sequence observed under three rates by electromagnetometry”, J. Acoust. Soc. of Am, **110** (**5**), pp. 2657.

Toda, T., Black, A.W. and Tokuda, K. (2007) “Voice conversion based on maximum likelihood estimation of speech parameter trajectory”, IEEE Trans. Audio, Speech & Lang. Processing, **15**(**8**), pp.2222-2235.

Togneri, R. and Deng, L. (2003) “Joint State and Parameter Estimation for a Target-Directed Nonlinear Dynamic System Model”, IEEE Trans. on Sig. Processing, **51**(**12**), pp. 3061-3070.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T. (2000) “Speech parameter generation algorithms for HMM-based speech synthesis”, Proc. of ICASSP, Vol.3, pp. 1315-1318, Turkey.

Toutios, A. and Margaritis, K. (2005a) “A Support Vector Approach to the Acoustic-to-Articulatory Mapping”, Proc. of Interspeech, Eurospeech, pp. 3221-3224, Portugal.

Toutios, A. and Margaritis, K. (2005b) “Learning Articulation from Cepstral Coefficients”, Proc. of SPECOM, October, 2005, Patras, Greece.

Van Leeuwen, D.A., Van den Berg, L.G. and Steeneken, H.J.M. (1995) “Human benchmarks for speaker independent large vocabulary recognition performance,” in Proceedings of Eurospeech, Vol. 2, pp. 1461–1464.

Vapnik, V. (1998) “Statistical Learning Theory”, Wiley, New York.

Varga, A. and Steeneken, H.J.M. (1993) “Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems”, Speech Communication, vol. 12, pp. 247–251.

WaveSurfer (2006) [url: http://www.speech.kth.se/wavesurfer/].

Westbury, J. (1994) “X-ray microbeam speech production database user’s handbook”, University of Wisconsin.

Wester, M., Greenberg, S. and Chang, S. (2001) “A Dutch Treatment of an Elitist Approach to Articulatory-Acoustic Feature Classification”, Proc. of Eurospeech, Aalborg, Denmark, pp. 1729-1732.

Wester, M., Frankel, J. and King, S. (2004) “Asynchronous Articulatory Feature Recognition Using Dynamic Bayesian Networks”, Proc. of Institute of Electronics, Information and Communication Engineers Beyond HMM Workshop, Vol. 104, pp. 37–42, SP2004-81-95, Kyoto, Japan.

Weston, J., Gretton, A. and Elisseeff, A. (2003) “SVM practical session - How to get good results without cheating”, Machine Learning Summer School, Tuebingen, Germany.

Windheuser, C., Bimbot, F. and Haffner, P. (1994) “A probabilistic framework for word recognition using phonetic features”, Proc. of ICSLP, pp. 287-290.

Wrench, A. (1999) “MOCHA-TIMIT”,
http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

Xiao, X., Li, J., Chng, E.S., Li, H. and Lee, C. (2010) “A Study on the Generalization Capability of Acoustic Models for Robust Speech Recognition”, IEEE Trans. Audio, Speech & Lang. Processing, Vol. 18, Iss. 6, pp. 1158-1169.

You, H., Zhu, Q. and Alwan, A. (2004) “Entropy-Based Variable Frame Rate Analysis of Speech Signals and its Application to ASR”, Proc. of ICASSP, pp. 549-552.

Yuan, J., and Liberman, M. (2008), “Speaker identification on the SCOTUS corpus”, J. Acoust. Soc. of Am., 123(5), pp. 3878.

Zachs, J. and Thomas, T.R. (1994) “A new neural network for articulatory speech recognition and its application to vowel identification”, Comp. Speech & Language, 8, pp. 189–209.

Zhang, Z. and Furui, S. (2004) “Piecewise-linear transformation-based HMM adaptation for noisy speech”, Speech Comm., Vol. 42, pp. 43-58.

Zhuang, X., Nam, H., Hasegawa-Johnson, M., Goldstein, L. and Saltzman, E. (2008) “The Entropy of Articulatory Phonological Code: Recognizing Gestures from Tract Variables”, Proc. of Interspeech, pp. 1489-1492.

Zhuang, X., Nam, H., Hasegawa-Johnson, M., Goldstein, L., and Saltzman, E. (2009), “Articulatory Phonological Code for Word Classification”, Proc. of Interspeech, pp. 2763-2766, Brighton, UK.