An Analysis-by-Synthesis Approach to Vocal Tract Modeling for Robust Speech Recognition

Submitted in Partial Fulfillment of the Requirements for

the Degree of

Doctor of Philosophy in

Electrical and Computer Engineering

Ziad A. Al Bawab

B.E., Computer and Communications Engineering, American University of Beirut

M.S., Electrical Engineering, University of California Los Angeles

Carnegie Mellon University

Pittsburgh, PA 15213

December 17, 2009

To my parents Hoda Itani and Abdulaziz Al Bawab

Abstract

In this thesis we present a novel approach to speech recognition that incorporates knowledge of the speech production process. The major contribution is the development of a speech recognition system that is motivated by the physical generative process of speech, rather than the purely statistical approach that has been the basis for virtually all current recognizers. We follow an analysis-by-synthesis approach. We begin by attributing a physical meaning to the inner states of the recognition system pertaining to the configurations the human vocal tract takes over time. We utilize a geometric model of the vocal tract, adapt it to our speakers, and derive realistic vocal tract shapes from electromagnetic articulograph (EMA) measurements in the MOCHA database. We then synthesize speech from the vocal tract configurations using a physiologically-motivated articulatory synthesis model of speech generation. Finally, the observation probability of the Hidden Markov Model (HMM) used for phone classification is a function of the distortion between the speech synthesized from the vocal tract configurations and the real speech. The output of each state in the HMM is based on a mixture of density functions. Each density models the distribution of the distortion at the output of each vocal tract configuration. During training we initialize the model parameters using ground-truth articulatory knowledge. During testing only the acoustic data are used.

In the first part of the thesis we describe a segmented phone classification experiment. We present results using analysis-by-synthesis distortion features derived from a codebook of vocal tract shapes. We create a codebook of vocal tract configurations from the EMA data to constrain the articulatory space. Improvements are achieved by combining the probability scores generated using the distortion features with scores obtained using traditional acoustic features.

In the second part of the thesis we discuss our work on deriving realistic vocal tract shapes from the EMA measurements. We present our method of using EMA data from each speaker in MOCHA to adapt Maeda's geometric model of the vocal tract. For a given utterance, we search the codebook for codewords corresponding to vocal tract contours that provide the best fit to the superimposed EMA data on a frame-by-frame

basis. The articulatory synthesis approach of Sondhi and Schroeter is then used to synthesize speech from these codewords. We present a technique for synthesizing speech solely from the EMA measurements. Reductions in Mel-cepstral distortion between the real speech and the synthesized speech are achieved using our adaptation procedure.

In the third part of the thesis we present a dynamic articulatory model for phone classification. The model integrates real articulatory information derived from EMA data into its inner states. It maps from the articulatory space to the acoustic space using an adapted vocal tract model for each speaker and a physiologically-based articulatory synthesis approach. We apply the analysis-by-synthesis paradigm in a statistical fashion. The distortion between the speech synthesized from the articulatory states and the incoming speech signal is used to compute the output observation probability of the Hidden Markov Model (HMM) used for classification. The output of each state in the HMM is based on a mixture probability density function. Each probability density models the distribution of the distortion at the output of each codeword. The estimation algorithm converges to a solution that zeros out the weights of the unlikely codewords for each phone. Hence, each state inherits an articulatory meaning based on these estimated weights and the transition from one state to another reflects articulatory movements. Experiments with the novel framework show improvements in phone classification accuracy over baseline accuracy obtained using a purely statistically-based system, as well as a close resemblance of the estimated weights to ground-truth articulatory knowledge.

To our knowledge this is the first work that applies the analysis-by-synthesis paradigm in a statistical fashion for phone classification. It is the first attempt to integrate realistic speaker-adapted vocal tract shapes with a physiologically-motivated articulatory synthesis model in a dynamic pattern recognition framework. It is also the first work to synthesize continuous speech waveforms solely from EMA measurements and to perform a speaker-independent analysis of highly speaker-dependent EMA phenomena.

Acknowledgments

Words cannot describe how grateful I am to my mentor, Richard M. Stern, for the support and confidence he provided me on this long journey. Rich believed in my ability to perform genuine research from day one. He assigned a very challenging research problem to me and after six years, together, we came up with this thesis. Rich was more than an advisor to me. I enjoyed learning from his wisdom and experience in life as much as I enjoyed learning and deeply understanding the basic issues related to signal processing and speech recognition from him. I enjoyed meeting with Rich on a weekly basis and discussing progress on research. In addition, I enjoyed discussing with Rich different subjects about academia, life, careers, politics, cultures, and many more. He was always available to chat.

Bhiksha Raj has been a big brother to me that I turned to for inspiration and motivation at all times. Most of the ideas you see in this thesis have stemmed from discussions with him. This work could not become a reality without Bhiksha's help. Bhiksha's positive attitude to life and research problems is unique. Whenever I felt down, Bhiksha was there to motivate me to pursue my ideas. Bhiksha was always available for discussions. We started this work when Bhiksha was in Boston working for Mitsubishi Electric Research Labs (MERL) using Skype and Powerpoint. I am very grateful to Bhiksha's flexibility and availability. He is a main architect of this thesis.

Bhiksha Raj and Rita Singh have been a great help to me on the speech recognition problem I wanted to solve. Together they are an encyclopedia on ideas related to speech recognition and have contributed to this field for more than a decade now. I feel lucky to have them around at Carnegie Mellon University (CMU) during my stay.

I am also grateful to Lorenzo Turicchia and to Sankaran Panchapagesan (Panchi) for their collaboration. Their expertise on articulatory synthesis and geometric modeling was very helpful to this thesis. Panchi's help in forwarding the articulatory synthesis package was the basic starting point for this work.

In addition to Rich, the chair of my PhD committee, and to Bhiksha, I would also like to thank the other members of the committee Alan W. Black and Tsuhan Chen for

their time and collaboration on this piece of work.

I have enjoyed the time I spent at CMU with my robust speech lab colleagues. I am thankful to Xiang Li, Evandro Gouvea, Yu-Hsiang Chiu (Bosco), Lingyun Gu, Chanwoo Kim, and Kshitiz Kumar for their nice company in the basement of Porter Hall. I am also grateful to discussions with Amr Ahmad, Sajid Siddiqi, Usman Khan, Mike Seltzer, Ravishankar Mosur, David Huggins-Daines, and Arthur Toth.

On the administrative side, I am very thankful to Carol Patterson, Elaine Lawrence, and Claire Bauerle for their continuous help.

My family has been great motivator for me to complete this work. I am very grateful to my mother Hoda and my father Abdulaziz for their love. I am also lucky to have my brother Mohammed and sister Amani's support all the time. I can not also forget the support of my uncles Salim and Bilal who were always there when I needed their help and guidance.

My friends were the unknown soldiers behind my progress and success. I am grateful to my big brothers Abdulkader Sinno, Basil AsSadhan, and Mohammed Noamany's continuous support and advice. I have enjoyed greatly the company of my best friends Tarek Sharkas, Firas Sammoura, and Ahmad Nasser. I am also grateful to all the members of the Muslim Students Association at CMU. They were all part of my other family in Pittsburgh.

Finally, I am very proud to be a member of an elite group of researchers at this small university in Pittsburgh that is revolutionizing the world.

The work in this thesis was supported by the National Science Foundation (NSF) grant IIS-0420866 and the Defense Advanced Research Projects Agency (DARPA) grants NBCH-D-03-0010 and W0550432.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Human speech is a product of several physiological parts working jointly to convey an acoustic message from the speaker to the listener. The main components of the speech production system are the lungs and glottis comprising the source function and the vocal tract comprising the filter function.

It has long been known that the physics of the vocal tract greatly constrains the set and sequence of sounds that a person may produce. The vocal tract comprises several mechanically coupled components, including articulators such as the lips, jaw, tongue, palate, and velum, in addition to the various bones, cartilages, and other soft tissues that affect sound production. Each of these has its own dynamics and physical characteristics such as compliance, resonance, mass, inertia, momentum, etc. These impose restrictions both on the static configurations that the vocal tract can take, as well as on the dynamics of the vocal tract itself, which constrains the set and sequences of sounds that a person may utter.

Automatic speech recognition (ASR) is the process of recognizing human speech by machines. State-of-the-art ASR systems do not explicitly or implicitly account for the restrictions mentioned above. Instead, the speech signal is usually treated in an entirely phenomenological manner: features that are derived for speech recognition are based on measurements of the spectral and temporal characteristics of the speech signal [3] without reference to the actual physical mechanism that generates it. Even spectral

estimation techniques such as linear predictive coding (LPC), that provide an empirical characterization of the output of the vocal tract, do not directly access the physics of the generating process – the relationship between the parameters that are estimated and the vocal tract is chiefly one of analogy.

In addition, the physical constraints that guide the articulators' movements over time give rise to various phenomena observed in speech production. Coarticulation occurs due to the effect of context on the current sound unit, or more accurately, on the current articulatory configuration since articulators move smoothly. Sloppiness (reduced effort) in spontaneous speech and faster speaking rates cause the articulators to miss their intended "targets" for a particular sound unit. Hence articulatory target undershoot (or overshoot) is also common, which gives rise to different acoustic observations. Asynchrony in the articulators' movements causes different articulatory configurations to overlap in time giving rise to pronunciation variations. The fact that some articulators are "free" to take the path of least resistance compared to "critical" articulators that should develop a specific configuration pertaining to a particular phone also leads to pronunciation variations. All of these high-level phenomena hinder the performance of ASR systems especially during conversational and unrehearsed speech.

The basic sound unit defined in speech recognition systems is the phone. Each phone corresponds to a particular acoustic distribution. Vocabulary entries are modeled using non-overlapping sequences of these abstract segmental units [4, 5]. This framework, shown in Figure 1.1, and the currently-used features cannot account for the phenomena observed above in spontaneous speech nor can they account for the physical constraints of speech production. For example, a slight variation in the articulatory configurations pertaining to a particular phone would cause a variation in the corresponding acoustic observation, which in turn may change the identity of the phone recognized. Hence, a new framework that incorporates *physical constraints* in the *articulatory space* would be expected to achieve more robust speech recognition performance.

1.1 Thesis Objectives and Framework

In this thesis we present a novel approach for speech recognition that incorporates knowledge of the speech production process. We discuss our contributions in moving from a purely statistical speech recognizer to one that is motivated by the physical generative process of speech. This process incorporates knowledge of the physical instantaneous and dynamic constraints of the vocal tract and knowledge of the various phenomena observed in speech production such as coarticulation, pronunciation variation, sloppiness, speaking rate, etc.

To achieve this, we will utilize an *analysis-by-synthesis* approach that is based on an explicit mathematical model of the vocal tract to represent the physics of the sound production process (*synthesis*) and constrain a statistical speech recognition system using this model. In addition, we will incorporate instantaneous and dynamic constraints on the articulatory configurations (*analysis*).

The physics and fluid dynamics of the vocal tract have been well studied and the equations characterizing them have been derived from first principles by a number of researchers including Flanagan [6] and Stevens [7]. A variety of mathematical models of the vocal tract that embody these equations has been proposed in the literature,

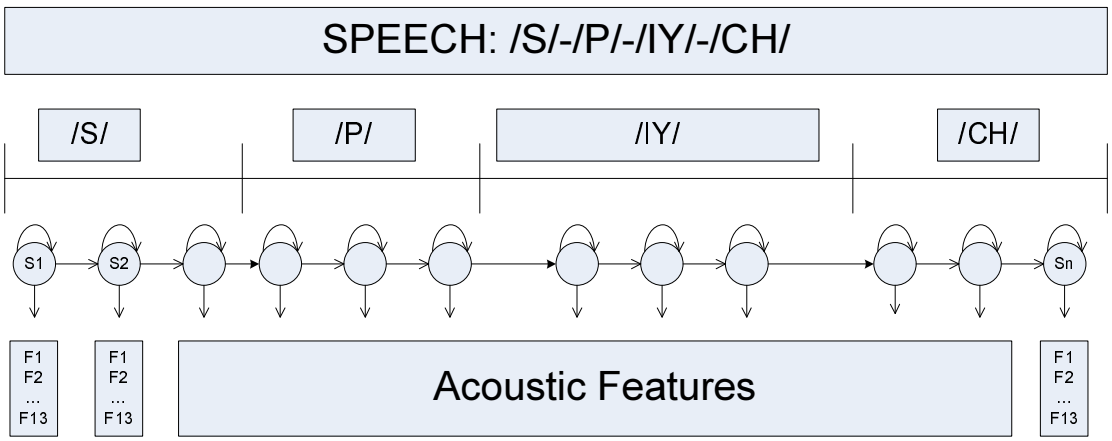


Figure 1.1: *Current concatenative framework for the word "SPEECH" as pronounced in cmudict. Each phone is modeled by a three-state HMM, and the acoustic observations are represented by 13-dimensional feature vectors.*

e.g. by Sondhi and Schroeter [8]. Although the actual number of free parameters in these equations is large, most sounds can be geometrically specified using only the seven parameters of Maeda's model [9], which describes the relevant structure of a capital section of the vocal tract as will be discussed in Section 2.2. Simple manipulations of these seven parameters can produce a variety of natural-sounding speech sounds.

We hypothesize that implicit knowledge of articulatory configurations would yield better speech recognition accuracy. Given a particular configuration of a vocal tract and information regarding the excitation to the tract, the models described above are able to synthesize plausible sounds that would have been produced by it. Conversely, by matching the signal synthesized by the model for a particular configuration to a given speech signal, it becomes possible to gauge the likelihood that a particular configuration might have produced that speech, or, alternatively, to obtain a metric for the "distance" between a given configuration and the one that actually produced the speech signal. The known relationships between articulator configurations and sounds will be utilized in conjunction with the rules of physics and physiology that constrain how fast articulators can move to generate sounds. By appropriate selection of the parameters of the equations that govern these motions, the effects of speech rate, spontaneity, speaker effects, accents, etc., can be systematically modeled. Unlike conventional speech recognition systems that discard the excitation function associated with the waveform and retain only the signal's coarse spectral shape, we will explicitly use the excitation signal extracted from the incoming speech, mimicking it closely. This model is described in Figure 1.2.

In the first part of the thesis, we account for the "instantaneous" or short-term physical generative process of speech production. We devise a technique to extract new articulatory features using the analysis-by-synthesis framework. We characterize the space of vocal tract configurations through a carefully-chosen codebook of configuration parameters in which each codeword is composed of a vector of Maeda parameters. Maeda uses seven parameters to represent a vocal tract shape. We derive these configurations from Electromagnetic Articulograph (EMA) data available in the MOCHA database [2]. MOCHA contains positional information from sensors placed on the speakers' lips,

incisors, tongue, and velum in addition to the speech recorded as the speakers read TIMIT utterances. Using a heuristic mapping that is independent of the model, the EMA measurements are converted to a Maeda parameters. Using Maeda's geometric model of the vocal tract, we compute the areas and lengths of the tubes model forming the vocal tract. Sondhi and Schroeter's articulatory synthesis model is used to compute vocal tract transfer functions of each of the configurations in the codebook and to excited them by a source function whose parameters (energy, pitch, etc.) are derived from the incoming speech signal. The synthesized speech signals for each configuration are compared to the actual incoming signal to obtain a vector of distances, each dimension of which represents the distance of the current signal to one of the configurations. The sequence of distance vectors obtained in this fashion represents the trajectory of the articulator configurations for the signal. After dimensionality reduction, the sequence of vectors are modeled by a statistical model used as supporting evidence and combined with a model based on conventional Mel-frequency cepstral coefficients (MFCCs). A fast analysis-by-synthesis approach is also developed in this part.

In the second part of the thesis, we present a method for adapting Maeda's model to the EMA data and we derive realistic vocal tract shapes from the measurements.

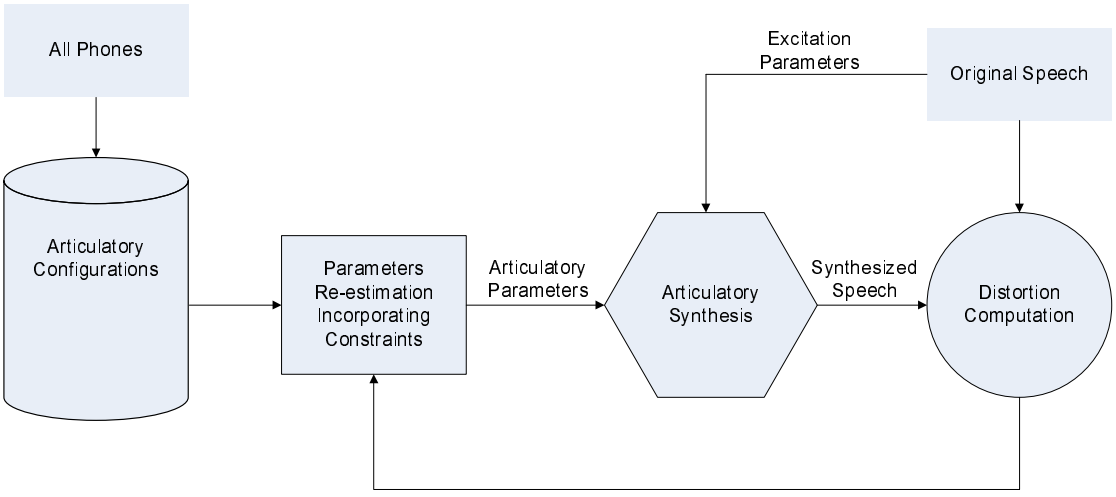


Figure 1.2: *Analysis-by-synthesis framework mimicking incoming speech by estimating the vocal tract parameters and using the original signal excitation parameters.*

We thus present a more principled approach for mapping the EMA data to Maeda parameters. Using the ensemble of all measurements for each speaker, we obtain the scatter of the distribution whose upper outline is used in the adaptation. For a given utterance, the algorithm we developed searches the codebook for the codewords whose vocal tract contours best fit the superimposed EMA data, on a frame-by-frame basis. Next, the area functions of the corresponding vocal tract shapes are computed. The articulatory synthesis approach of Sondhi and Schroeter is then applied to synthesize speech from these area functions. We decouple the source model from the transfer function, which improves the quality of synthesis and helps speed up the approach. We have thus presented a technique for synthesizing speech solely from EMA data and without any statistical mapping from EMA to acoustic parameters.

In the third part, we impose physical constraints on the dynamic configurations of the vocal tract and utilize a data-driven approach to learn these constraints in a dynamic framework. Each state will represent a combination of codewords, which will be structured in a way that attributes physical meaning to the state. For example, in the Hidden Markov Model (HMM) framework, the states will resemble different articulatory configurations rather than abstract segmental units like phones as they currently do. Statistical dependencies between the states, that capture dynamic relationships (similar to transition probabilities and state-dependent symbol probabilities in HMMs), will be learned through a maximum likelihood approach. By comparing the signal synthesized by applying the vocal tract parameter values represented by the codewords to the mathematical model of the vocal tract to the incoming speech signal, we derive a "synthesis error". The "synthesis error" and the ground-truth articulatory information will guide the propagation through the articulatory space and help learn the codeword weights and state transitions as we will explain in Chapter 5.

In summary, we will incorporate models of the vocal tract directly into the recognition process itself. This thesis includes three main contributions. The first contribution is a feature-based approach that attempts to capture the location of the articulatory configuration for a given frame in the entire space of vocal tract configurations through

a feature vector. The second contribution is a procedure for deriving realistic vocal tract shapes from EMA measures using an adapted geometric model of the vocal tract. The third contribution is a dynamic articulatory model for phone classification that explicitly represents articulatory configurations through its states.

1.2 Thesis Outline

In this chapter we introduced at a high level the problem we are solving and the solutions we develop in this thesis. In Chapter 2 we present the basic background the reader of this thesis needs. We discuss Maeda's model, the MOCHA EMA data, and the Sondhi and Schroeter articulatory synthesis approach.

In Chapter 3 we discuss our first main contribution, deriving the analysis-by-synthesis distortion features. We present phone classification results and an approach for deriving fast analysis-by-synthesis features.

In Chapter 4 we discuss our second contribution, deriving realistic vocal tract shapes from EMA data. We describe how we adapt Maeda's geometric model of the vocal tract to the EMA data and then search for vocal tract shapes on a frame-by-frame basis. We also describe how we can synthesize speech based on EMA data only.

In Chapter 5 we discuss our third contribution, using the analysis-by-synthesis distortion features in a probabilistic framework. We integrate the adapted vocal tract shapes, the adapted transfer functions, and analysis-by-synthesis features into the Hidden Markov Model (HMM) used for phone classification.

In Chapter 6 we propose future research directions and in Chapter 7 we conclude our work and summarize our contributions.

Chapter 2

Background Review

In the past few decades there has been a plethora of scientific research analyzing the speech production process for speech synthesis, speech recognition, and other speech disciplines. Acoustic phonetics [7] and speech analysis [6] are mature fields now. As mentioned in Chapter 1, state-of-the-art speech recognition systems use a small fraction of this knowledge and rely heavily on statistical modeling of abstract speech units. They model speech as a concatenation of non-overlapping segmental units referred to as phones. This framework makes it hard to account for the various phenomena discussed above. These phenomena occur in the *articulatory space* and their effects are observed in the *acoustic space*.

Gesturalists argue for the gesture (articulatory configuration) as the basic unit of speech production and perception. On average, a human produces 150-200 words a minute, which corresponds to 10-15 units per second. It would be impossible for humans to achieve this rate of sound production and maintain speech intelligibility if we follow a concatenative approach. Coarticulation is viewed as the means to achieve this rate in perception and production [10]. For example, a rounded /s/ indicates to the listener that the next vowel /u/ would be rounded like in "student" and cause the speaker to assimilate the roundness of the lips while producing /s/ to minimize production effort.

Liberman and Mattingly [11] in their "motor theory" proposal hold that the articulatory space is essential in retrieving the symbolic sequence (phones) from the acoustic

signal and that the articulatory gesture is the basic unit of perception. The mapping from the symbolic units to the gestures can be invariant. The articulatory targets for a particular phone are known. Nevertheless, achieving them depends on speaking effort, rate, style, etc., which in turn depends on the physics of the articulators and their ability to evolve from one target to another. Add to this the fact that the evolution of articulators is not synchronized, neither within nor between phone segments. Therefore, there will not be a clear correspondence between the segmentation in the phonetic and the acoustic domains. Hence, the mapping between the symbolic sequence and the observed acoustics can be arbitrary complex [12] without an intermediate articulatory space such as the one shown in Figure 2.1. In essence, control over articulatory parameters is more direct than control over acoustic observations since they resemble control over muscle movements. Another theory that argues for an overlapping gestural representation is the articulatory phonology theory developed by Browman and Goldstein [13].

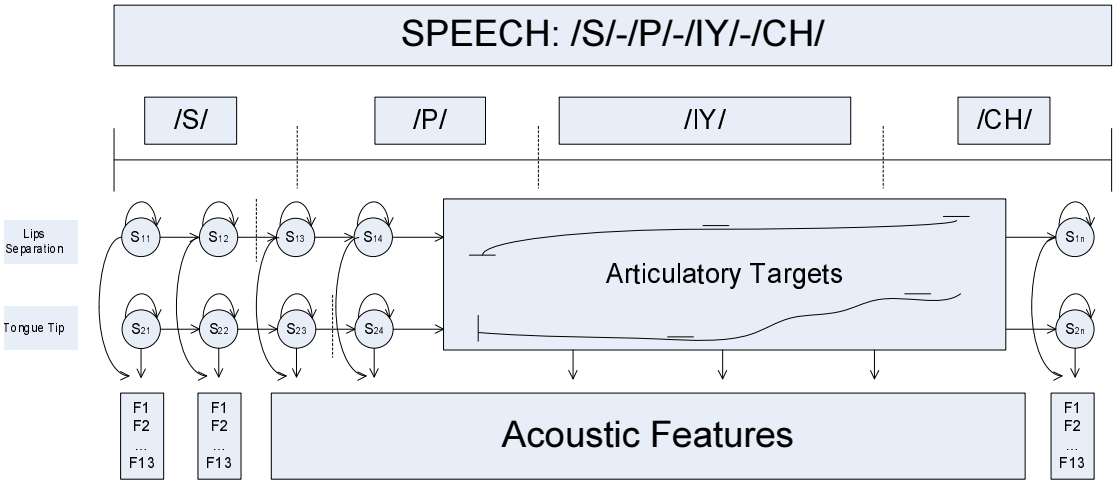


Figure 2.1: *A hypothetical framework allowing articulators to flow independently and asynchronously. The horizontal dashes represents critical articulatory targets that are neither totally achieved nor synchronized in time. The vertical dash represents a free articulatory target.*

2.1 Articulatory Modeling for Speech Recognition

Recently, researchers have started incorporating articulatory modeling into speech recognition systems for many of the reasons mentioned above. Speech production information is added in the form of *phonological features* or *articulatory parameters*, as we choose to distinguish between the two in this thesis. Phonological features are rather discrete phone related features *e.g. manner* (vowel or fricative), *place* (high or front), *voicing*, and *rounding*. Articulatory parameters are continuous measurements of the articulatory positions.

Articulators can be divided into three groups based on their role in producing each phone: *critical*, *dependent*, and *redundant* articulators [14]. A critical articulator is one whose configuration is essential for the production of the current sound, *e.g.* the position of the tongue tip in producing /s/. Dependent articulator configurations are affected by the configuration of the critical articulators while the redundant articulators are free to take any position and can be considered to be in a "do not care" state. Papcun *et al.* [15] have showed using x-ray microbeam data that articulators which are considered free for a particular phone have much higher variance compared to those that are critical. On the other hand, Maeda [9] has shown that articulators can compensate for each other acoustically.

Erler and Freeman [12] proposed an articulatory feature model (AFM) for speech recognition, which is an HMM-based system with states that resemble phonological configurations, attributing a physical meaning to them. They introduced three feature evolution constraints. *Dynamic* constraints mimic the limitations of the articulatory system such as the rate of change, maximum movement, and how closely a target is approached or skipped. *Instantaneous* constraints allow only the configurations that are physically realizable. *Propagation* constraints control the range of propagation rules set on articulatory movements over a period of time and the configurations of the "underspecified" or free articulators. In a similar approach, Livescu [16] used dynamic Bayesian networks (DBNs) to model asynchronous phonological feature streams.

Analysis-by-synthesis approaches have previously been applied to speech recognition.

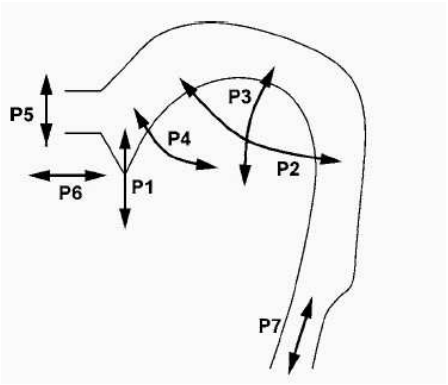
Blackburn [17] used an articulatory codebook that mapped phones generated from N-best lists to articulatory parameters. He linearly interpolated the articulatory trajectories to account for coarticulation and used artificial neural networks (ANNs) to map these trajectories into acoustic observations. Each hypothesis was then rescored by comparing the synthesized features to the original acoustic features. Deng [18] used Kalman filters to smooth the hidden dynamics (represented by vocal tract resonances (VTRs) or pseudo formants) generated using the hypothesis segmentation, accounted for coarticulation by smoothing the trajectories, and introduced different statistical mappings from VTRs to acoustics for rescoring.

Other attempts at incorporating real articulatory measurements have used DBNs [19, 20] to model jointly the articulatory and acoustic distributions and have also used linear dynamic models (Kalman filter) [21]. For a thorough review of the literature on incorporating articulatory information into statistical systems for speech recognition, the reader is referred to [22].

2.2 Articulatory Speech Synthesis

Modeling the articulatory dynamics and relying on statistical mapping from the articulatory states to acoustic observations has been problematic for the approaches described above. It is hard to align articulatory states with acoustic parameters since ground truth is usually not available. Even if parallel data were available, it would not be enough to learn a distribution that can generalize to unseen configurations or even different speakers. We propose the use of articulatory synthesis to compute the observation, or what statisticians refer to as the *emission model*.

Maeda's model [9] uses seven parameters to describe the vocal tract shape and to compute the cross-sectional areas of the acoustic tubes used to model speech generation. Using a factor analysis of 1000 frames of cineradiographic and labiofilm data, Maeda derived a representation of the vocal tract profile as a sum of linear basis vectors or components in a semipolar coordinate space spanning the midsagittal plane of the vocal tract. These components are described in Figure 2.2.



Parameter	Description	Movement
p1	jaw position	vertical
p2	tongue dorsum position	forward or backward
p3	tongue dorsum shape	roundedness
p4	tongue tip	vertical
p5	lip height	vertical
p6	lip protrusion	horizontal
p7	larynx height	vertical

Figure 2.2: *Maeda parameters describing the geometry of the vocal tract. Figure extracted from [1].*

Maeda's model converts each vector of articulatory configurations to a vector of areas and lengths of the sections of the acoustic tube describing the shape of the vocal tract. The Sondhi and Schroeter model [8] uses the chain matrices approach to model the overall transfer function of the vocal tract. Specifically, the transfer function of each section is modeled by a matrix whose coefficients depend on the area and length of the section and on the loss parameters. The input (and output) of the matrix is the pressure and volume velocity in the frequency domain. The transfer function represents the wave equation at each section. The overall transfer function is the product of the matrices. The Sondhi and Schroeter model also allows for nasal tract coupling to the vocal tract by adjusting the velum opening area.

The glottal source and interaction with the vocal tract is modeled in the time domain using the two-mass model of vocal cords developed by Ishizaka and Flanagan [23]. The parameters of this model are the lung pressure Ps, the glottal area A0, and the pitch factor Q. The overall transfer function must be excited in order to generate speech.

2.3 The MOCHA Database

The recent availability of databases such as MOCHA [2], which consists of a set of real articulatory measurements and the corresponding audio data, opens new horizons for better understanding of articulatory phenomena and for further analysis and modeling

of the speech production process.

The MOCHA database contains data from 40 speakers reading 460 TIMIT utterances (in British English). The articulatory measurements include electromagnetic articulograph (EMA), electroglottograph (EGG), and electropalatograph (EPG) measurements. The EMA channels include (x, y) coordinates of nine sensors directly attached to the lower lip (LL), upper lip (UL), lower incisor (LI), upper incisor (UI), tongue tip (TT), tongue body (TB), tongue dorsum (TD), soft palate (velum, VL), and the bridge of the nose.

2.4 Discussion

The work described above which models phonological feature streams is limited by the inherent assumptions that were made. The use of discrete, quantized, and abstract phonological features makes it hard to incorporate transition rules and leads to weak dynamic modeling. The mapping from the phone to its corresponding phonological target is not accurate. It depends on canonical information available in linguistic sources and not on the acoustic observation or the task at hand. The authors of this approach mention the lack of real articulatory measurements parallel to acoustic data which would have provided better initialization of the models. It would also allow for learning of realistic instantaneous and dynamic constraints rather than using rule based ones. In addition, using phonological features does not allow for partially achieved targets (*i.e.* accounting for target undershoot or overshoot).

Researchers using real articulatory features model a joint probabilistic distribution of the articulatory measurements and the acoustic parameters instead of using a physical model of how the acoustics are generated from the articulatory configurations [19, 20]. Hence their approach can only work on limited databases with parallel data and will not generalize to unseen articulatory configurations.

All of approaches mentioned above are *phenomenological*. They attempt to apply constraints based on inferences from observed phenomena. In contrast to these methods, the approach we will follow is a true analysis-by-synthesis technique that actually models

the articulatory configurations and movements of the vocal tract and synthesizes speech based on the physics of sound generation. Consequently, it provides the framework to model various phenomena like speaking rate and stress explicitly through their effect on the parameters and the dynamics of the explicit mathematical representation of the vocal tract.

The recent availability of databases like MOCHA provides many advantages that were not available before. With real articulatory data, realistic instantaneous and dynamic constraints can be learned and can help bootstrap the dynamic models. In addition, better synthesis techniques can be devised from the articulatory information to the acoustic data since the ground truth is known. While such approaches were not considered feasible in the past due to computational considerations, modern computers now make it feasible to incorporate highly computationally-intensive physical models of synthesis into the recognition process.

Chapter 3

Analysis-by-Synthesis Features

In this chapter we account for the "instantaneous" or short-term physical characteristics of the physical generative process. We want to answer the questions of what the possible vocal tract configurations are and how sound is generated from these configurations. We attempt to answer these questions through a feature computation process.

The first step in the analysis-by-synthesis approach to is to derive the features that will be used for recognition. These features would convey information about the articulatory configurations. Since the only observation we have is the acoustic signal, these features implicitly convey information about the articulatory space through an acoustic distance measure between the synthesized and incoming speech.

In our first attempt in deriving these features, we characterize the space of vocal tract configurations through a carefully-chosen codebook of configuration parameters. Vocal tract models for each of the configurations are excited by a source signal derived from the incoming speech signal. The synthesized speech signals from each vocal tract configuration are compared to the actual incoming signal to obtain a vector of distances, each dimension of which represents the distance of the current signal from one of the configurations. The sequence of distance vectors thus obtained represents the trajectory of the articulator configurations for the signal. After dimensionality reduction using LDA, the sequence of vectors is modeled by a statistical model such as a GMM used as supporting evidence and combined with a conventional MFCC-based GMM for a

segmented phone recognition task.

3.1 Generating an Articulatory Codebook

We use the Maeda parameters described in Figure 2.2 as a seven-dimensional representation of vocal tract configurations. EMA measurements from the MOCHA database described in Section 2.3 are converted to these seven-dimensional vectors. To do so, we have developed a model-independent geometric mapping from the EMA measurements to Maeda parameters. By model-independent we mean that we do not superpose the EMA data onto Maeda's model to find the best matching parameters as we do in Chapter 4. For p1 we compute the distance between the lower and upper incisors. For p2, we use the horizontal distance between the tongue dorsum and the upper incisor. For p3 we compute the angle between the line joining the tongue tip and the tongue body, and the line joining the tongue body and the tongue dorsum. For p4 we compute the vertical distance between the upper incisor and the tongue tip. For p5 we compute the distance between the upper and lower lips. For p6 we compute the distance between the midpoint of the upper and lower incisors and the line joining the upper and lower lips. Since we are only using the EMA data, we set p7, which pertains to the larynx height, to zero in the rest of the experiments. These parameters are then normalized using their mean and variance, per utterance, to fall within the [-3,+3] range that is required by Maeda's model. Using a linear mapping, the mean plus standard deviation value is mapped to +2 and the mean minus standard deviation is mapped to -2. The regions mapped outside the [-3,+3] range are clipped. We use the energy in the audio file to set the starting and ending time of the normalization. This way we exclude the regions where the EMA sensors are off from the steady state position before and after the subject is moving his or her articulators. Table 3.1 summarizes this mapping procedure.

Once all measured articulatory configurations are converted to their corresponding Maeda equivalents, we compute a codebook of articulatory parameters. Since p7 is not measured, we do not consider it in this process. The EMA data, and hence the derived Maeda parameters, are aligned with the audio data. To cancel out effects of

Table 3.1: *Model-independent approach for mapping EMA to Maeda parameters.*

Maeda Parameters	Parameters Control	Derivation from EMA
p_1	jaw vertical position	distance(LI, UI)
p_2	tongue dorsum position	distance(UIx, TDx)
p_3	tongue dorsum shape	angle([TT TB], [TB TD])
p_4	tongue tip position	distance(UIy, TTy)
p_5	lip height	distance(UL, LL)
p_6	lip protrusion	distance (UI,[UL LL])
p_7	larynx height	zero

varying speech rate and phone length on the set of available articulatory configurations, we sample the sequence of Maeda parameter vectors to obtain exactly five vectors from each phone. To do so, the boundaries of all phones in the data must be known. In our work these are obtained by training a speech recognizer (the CMU Sphinx system) with the audio component of the MOCHA database and forced-aligning the data with the trained recognizer to obtain phone boundaries.

We sample each phone at five positions: the beginning, middle, end, between beginning and middle, and between middle and end, and read the corresponding Maeda parameter vectors. We perform k-means clustering over the set of parameter vectors obtained in this manner. We designate the vector closest to the mean of each cluster as the codeword representing the cluster. This is done to guarantee that the codeword is a legitimate articulatory configuration. The set of codewords obtained in this manner is expected to span the space of valid articulatory configurations.

3.2 Deriving Articulatory Features

Once a codebook spanning the space of valid articulatory configurations is obtained, it is used in an analysis-by-synthesis framework for deriving a feature vector.

For each incoming frame of speech, a corresponding frame of speech is generated by the synthesis model for the articulatory configuration defined by each codeword.

Thus there are as many frames of speech synthesized as there are codewords in the codebook. Each frame of synthesized speech is compared to the incoming signal to obtain a distortion value[1]. We use the Mel-cepstral distortion (MCD), as defined in Equation 3.1, between the incoming and synthesized speech as the distortion metric, where $\mathbf{c}$ is the vector of MFCCs.

$$MCD(\mathbf{c}_{incoming}, \mathbf{c}_{synth}) = \frac{10}{\ln 10}\sqrt{2\sum_{k=1}^{12}(c_{incoming}(k) - c_{synth}(k))^2} \tag{3.1}$$

The set of distortion values effectively locates the signal in the articulatory space. A vector formed of the distortion values thus forms our basic articulatory feature vector. The process of creating articulatory feature vectors is shown in Figure 3.1.

The articulatory feature vector obtained in this manner tends to be high-dimensional – it has as many dimensions as codewords. Its dimensionality is then reduced through Linear Discriminant Analysis (LDA). Other linear or non-linear dimensionality reduction mechanisms may also be employed.

3.3 Experiments and Results

We conducted a number of experiments to evaluate the usefulness of the proposed articulatory feature extraction method for speech recognition. In order to avoid obfuscating our results with the effect of lexical and linguistic constraints that are inherent in a continuous speech recognition system, we evaluate our features on a simple phone classification task, where the boundaries of phones are assumed to be known. All classification experiments are conducted using simple Gaussian mixture classifiers.

We choose as our data set the audio recordings from the MOCHA database itself, since it permits us to run "oracle" experiments where the exact articulatory configurations for any segment of sound are known. Of the 40 speakers recorded in MOCHA,

[1]We use the implementations of Maeda's model and the Sondhi and Schroeter model provided with the articulatory synthesis package developed by Riegelsberger [24]. The later work in this thesis is based on a modifications of these models.

data for only ten have been released. Of the ten, data for three have already been checked for errors. We checked the data from the remaining seven speakers ourselves and retained a total of nine speakers for our work: "faet0", "falh0", "ffes0", "fjmw0", "fsew0", "maps0", "mjjn0", "msak0", and "ss2404". Five of the speakers are females and four are males. We checked the EMA, the audio, and the corresponding transcript files for the nine speakers. We discarded the utterances that had corrupted or missing EMA channels, corrupted audio files, or incorrect transcripts. We ended up with 3659 utterances, each around 2-4 secs long. We chose to test on the female speaker "fsew0" and the male speaker "maps0" and train on the rest. All experiments are speaker independent. The amount of training utterances is 2750 and testing utterances is 909. Only data from the training speakers were used to compute the articulatory codebook. The codebook consisted of 1024 codewords after clustering 425673 articulatory vectors that were sampled from all the phones. The articulatory data of the test speakers have not been used. The total number of phone-segments used in classification is 14310 for speaker "fsew0", 14036 for speaker "maps0", and 28346 for both speakers.

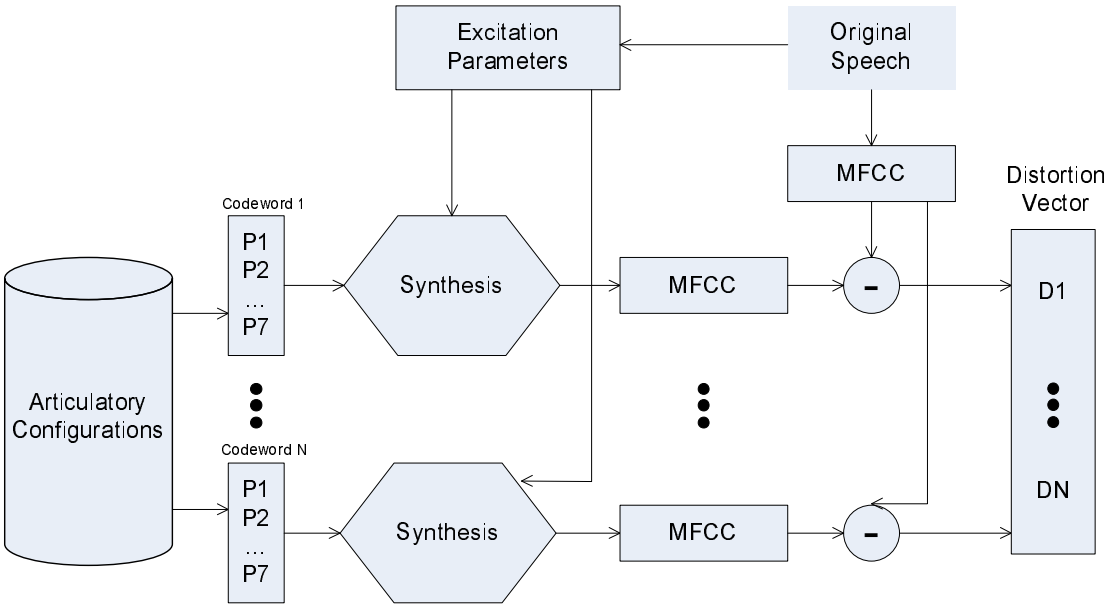


Figure 3.1: *Framework for deriving the analysis-by-synthesis distortion features. Only two codewords are shown explicitly in the illustration.*

In all experiments in this section the audio signal was represented as 13-dimensional MFCC vectors. We trained a Gaussian mixture density with 64 Gaussians to represent each phone. Cepstral mean normalization (CMN, Atal [25]) was applied. No first-order or second-order derivatives were used as they were not found to be useful within the GMM framework.

3.3.1 An Oracle Experiment

We begin with an oracle experiment assuming that the exact articulatory configuration (expressed as a vector of Maeda parameters) for each frame of speech is known. We obtain it directly from the EMA measurement for the frame. The articulatory feature (AF) vector for any frame of speech is obtained simply by computing the Mahalanobis distance between the known Maeda parameter vector for the frame and each of the 1024 codewords in the codebook. The variances used in the distance are computed for each cluster after the clustering stage. We reduce the dimensionality of the resultant 1024-dimensional vectors to 20 dimensions using LDA. A mixture of 32 Gaussians is trained to represent the distribution of these 20-dimensional vectors for each phone. The phone $\hat{C}$ for any segment is estimated as:

$$\hat{C} = \text{argmax}_C P(C)P(MFCC|C)^{\alpha}P(AF|C)^{(1-\alpha)} \tag{3.2}$$

where C represents an arbitrary phone, and $MFCC$ and AF represent the set of MFCC features and articulatory features for the segment respectively. α is a positive number between 0 and 1 that indicates the relative contributions of the two features to classification. We varied the value of α between 0 and 1.0 in steps of 0.05, and chose the value that resulted in the best phone error rate (PER). The classification results and the optimal value of α are shown in Table 3.2.

We note that feature vectors obtained with oracle knowledge of the vocal tract configuration can result in significant improvements in classification performance in combination with MFCCs, although by themselves they are not very effective.

Table 3.2: *PER using MFCC, AF based on oracle knowledge of articulatory configurations, and a combination of the two features.*

Features (dimension)	fsew0	maps0	Both
MFCC + CMN (13)	64.2%	68.1%	66.1%
AF + LDA (20)	77.5%	85.8%	81.6%
Combination ($\alpha = 0.85$)	55.2%	62.9%	59.0%
Relative Improvement	14.0%	7.7%	10.8%

3.3.2 Synthesis with Fixed Excitation Parameters

As explained in Section 3.2, the articulatory feature vector is computed as the vector of Mel-cepstral distortions between the speech signal and the signals generated by the Sondhi and Schroeter model of the vocal tract. The latter, in turn, requires the vocal tract to be excited. In this experiment we assume that the excitation to the synthetic vocal tract is fixed, *i.e.* the synthesis is independent of the incoming speech itself. This may be viewed as a worst-case scenario for computing features by analysis-by-synthesis.

In this experiment we fixed the excitation parameters [Ps, A0, Q] described in Section 2.2 to the values of [7,0.05,0.9] for voiced excitation and [7,0.15,0.7] for unvoiced excitation. Since the synthesis was independent of the incoming signal, two MFCC vectors were generated from each codeword, one from each excitation. Both synthetic MFCCs were compared to the MFCCs of the incoming speech. Since the energy level in the synthesized speech is the same for all codewords, $c(0)$ (zeroth cepstral term) was not considered when computing the distortion. Since two distortion values were obtained from each codeword, the final articulatory distortion feature (Dist Feat) vector has 2048 dimensions that were reduced to 20 dimensions using LDA.

The rest of the details of the experiment, including the specifics of dimensionality reduction, distributions estimated, and likelihood combination were identical to those in Section 3.3.1. The results of this experiment are summarized in Table 3.3.

We note that even in this pathological case, the combination of the articulatory features with MFCCs results in a significant improvement in classification, although it is

Table 3.3: *PER with Dist Feat computed using two fixed excitation parameters.*

Features (dimension)	fsew0	maps0	Both
MFCC + CMN (13)	64.2%	68.1%	66.1%
Dist Feat + LDA (20)	65.9%	72.3%	69.1%
Combination ($\alpha = 0.25$)	60.8%	65.5%	63.1%
Relative Improvement	5.3%	3.8%	4.5%

much less than what was obtained with oracle knowledge. The value for α obtained in this experiment looks counter-intuitive, suggesting that the system devotes 75% attention to the articulatory features. This could be attributed to the way the Dist Feat are extracted. In this experiment, we compute the distortion with respect to a fixed set of synthesized speech parameters and derive the new features. This also affects the performance of LDA projection which might not have been optimal. In the next experiment, we compute the distortion with respect to a variable set of synthesized speech parameters. Nonetheless, these results are in line with those of the previous experiment as well as with the results of the next one.

3.3.3 Excitation Derived from Incoming Speech

Here we actually attempt to mimic the incoming signal using the various codewords, in order to better localize the incoming signal in articulatory space. To do so, we derive the excitation signal parameters [Ps,A0,Q] from the original signal. Ps (lung pressure) is proportional to the rms energy. A0 and Q are proportional to the pitch. These excitations are then employed to synthesize signals from each of the 1024 articulatory configurations, which are used to derive a 1024-dimensional articulatory distortion feature vector. As before, the dimensionality of this vector is reduced to 20, prior to classification. $c(0)$ was not considered when computing the distortion. All other details of the classification experiment remain the same as in Section 3.3.1. Table 3.4 summarizes the results of this experiment.

We observe that in this "fair" test, the articulatory distortion features are effective

Table 3.4: *PER with Dist Feat computed using excitation parameters derived from the incoming speech.*

Features (dimension)	fsew0	maps0	Both
MFCC + CMN (13)	64.2%	68.1%	66.1%
Dist Feat + LDA (20)	63.2%	73.1%	68.1%
Combination ($\alpha = 0.6$)	56.9%	64.2%	60.5%
Relative Improvement	11.3%	5.7%	8.5%

at improving classification. Not only are the distortion features by themselves quite informative (as indicated by the PER obtained with them alone), they also appear to carry information not contained in the MFCCs. Interestingly for speaker "fsew0", the PER achieved with articulatory distortion features alone is 1% better than with MFCCs.

3.4 Further Optimization of the Setup

3.4.1 Optimizing the Number of Gaussian Mixtures

We further optimize the setup by increasing the number of Gaussian mixtures used to model the baseline MFCC features and the analysis-by-synthesis distortion features. The best classification results are achieved using 128 mixture components for each type of feature. In addition, appending $c(0)$, the energy coefficient, into the distortion features and employing CMN improves the performance. Improvements are achieved for both the distortion features alone and in combination with the baseline MFCC features. Table 3.5 shows the classification results using the same features in Subsection 3.3.3 but with a different setup.

3.4.2 Fast Analysis-by-Synthesis Distortion Features

In the previous section, we used the analysis-by-synthesis distortion features derived from a codebook of Maeda parameters. For each frame of incoming speech we used Maeda's model to convert the codeword to an area function and we then applied Sondhi

Table 3.5: *PER using 128 GMM for each type of features and including* $c(0)$ *and applying CMN on the distortion features.*

Features (dimension)	fsew0	maps0	Both
MFCC + CMN (13)	63.6%	67.1%	65.4%
Dist Feat + LDA + c(0) + CMN (21)	58.3%	69.3%	63.7%
Combination ($\alpha = 0.5$)	54.2%	62.7%	58.4%
Relative Improvement	14.8%	6.6%	10.7%

and Schroeter's chain matrices approach to convert the area function to a vocal tract transfer function. We also use the source information in the frame to synthesize speech. All of this is part of the "Synthesis" block of Figure 3.1. The source modeling technique of Sondhi and Schroeter is based on the two-mass model of vocal cords developed by Ishizaka and Flanagan [23]. In this approach, the source model is coupled with the vocal tract. This whole framework turned out to be very computationally intensive. We made two main modifications that improved the computations tremendously (from a week to a couple of hours worth of features computation for 3659 utterances, each around 2-4 secs long) with small degradation in classification accuracy.

The first modification is in the synthesis model and is explained in Section 4.5. The second modification is the use of a codebook of transfer functions rather than a codebook of Maeda parameters. In an off-line procedure, we use Maeda's model to convert the codebook of Maeda parameters to a codebook of area functions and use Sondhi and Schroeter's chain matrices approach to convert the area functions to a codebook of transfer functions, $\{H\_Tract, H\_Frication\}$ for each codeword. The entire vocal tract transfer function, *H_Tract*, including the nasal tract is used for voiced frames. For unvoiced frames, we use the Sondhi and Schroeter frication transfer function *H_Frication*. The codebook stores the impulse response of $h\_Tract = \{h_1^T, h_2^T, \ldots, h_L^T\}$ and $h\_Frication = \{h_1^F, h_2^F, \ldots, h_L^F\}$. L is the length of the impulse response. In computing the analysis-by-synthesis distortion features, we use these transfer functions in the manner shown in Figure 3.2. This saves a lot of unnecessary computation at run

time. The impulse response is converted to the frequency domain using the Fast Fourier Transform (FFT) and multiplied by the generated source signal in the "Fast-Synthesis" block using the overlap-add approach to synthesize speech from each codeword.

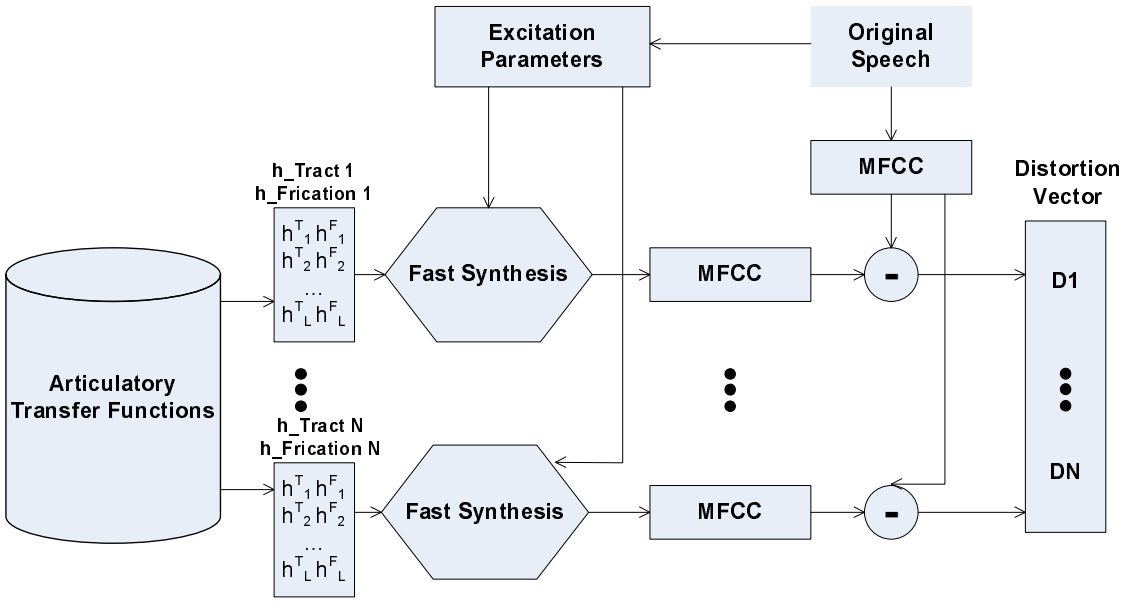


Figure 3.2: *A fast dynamic analysis-by-synthesis distortion framework.*

Table 3.6 shows the phone classification error rates using the set of features extracted by this new faster approach. CMN is applied to the distortion features. Adding $c(0)$ didn't improve the performance so we excluded it. The classification performance was slightly degraded but the reduction in computations was high. The degradation in classification accuracy can be explained by the new synthesis approach that we followed which is not optimal as explained in Section 4.5.

3.5 Discussion

Our results indicate that the analysis-by-synthesis features we introduce in this chapter do carry information that is complementary to that contained in the MFCCs. A 10.7% reduction in phone classification error is achieved when combining the new features with the baseline MFCC features. More importantly, the phone classification results indicate that

Table 3.6: *PER using fast analysis-by-synthesis distortion features (Fast Dist) and 128 GMM for each type of features and applying CMN on the distortion features.*

Features (dimension)	fsew0	maps0	Both
MFCC + CMN (13)	63.6%	67.1%	65.4%
Fast Dist + LDA + CMN (20)	61.6%	67.7%	64.6%
Combination ($\alpha = 0.55$)	56.9%	63.1%	59.9%
Relative Improvement	10.6%	6.0%	8.3%

articulatory configurations are intrinsic to phone identities. The articulatory features are, in effect, *knowledge-based* representations of the speech signal. Our experiments might thus indicate the potential value of combining physiologically-motivated systems based on the knowledge of speech production within the statistical framework of speech recognition. This argument is further supported by the fact that while such approaches were not considered feasible in the past due to computational considerations, modern computers make the incorporation of even highly computationally-intensive physical models of synthesis into the recognition process feasible.

The experiments we report on in this chapter use a very simple statistical model, aimed at highlighting the contributions of these features. The baseline results of our experiments are not optimal. In the GMM framework we are not modeling the transition probability of the states as in HMMs. Also we model context-independent (CI) phones that are segmented from connected speech, rather than using detailed triphone modeling. In addition, we do not use a phone language model or include the first and second order derivatives of the features for the reasons mentioned. In Chapter 5, we describe means of improving this framework. It is our hope that these improvements will also carry over to fully-featured HMM-based large vocabulary systems as well.

In addition to improving the framework, we invest our efforts in the next chapter to derive realistic vocal tract shapes from the EMA measurements. Rather than rely on a heuristic mapping from EMA to Maeda as we did in this chapter, we employ geometric adaptation and profile fitting of the vocal tract model into the EMA data.

The motivation is that we will be modeling more accurately the geometry of the vocal tract of each speaker and tracking closely the movements of the articulators. This will provide enhanced modeling of articulation to our framework.

Chapter 4

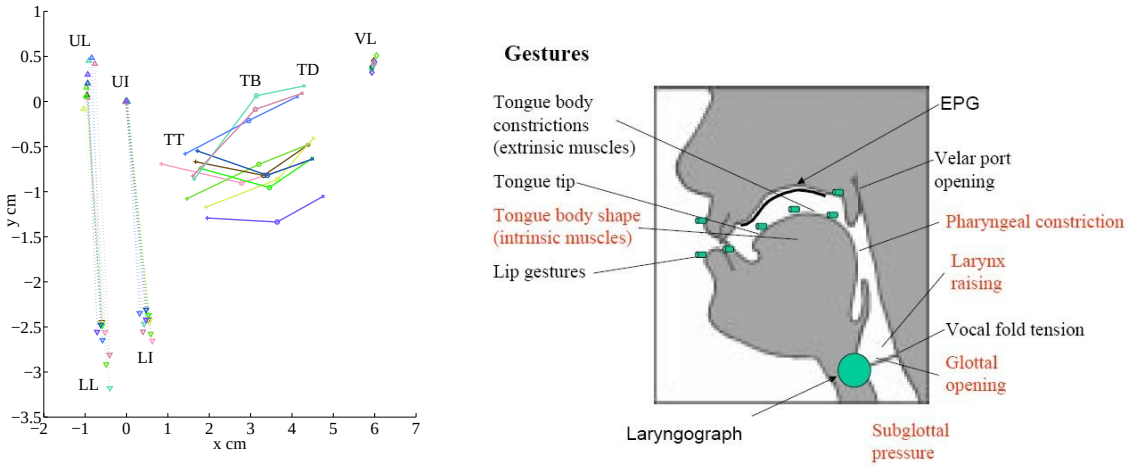
Deriving Realistic Vocal Tract Shapes from EMA Measurements

4.1 Introduction

ElectroMagnetic Articulography (EMA) has lately been gaining popularity among researchers as a simple technique for measuring the mechanism of speech production [2]. EMA, originally developed in the University of Göttingen in 1982, comprises a set of sensors placed on the lips, incisors, tongue, and velum of the speaker. A set of transmitters generates magnetic fields at multiple frequencies, each of which induces a current in the sensors. By measuring the levels of generated current, the (x, y) coordinates of each of the sensors can then be determined. Each EMA measurement thus consists of a set of such position coordinates, one from each sensor.

Figure 4.1 illustrates the positions of the sensors and the typical measurements obtained from the MOCHA database [2]. As the person speaks, a sequence of EMA measurements is obtained from the sensors. This sequence of measurements is assumed to provide at least a partial characterization of the speech production process. In addition to the EMA data, MOCHA also contains information on the contact of the tongue with the upper palate, the electro-palatography (EPG). It also contains information about voicing recorded using electro-glottography (EGG). In our work, we use the EMA data

only.



(a) EMA Data

(b) MOCHA Sensors

Figure 4.1: *(a) EMA measurements sampled from the MOCHA database. Notation of EMA data used is: upper lip (UL), lower lip (LL), upper incisor (UI), lower incisor (LI), tongue tip (TT), tongue body (TB), tongue dorsum (TD), and velum (VL). (b) MOCHA apparatus showing the EMA sensors in green (this figure is excerpted from [2]).*

But exactly how reliable are these measurements and how much do they tell us about the vocal tract that produces the speech? The EMA measures only the locations of a very small number of points on the vocal tract, typically four locations for the lips and incisors, one location on the velum, and merely three locations on the tongue. The vocal tract, on the other hand, is a complex three dimensional object that cannot be fully characterized by a small number of points. Furthermore, the precise location of the EMA sensors themselves is also highly uncertain and impossible to calibrate with respect to the vocal tract. Although the sensors on the tongue are placed at calibrated distances from one another, the elasticity and complexity of tongue structure prevents their actual positions, both along the tongue surface and relative to overall tongue structure, from being precisely known.

Given these various factors, it is natural to question the usefulness of these measurements as a characterization of the speech generation process. Clues may be found in

work by researchers who have previously shown that the EMA measurements are reliable cues to the speech signal itself. Toda *et al.* [26] have produced speech from EMA measurements using learned statistical dependencies between them and the corresponding speech signals, demonstrating that these measurements do indeed relate to the *output* of the speech generation process. Toth and Black [27] experimented with using EMA for voice transformation. While these experiments do provide indirect evidence of the relation of EMA measurements to the speech production mechanism, it is still not clear that they provide direct information about the shape of the speaker's vocal tract itself.

In this chapter we attempt to derive actual characterizations of vocal tract shapes from EMA measurements. Since the EMA itself comprises only a small set of sensor locations, we use a model-based approach to estimate the complete vocal tract configuration from them. Specifically, we use the model proposed by Maeda [9], which represents a mid-sagittal profile of the vocal tract in terms of seven parameters.

One simple approach to arriving at a vocal tract configuration in this manner is to determine the specific set of values for the seven Maeda parameters that best explains the measured EMA sensor positions [28]. This, however, is insufficient. Maeda's vocal tract model is not generic; it was originally developed using 1000 frames of cineradiographic and labiofilm data from only two female speakers. It must be *adapted* to the speakers in MOCHA. The specific aspects of the model that are adapted are the location of the center of the grid, the tilt of the oral cavity, and the length of the vocal tract. This is done by comparing the geometry suggested by the ensemble of all EMA measurements for the speaker to that defined by the model. The actual locations of individual sensors need not be known; hence the procedure is robust to variations and inconsistencies in sensor placement.

Once Maeda's model is adapted to the speaker, the actual vocal tract configuration corresponding to any set of EMA measurements is obtained through a simple codebook search. We use a codebook of Maeda model parameters that describes a large sampling of possible vocal tract shapes. For each EMA measurement, we select the vocal tract shape that is geometrically closest to the set of position coordinates represented in it. In order

to ensure that the estimate of the vocal tract is based entirely on geometric principles, since the EMA measurements are geometric in nature, we do not use the audio recordings of the speech signal in the first pass of the adaptation. For other adaptation approaches, the reader is referred to the work in [29, 30].

The "truthfulness" of the estimated vocal tract configurations can now be evaluated by synthesizing speech from them using an articulatory synthesis model and comparing synthesized speech to the actual speech signal produced during the utterances. We specifically use a modified version of the Sondhi and Schroeter model [8] for this purpose. Experiments show that the synthesized speech is similar to the actual speech, both perceptually and in terms of the Mel-cepstral distortion (MCD) metric [26] as we report in this chapter. Yet, we have not performed standard perceptual studies that depend on subjective human evaluation of the synthesis.

4.2 Maeda's Geometric Model

Maeda's model is composed of a two-dimensional semi-polar grid spanning the midsagittal plane of the vocal tract. The grid is composed of the red lines in Figure 4.2. The grid is made of 6 linear sections in the tongue region, 11 polar sections in the velum region, and 13 linear sections in the larynx region. It is defined by a set of parameters: the *Origin*, the width of each section d, and the angle of the polar region θ. The vocal tract itself is composed of an upper profile and a lower one. The upper profile shown in blue consists of the upper lip and incisor, upper palate, and pharynx and larynx outer wall. The inner profile consists of the lower lip and incisor, tongue, and pharynx and larynx inner wall and is shown in green.

Maeda uses seven parameters to generate the overall profile of the vocal tract. The formulation in Equation 4.1 summarizes the procedure in pseudo MATLAB code. p_1 is related to the jaw, p_2, p_3, and p_4 to the tongue, p_5 and p_6 to the lips, and p_7 to the larynx. The bases $[B_{larynx}\ B_{uwall}\ B_{tong}\ B_{lips}]$ and offsets $[O_{larynx}\ O_{uwall}\ O_{lips}]$ are derived from the speaker-specific vocal tract profiles Maeda extracted from the 1000 images. The vocal tract parameters are normalized within the [-3,+3] range and reflect standard deviations

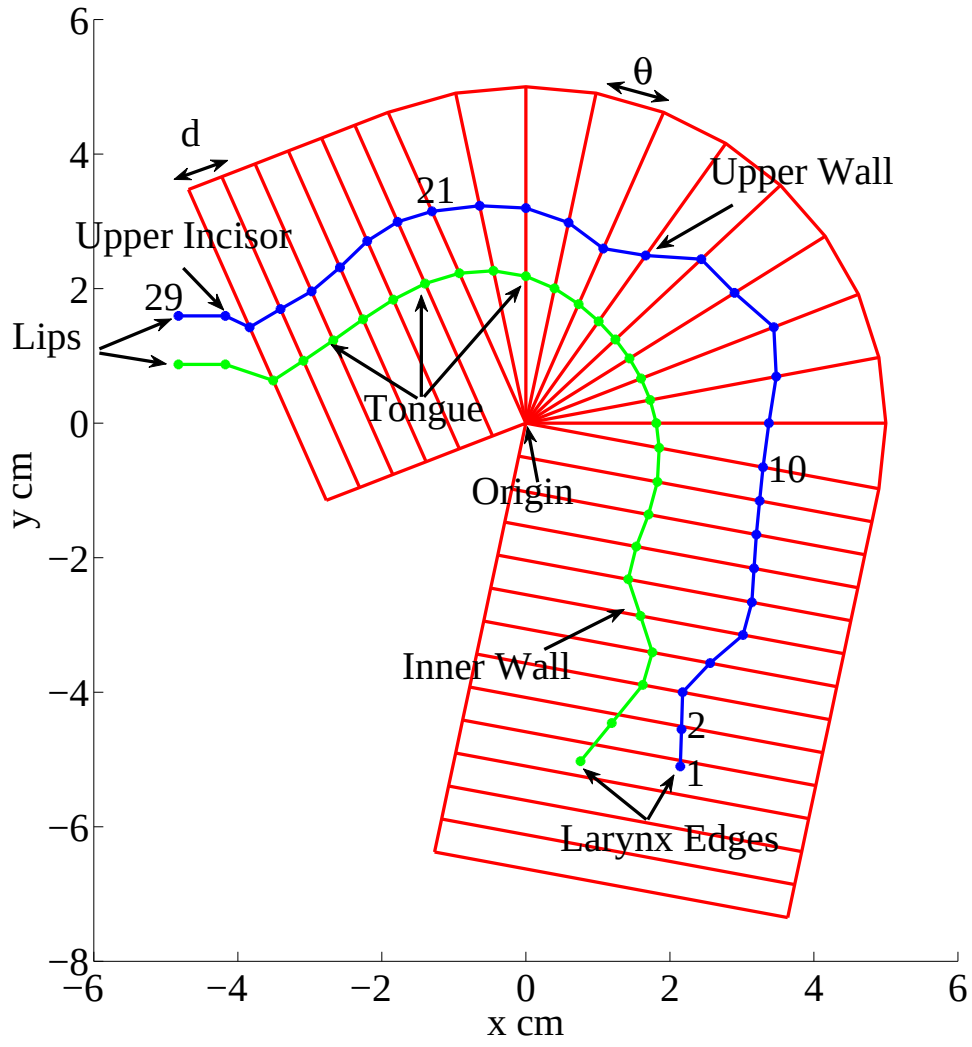


Figure 4.2: *Maeda's model composed of grid lines in red, vocal tract upper profile in blue, and vocal tract lower profile in green corresponding to the steady state shape with all p values set to zero.*

from the means of the bases rather than absolute numbers. The bases are multiplied by the parameters and then added to the offsets to generate different shapes. A 29-dimensional vector is computed using the formulation in Equation 4.1 and projected onto the grid lines, except for the four points describing the larynx edges and the four points describing the lips. The vocal tract profiles are composed of the lines joining the projected points.

$$
\begin{aligned}
Larynx &= B_{larynx} * [p_1\ p_7]' + O_{larynx} \\
UpperWall &= Proj(B_{uwall} + O_{uwall}) \\
Tongue &= Proj(B_{tong} * [p_1\ p_2\ p_3\ p_4]' + O_{uwall}) \\
Lips &= B_{lips} * [p_1\ p_5\ p_6]' + O_{lips}
\end{aligned} \tag{4.1}
$$

Using the seven Maeda parameters with the current model will create vocal tract shapes and generate sounds pertaining to the two speakers from whom the bases and offsets are derived. In order to make the model generate sounds pertaining to different speakers, it has to be able to match their vocal tract shapes. Hence it is necessary to adapt Maeda's model to the EMA data. Since the EMA data are purely geometric, we must ensure that the geometry of the Maeda model is accurate enough to be able to characterize the EMA measurements.

Note that the upper palate defined by $UpperWall$ in Equation 4.1 and shown in blue in Figure 4.2 is independent of the seven Maeda parameters. It is just a projection of the sum of B_{uwall} basis and the O_{uwall} offset into the geometric grid. Different parameters of the grid lead to different projected shapes. Hence, we choose to adapt the grid parameters and use the bases and offsets without adaptation. Table 4.1 lists the four grid adaptation parameters and the seven vocal tract shape-describing parameters.

4.3 Vocal Tract Model Adaptation to EMA Data

The grid parameters of Maeda's model are derived from the 1000 x-ray images of two female speakers. Hence they need to be adapted to be used with EMA data from

Table 4.1: *Maeda's model adaptation parameters and vocal tract shape parameters.*

Grid Adaptation Parameters	k, dx, dy, β
k	grid width stretching or compression factor
dx	origin x-axis translation (cm)
dy	origin y-axis translation (cm)
β	polar grid separation angle increment or decrement (radians)
Vocal Tract Shape Parameters	p_1, p_2, $\ldots p_7$
p_1	jaw vertical position
p_2	tongue dorsum position (forward or backward)
p_3	tongue dorsum shape (roundedness)
p_4	tongue tip vertical position
p_5	lip height
p_6	lip protrusion
p_7	larynx height

MOCHA. We follow a geometric adaptation procedure in this chapter and explain details of our approach. We start by describing how we processed the EMA data for each speaker. Since we are using the same adaptation process for all the speakers in MOCHA, we need to remove inconsistencies in the measurements in centering the data around a fixed reference and aligning it too. Once the data for each speaker are processed, we apply the adaptation procedure.

The adaptation process for each speaker can be summarized as follows. First we estimate the upper palate and mouth opening of each speaker in MOCHA. We then compute the average distance from the estimated upper palate to the Maeda model upper palate. Finally we modify the grid adaptation parameters and choose the set that yields the least average geometric distance between the two upper palates.

4.3.1 EMA Data Processing

The MOCHA database currently has recordings for only 10 speakers. Data for three speakers have been checked and processed by the creators of the database [2]. One of the remaining seven speakers has corrupted files. We have checked the data for the other six speakers in addition to the first three. We have also applied our own processing on the data as we explain here. The EMA measurements are recorded at 500-Hz sampling rate. We downsample the data to 100 Hz, the same sampling rate used for processing the acoustics (as per frame rate). We apply the MATLAB routine *decimate* which applies low-pass filtering beforehand to avoid aliasing. We apply the additional processing of the EMA data described below on a frame-by-frame basis. In order to use a geometric model with the data, we need to process the EMA consistently for all the speakers.

Centering the Data

For the three EMA speakers checked by the creators of MOCHA, the upper incisor (UI) is considered to be the center of the measurements. For all the utterances from the nine speakers, we subtract the UI from the EMA measurements on a frame-by-frame basis. Hence, we center the EMA measurements for all the speakers around the UI.

Aligning the Data

After centering the measurements, we noticed that different speakers have different tilt of their heads while recording the EMA. To align the data consistently for all the speakers, we align the sensor located at the bridge of the nose with the upper incisor. Hence, we compute the angle needed to rotate the EMA measurements such that the bridge of the nose and the upper incisor are vertically aligned. As with centering, we apply this rotation on a frame-by-frame basis. Rotation by angle α is applied using a matrix of the form $[cos\alpha\ sin\alpha; -sin\alpha\ cos\alpha]$.

Interpolating between the Tongue Sensors

The sensors on the tongue are placed at the tongue tip, body, and dorsum. The sensors are separated by 2 cm. Because our work is based on the geometry of the vocal tract, we need an accurate specification of the geometry of the oral cavity. We attempt to learn the geometry of the oral cavity from the distribution of all the measurements from the tongue sensors. To get a more detailed specification, we interpolate between each pair of sensors and add the interpolated measurements to the distribution. This helps covering the gap between each pair of sensors. Although it is not from a real sensor, yet the interpolated measurements should be a reasonable estimate due to the physiology of the tongue and to the geometric constraints of the oral cavity.

Data Cleaning Results

Figure 4.3 shows the scatter plot of all the raw measurements from speaker "falh0". Each dot on the scatter plot belongs to a frame from one utterance. We use all the frames from the checked utterances. The measurements from the velum sensor are shown in red, those from the bridge of the nose are in magenta, and the rest of the sensors are in blue. From the collection of measurements from the sensors on the tongue, one can infer the geometry of the oral cavity as indicated in the figure. Also shown in the figure is the gap between the sensors, especially in the upper palate.

Figure 4.4 shows the scatter plot of the measurements after processing. The figure shows the measurements centered around the upper incisor. All the measurements from the upper incisor sensor are mapped to $(0, 0)$. In addition, the bridge of the nose and the upper incisor are vertically aligned. The gap between the sensors is now covered by the measurements due to interpolation. One additional interesting thing to note in the figure is the shape of distribution of the velum measurements. It shows the velum moving as a valve that opens for nasalized sounds and closes for non-nasals. The few measurements that fall outside the scatter are due to interpolation errors. Some are also due to an uncommon sensor location, usually at the beginning or end of utterances when the speaker is not speaking.

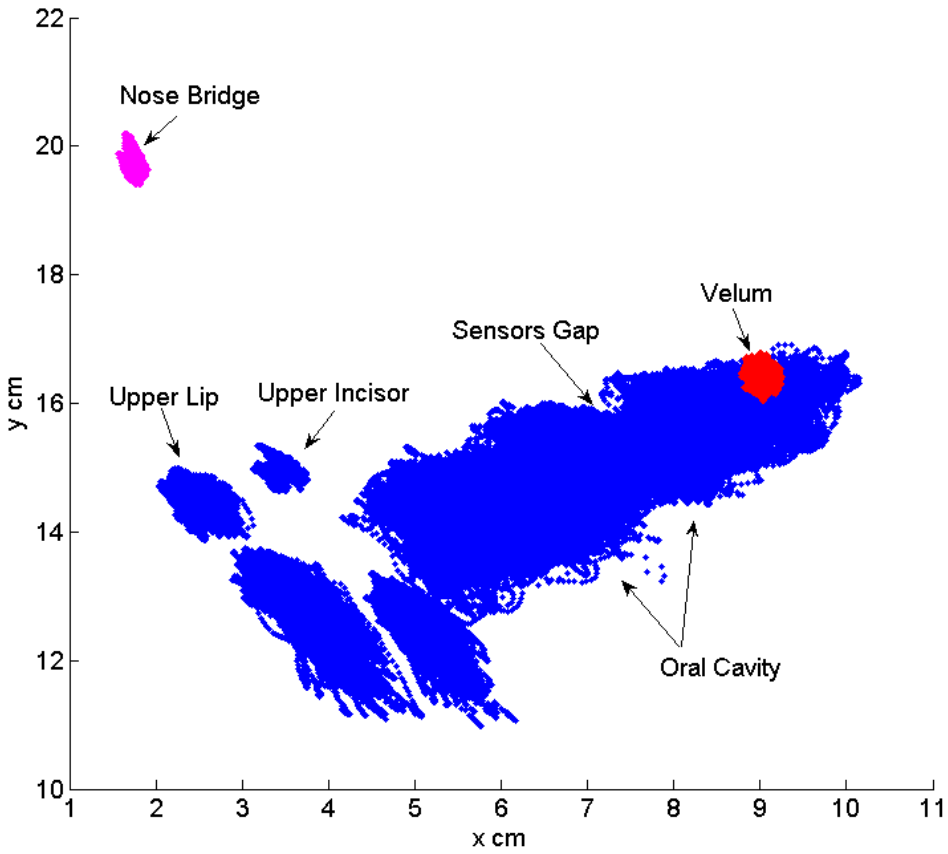


Figure 4.3: *Scatter plot of raw EMA measurements for all sensors for all the frames of data collected from speaker "falh0".*

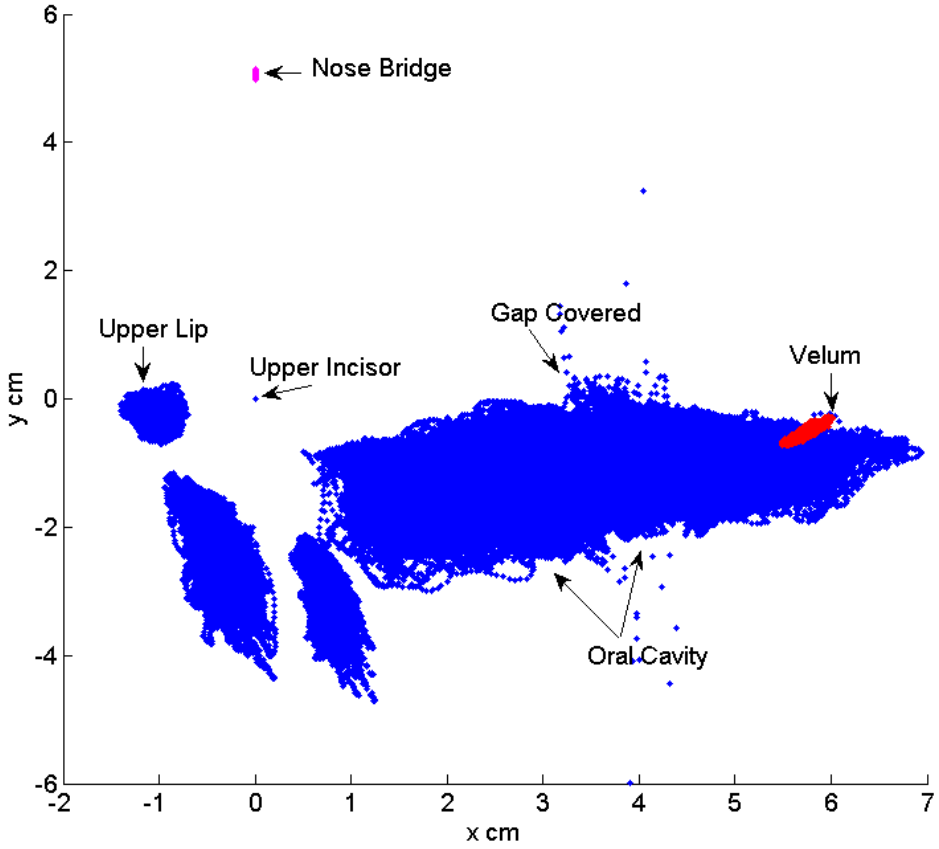


Figure 4.4: *Scatter plot of EMA measurements from speaker "falh0" after centering, rotation, and interpolation.*

4.3.2 Estimating Speaker Upper Palate and Mouth Opening from EMA

We estimate the upper wall of the EMA data using the distributions of the sensor positions for all the frames available for the speaker. Figure 4.5 shows the scatter plot of all the measurements from speaker "msak0", who will be our adaptation example speaker. The red circles in the figure are the mean locations of each sensor's measurements. The red lines with '+' signs are two-standard deviations from the mean of the distribution. These red lines summarize the movements of the sensors in two orthogonal directions. Refer to the caption in Figure 4.1.a for the notation used here.

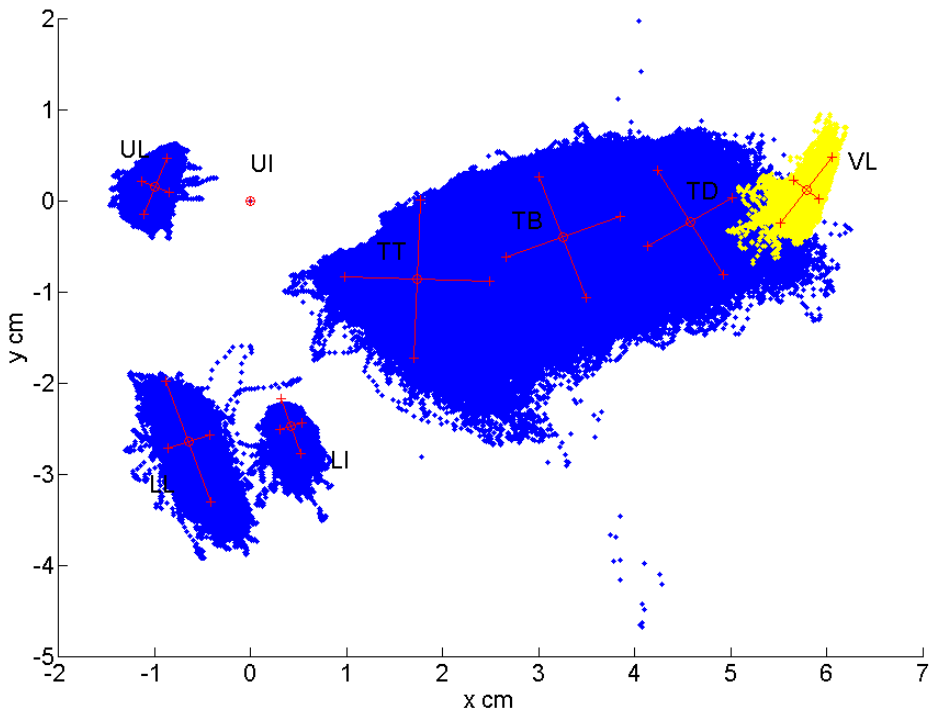


Figure 4.5: *Scatter plot of EMA measurements from speaker "msak0" after centering, rotation, and interpolation. The red lines show two-standard deviations from the mean of the sensor movements.*

Next we compute a histogram of the distribution of the data. Similar to the spec-

trogram, the two-dimensional plot of Figure 4.6 conveys three-dimensional information, the third being the density of the distribution reflected by the color intensity. Note also the distribution of the interpolated measurements making two of the five globes in the oral region.

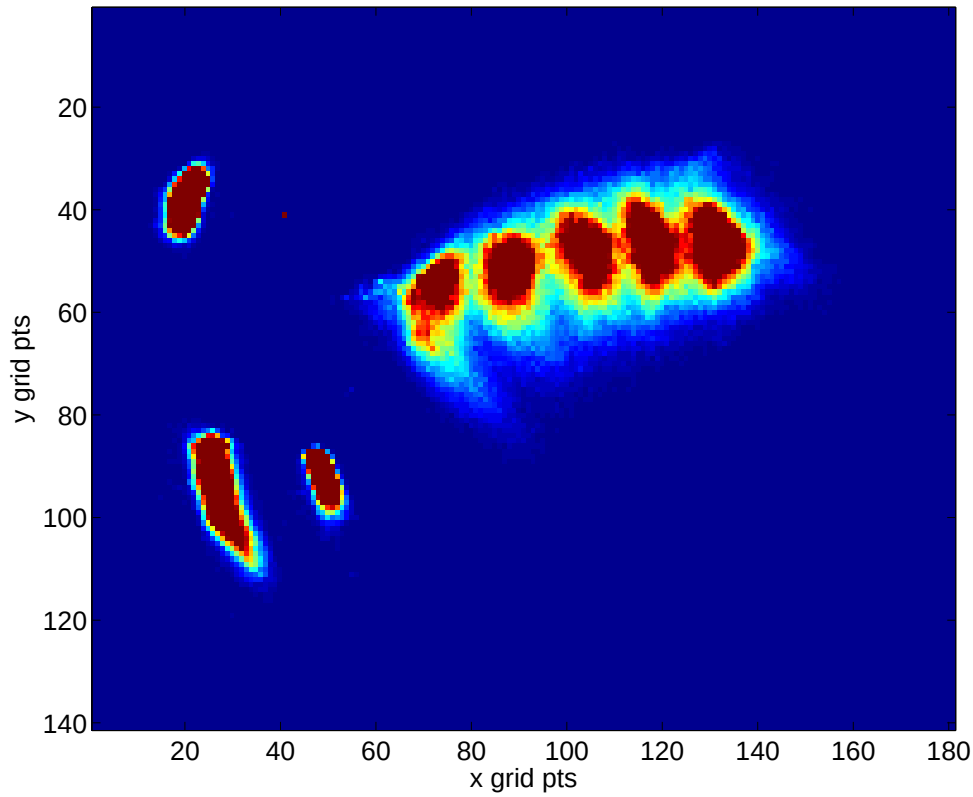


Figure 4.6: *Histogram of the distribution of the sensors. The color becomes darker as distribution becomes more dense.*

We then compute a smoothed histogram of the positions of these sensors and label five disconnected regions: UL, LL, UI, LI, and the mouth cavity using the MATLAB command *bwlabel*. The biggest of these regions is the mouth cavity composed of the region of the TT, TB, and TD. We set the highest points in the mouth cavity as the upper wall estimated from the EMA. We add to this estimated upper wall the mean location of the velum sensor, VL. We also set the top ten points at the left side of this connected region as the mouth opening. Figure 4.7 shows the smoothed histogram with

the estimated EMA upper wall.

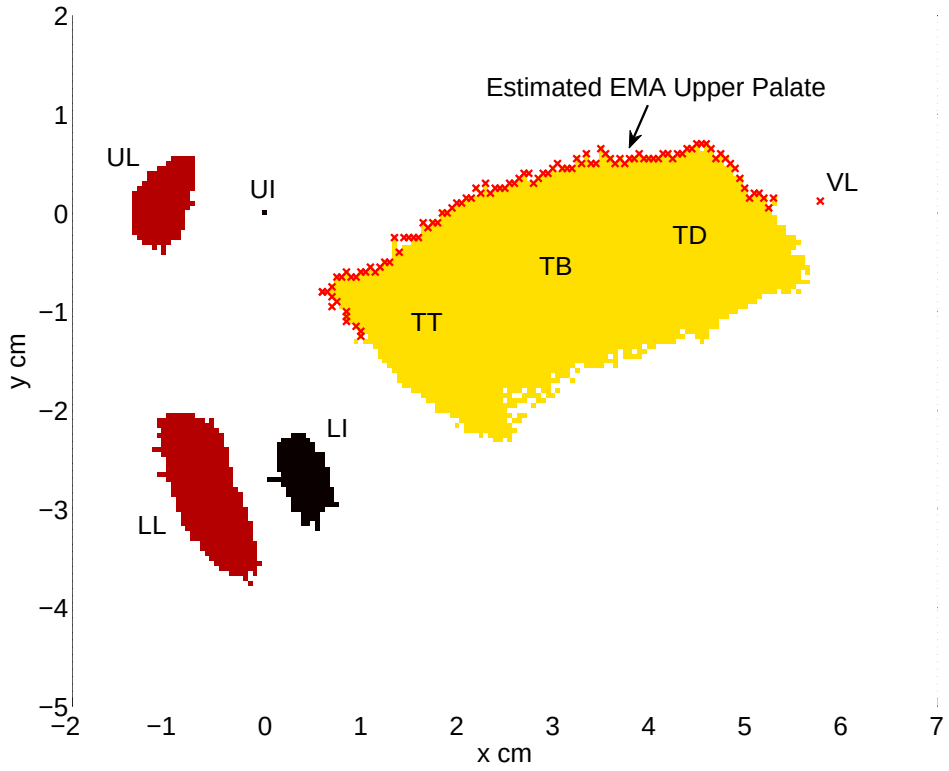


Figure 4.7: *Smoothed histogram of the distribution of the sensors. The red crosses show the estimated upper palate and mouth opening.*

4.3.3 Translating the Origin of Maeda's Model

We follow an approach similar to the one by McGowan [29] by superimposing the EMA data onto the Maeda model semi-polar coordinate space. We first need to match the coordinates of the two systems. In MOCHA, the sensor placed on the upper incisor is used as the origin [2]. In Maeda's model, the upper incisor is at a fixed location defined by $(INCI_x, INCI_y)$. Hence we translate Maeda's model coordinates such that the new origin coincides with the upper incisor. During adaptation, we shift the origin (labeled *Origin* in Figure 4.2) by (dx, dy). This will shift the whole grid and will also change the projection of the Maeda model upper palate as we explain below.

4.3.4 Adapting the Grid of Maeda's Model and Fitting the Estimated EMA Palate

The Maeda model parameters that we choose to adapt are the origin of the grid *Origin*, the width of each of the grid's linear sections d, and the angle between the polar sections of the grid θ. These parameters are labeled in Figure 4.2. The *Origin* is adapted by the shifting it by (dx, dy). The linear section's width d is adapted by expanding or compressing it by a factor of k. The polar section's angle θ is adapted by a small increment β in radians. Equation 4.2 describes mathematically the adaptation procedure.

$$\begin{aligned} d' &= d(1-k) \\ \theta' &= \theta + \beta \\ Origin &= -(INCI_x + dx) - j(INCI_y + dy) \end{aligned} \tag{4.2}$$

The adaptation is based on an optimization procedure that is summarized in Equation 4.3. We minimize the distance between a function of the EMA data and a function of the model parameters. We choose a range over which we vary each of the four grid adaptation parameters: *k, β, dx, dy*. For each combination we compute the average geometric distance between all the points on the estimated EMA and the adapted Maeda upper walls. These distances are shown in magenta in Figure 4.8. We choose the set of parameters with the least average distance. Note that the value of k reflects vocal tract stretching or compression with respect to the standard Maeda model. The value of β reflects oral tract tilt. Similarly the value of (dx, dy) corresponds to shifting the grid horizontally and vertically, respectively.

$$\{\hat{k}, \hat{\beta}, \hat{dx}, \hat{dy}\} = argmin_{k,\beta,dx,dy}(|f(EMA) - f(model)|) \tag{4.3}$$

4.3.5 Lips Translation

The EMA lip sensors are placed outside the mouth on the tip of the lips [2]. This means that even when the lips are closed there is still a vertical gap between the two sensors

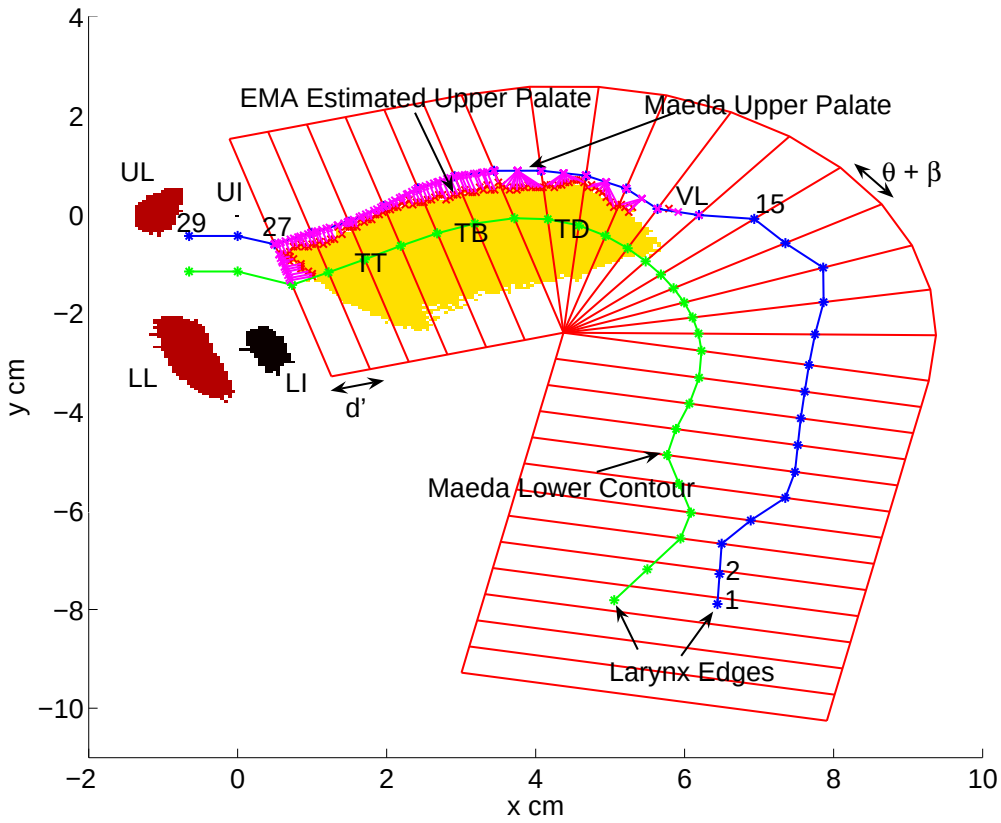


Figure 4.8: *Vocal tract adaptation showing the smoothed scatter plot of the distribution of the EMA data (yellow, red, and black) and the superimposed Maeda model (red grid lines). The green contour is for the steady state Maeda lower contour and the blue contour is the adapted Maeda upper contour resembling lips, palate, and larynx.*

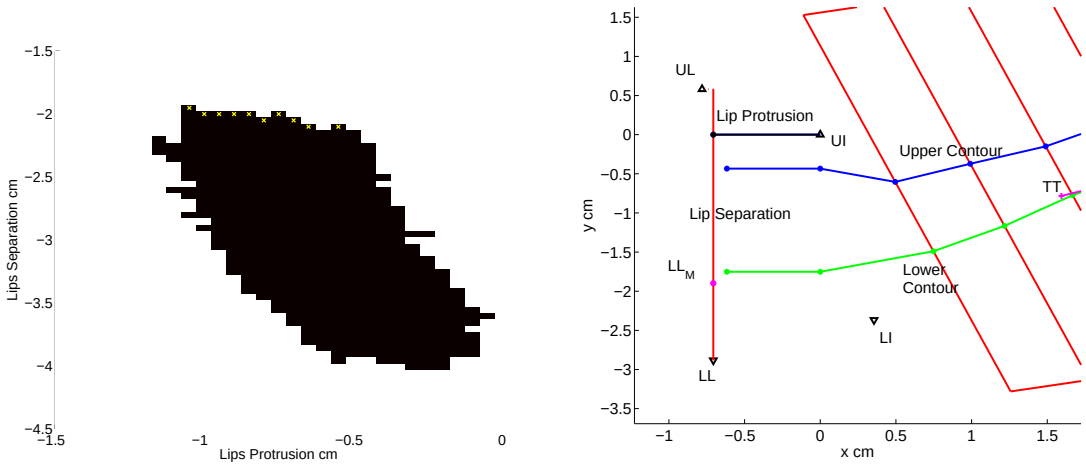
as evident in Figure 4.7. We estimate the minimum lip separation, Lip_{sep}, which is the gap between the sensors when the lips are closed.

Figure 4.9a shows the smooth distribution of the lip separation and lip protrusion for all data from speaker "msak0". The protrusion is defined as the horizontal distance between the UI and the UL or the LL, whichever is smaller. The separation is defined as the vertical distance between UL and LL. The distances are negative values since they are to the left and bottom of the origin. It is clear from the distribution that the minimum lip separation is 2 cm as indicated by the yellow points. The figure also shows that the more the lips protrude, the less separated they are. Once Lip_{sep} is estimated, 2 cm in this case, the lip translation is performed for each frame of data as shown in Figure 4.9b.

In Maeda's model, the outermost lower lip point has two degrees of freedom proportional to protrusion and separation. We map the four EMA measurements (UL, LL, UI, and LI) to a point LL_M in the vicinity of Maeda's outermost lower lip point on the green contour shown in Figure 4.9.b. The protrusion is shown by the horizontal black line in the figure. The separation is shown by the vertical red line. Lip_{sep} is then subtracted from the separation measure and the LL_M is found. Figure 4.9.b also shows the upper and lower contours of Maeda's model forming the lips tubes. In Maeda's model the two contours are allowed to close at the lips. To account for this, we subtract Lip_{sep} from the EMA measurements.

4.3.6 Velum Location and Nasal Tract Opening Area

The Sondhi and Schroeter model allows for nasal tract coupling to the vocal tract by adjusting the velum opening area. The location of the velum VL_{loc} is set after adaptation to the grid section number that precedes the one which contains the mean of VL counting from the glottis to the lips. This is because the velum sensor is placed on the soft palate [2]. The velum opening area is estimated from the ordinate of the velum sensor VL_y. For each utterance, the nasal tract is opened in proportion to how much the value of VL_y is below its mean over the utterance. The lower the velum, the more nasalized is



(a) Distribution (b) Translation Procedure

Figure 4.9: *(a) Smoothed histogram distribution of lip protrusion and separation. (b) Lip translation for one frame of data, also showing the Maeda model upper and lower lip contours.*

the sound and hence the larger is the opening area. Figure 4.10 shows the values of VL_y and the corresponding estimated values of nasal tract opening area, normalized between $[0, 1]$. The mean value of the ordinate over the utterance is shown in red.

4.3.7 Adaptation Results

As described in Subsection 4.3.4, we vary *k, β, dx, dy* until the projected Maeda upper palate best matches the estimated EMA upper palate. Before adaptation the average distance between the two contours is 0.67 cm and the values of *k, β, dx, dy* are {0, 0 rad, 0 cm, 0 cm}. This distance is the average of the distances from each point on the EMA upper palate (red) to the Maeda upper contour (blue), shown in Figure 4.8 in magenta. The average distance between the two contours after adaptation is 0.25 cm and the values of *k, β, dx, dy* are {-0.08, -0.0105 rad, -0.2 cm, 0.6 cm}. This means that the length of the vocal tract is extended by 8% and that the grid is shifted by 0.2 cm to the left and 0.6 cm upward. These numbers make sense since Maeda's model is based on images from two female speakers and the MOCHA speaker "msak0" is a male

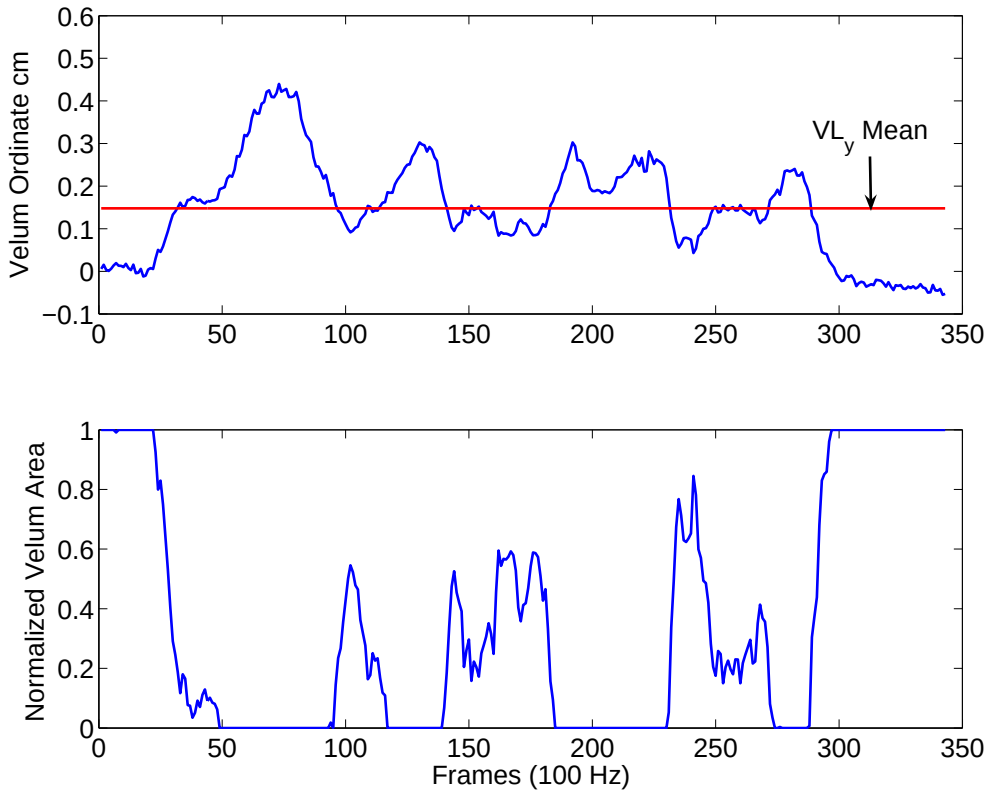


Figure 4.10: *Top figure shows the values of the ordinate of the velum sensor for utterance "Jane may earn more money by working hard". The bottom figure shows the values of the estimated nasal tract opening area.*

speaker. In addition, a close match is attained between the two upper palate contours. One constraint we apply here is that all the points on the EMA estimated upper palate (in red) should be below the adapted Maeda model upper contour (in blue). We later relax this constraint. Note also that the mean of the velum sensor locations, VL, almost falls on the adapted Maeda upper palate. In addition, the figure shows the distances between the estimated mouth opening and the left side of the grid. This distance is added in the overall measure to ensure correct matching.

4.4 Vocal Tract Profile Fitting

In Section 3.1, an approach for mapping EMA data to Maeda parameters is described for the purpose of speech recognition. It uses a heuristic mapping from EMA directly to Maeda parameters without actually using the model. For example, p_5, is simply the normalized distance between the UL and LL.

The work in this chapter describes a more principled approach that searches for the best fit of the EMA data to the adapted Maeda model vocal tract contours. We use a uniform codebook of Maeda parameters that represents different vocal tract shapes. For each frame of EMA data, we search for the best geometric fit. The best fit of the tongue and lip contours found for each frame of EMA data is then used in articulatory speech synthesis.

4.4.1 Codebook Design

We create a uniform codebook composed of 164000 codewords, where p_1 to p_4 take values of {-3, -2.25, -1.5, -0.75, 0, 0.75, 1.5, 2.25, 3}. p_5 and p_6 take the values of {-3, -1.5, 0, 1.5, 3}. p_7 is set to zero since we do not have EMA data to estimate the larynx position. Some of the shapes in this codebook have a constriction in the larynx region as shown in Figure 4.11. We remove the shapes with an area less that 2.5 cm^2 in the first 14 tubes and are left with 18525 codewords for the "msak0" speaker. We believe that the removed shapes are unlikely to occur for English speech.

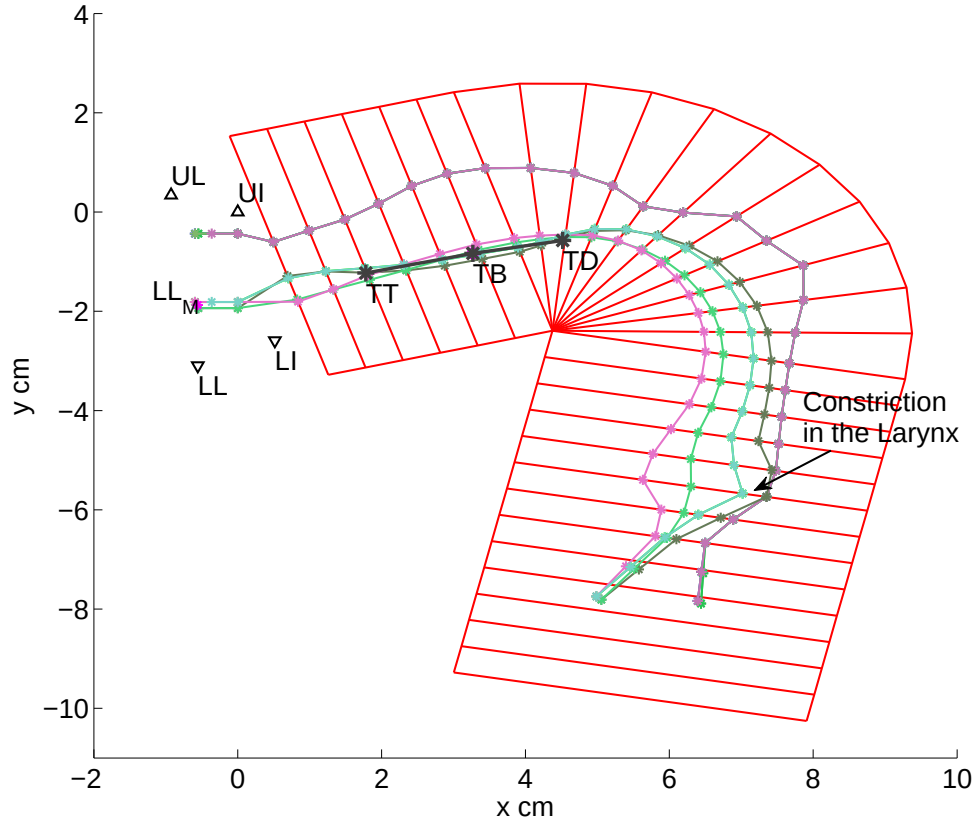


Figure 4.11: *The top four best-matching vocal tract shapes for a frame of EMA data. The EMA points are shown in solid black. The four shapes match the EMA well, but some of them exhibit a constriction in the larynx region.*

4.4.2 Searching Vocal Tract Shapes

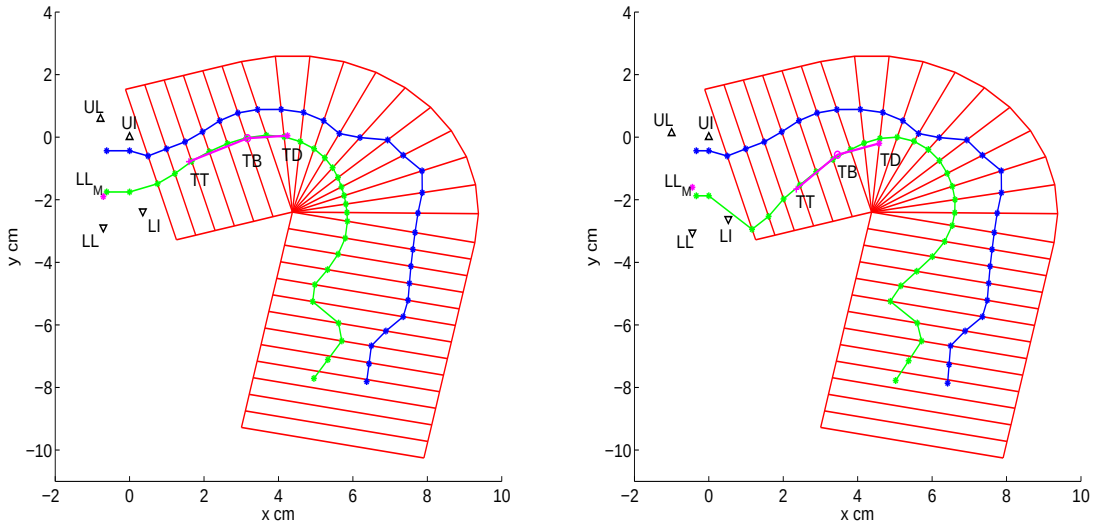
For each codeword, we compute the vocal tract profile and project it onto the adapted semi-polar coordinate space. For the EMA data, we first translate the UL, LL, UI, and LI to the LL_M point as described in Subsection 4.3.5. Then we compute the distance from this point to the outermost lower lip point in the lower contour provided by the given codeword. Thus, we compute the first distance pertaining to the lips. Then, for the TT, TB, and TD EMA points, we first find the grid section number in which each of these points falls. We then compute the distance of the EMA point to the segment of the lower vocal tract contour that falls within this grid section, which enables us to find the three other distances. The overall geometric distance between the given frame of EMA data and the vocal tract contour of the given codeword is the mean of the above four distances. We choose the codeword that yields the least distance.

4.4.3 Search Results

Note in Figure 4.8 that the minimum separation between the lips, Lip_{sep}, is estimated to be 2 cm. This measure is used to first translate the lips to the LL_M point. Figure 4.12 shows results of the search for the vocal tract shapes of the EMA data belonging to two frames in the middle of phones {'II', '@@'} in the words "Seesaw = /S-II-S-OO/" and "Working = /W-@@-K-I-NG/" respectively. It is clear that the resulting vocal tract shapes fit the projected EMA data well and reflect the articulatory characteristics of the two phones. Phone 'II' is a high front vowel and phone '@@' is a high back vowel. Note the difference in the lip opening and the protrusion. The solution vector of the Maeda parameters for Figure 4.12a is $[2.25, 0, 0, -2.25, 1.5, 0, 0]$ and for Figure 4.12b is $[0.75, 1.5, 2.25, -3, 1.5, -1.5, 0]$.

4.5 A Modified Synthesis Model

Once we find the best matching vocal tract shape, we convert it to the areas and lengths of the tubes forming the sections of the vocal tract. We follow Maeda's approach in computing the effective length and area of each tube bounded within the upper and lower

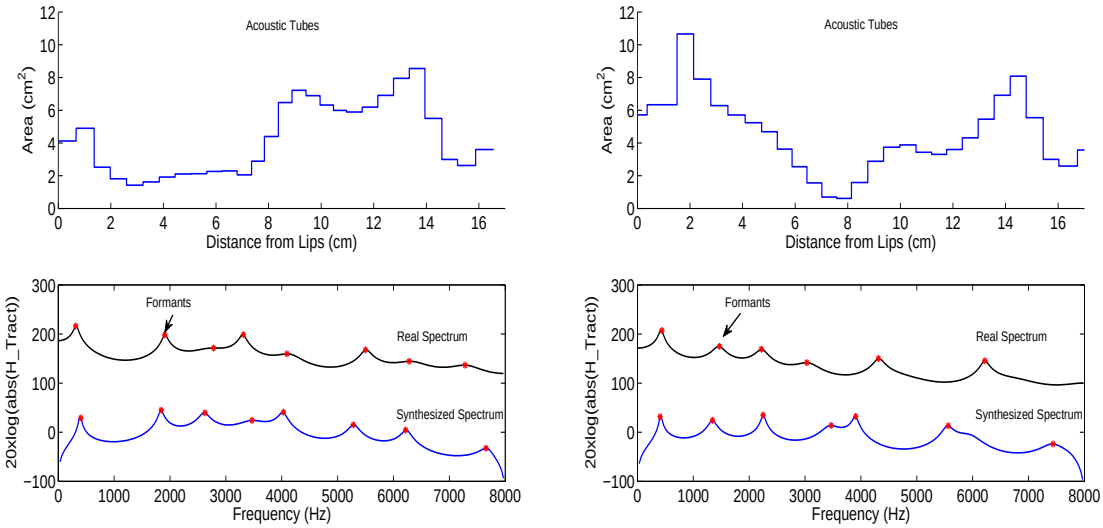


(a) 'II' (b) '@@'

Figure 4.12: *Search results for two EMA frames for 'II' in "Seesaw = /S-II-S-OO/" and '@@' in "Working = /W-@@-K-I-NG/". The EMA points used are LL_M, TT, TB, and TD shown in magenta. The resulting inner and upper walls of the matched shapes are in green and blue respectively.*

contours and the grid lines. We then feed these areas and lengths to an articulatory speech synthesizer. We use the Sondhi and Schroeter model [8] which uses the chain matrices approach to derive the overall transfer function of the vocal tract. The top plots in Figure 4.13 show the resulting areas and lengths of the sections of the vocal tract contours of Figure 4.12. The lower plots show the transfer functions obtained using the Sondhi and Schroeter model. The formant frequencies are marked in red. In addition, the plots show the LPC smoothed spectra of the real speech for the two examples. A close match between real and synthesized spectra is attained for the two vowels, especially for the first three formants.

We then replace the source modeling of Sondhi and Schroeter that uses the two-mass model of vocal cords developed by Ishizaka and Flanagan [23] with a modified version. The new approximation decouples the vocal tract from the glottis. The transfer function is obtained from the Sondhi and Schroeter entire vocal tract transfer function,



(a) 'II' (b) '@@'

Figure 4.13: *Upper plots show the areas and lengths of the sections of the acoustic tubes of the vocal tract contours in Figure 4.12. Lower plots show the corresponding transfer functions synthesized using the Sondhi and Schroeter approach in blue. The formants are shown in red. In addition, the LPC smoothed spectra of the real speech from the two examples is shown in black.*

H_Tract, including the nasal tract for voiced frames. For unvoiced frames the Sondhi and Schroeter frication transfer function *H_Frication* is used. For generating the source signal, we use the Rosenberg glottal pulse model [31] for voiced frames and random noise for unvoiced frames as shown in Figure 4.14. We extract the energy and pitch from the original speech signal and use them in generating the source signal. This approach improves the synthesis quality and is faster than our previous approach. Nevertheless, it is still an approximation since some phones, like voiced fricatives (*e.g.* 'Z'), are not well synthesized this way. In general, the synthesis is not as good for fricatives as it is for vowels. Equation 4.4 describes the overlap-add approach we follow to convolve the source signal with the transfer function. The operations are computed as multiplications in the frequency domain. Results are converted back to the time domain using the inverse Fast Fourier Transform (iFFT). Figure 4.14 shows the building blocks used in this procedure.

This approach improves the synthesis quality and is faster than our previous approach [28].

$$Voiced\_Speech = overlapp\_add(H\_Tract, Rosenberg\_Glottal\_Pulse)$$
$$Unvoiced\_Speech = overlapp\_add(H\_Frication, Random\_Noise) \quad (4.4)$$

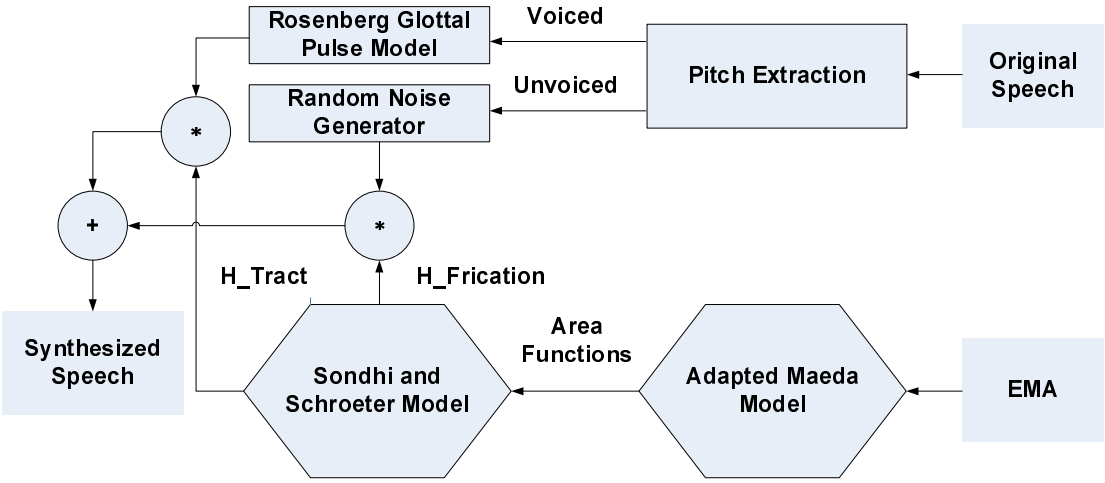


Figure 4.14: *A fast and dynamic articulatory synthesis framework.*

4.5.1 Synthesis Results

We estimate the nasal tract opening area from the EMA measurements of the velum sensor VL as described in Subsection 4.3.6. Since Maeda's model is independent of the Sondhi and Schroeter model, the exact location of the velum is hard to determine. Experiments have shown that setting VL_{loc} to 8 yields the best synthesis quality.

Figure 4.15 shows a step-by-step analysis of the synthesis procedure for the utterance: "Those thieves stole thirty jewels". The upper plot is the spectrogram of the real speech spoken by speaker "msak0". The second plot shows the root-mean-squared (rms) energy of the real speech. The third plot shows the pitch frequency and the fourth plot shows the source signal generated by our approach. Note that the source amplitude is weighted by the energy value. When pitch is zero, meaning that the speech is unvoiced, the source function is random noise. Otherwise, the source is a train of glottal pulses separated

by the reciprocal of the pitch value. The fifth plot is the transfer function derived from the EMA measurements using Maeda's model and Sondhi and Schroeter's model. Note also that the transfer function switches from vocal tract function *H_Tract* to the frication function *H_Frication* depending on the pitch. Finally, the lower plot shows the synthesized speech, which is the convolution of source and the transfer function.

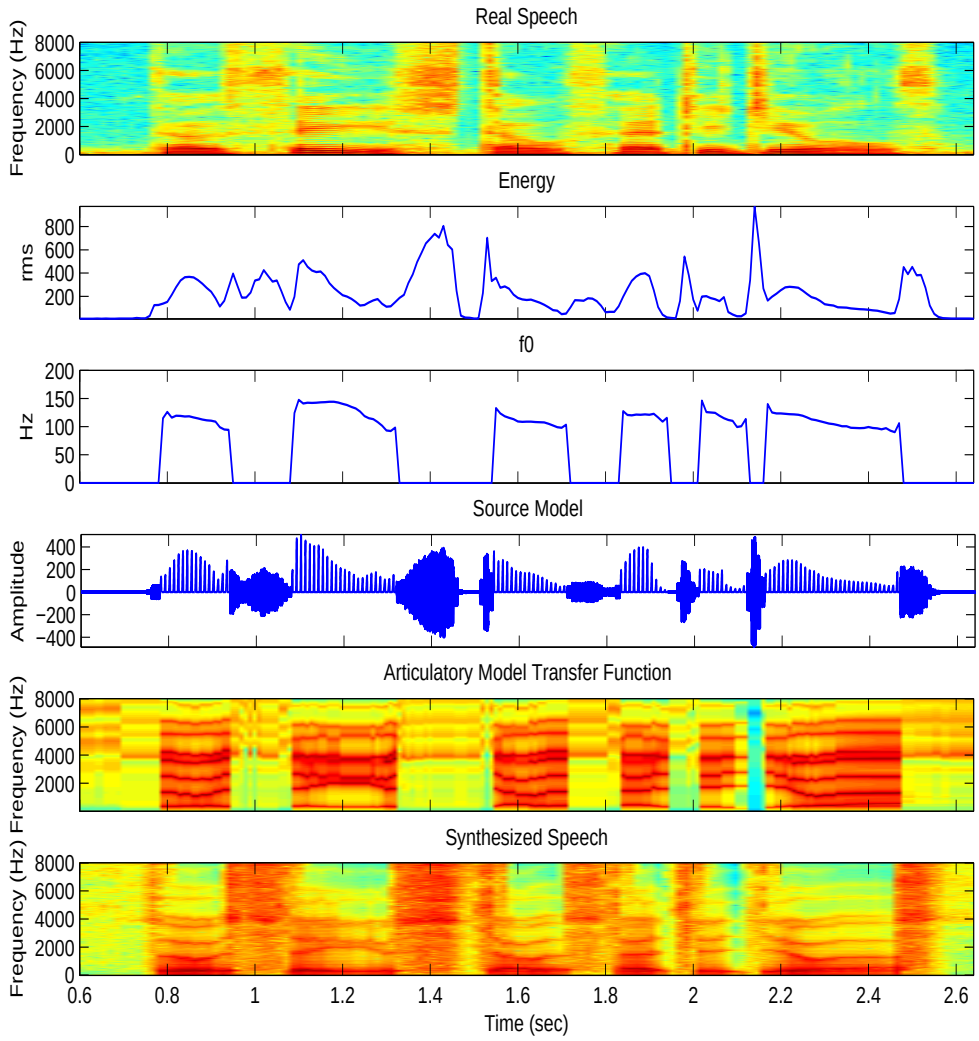


Figure 4.15: *Analysis of the synthesis procedure for the utterance: "Those thieves stole thirty jewels".*

Figure 4.16 shows the spectrogram of the real speech and the speech synthesized

from the EMA measurements in more detail. It is clear in the figure that the synthesized spectrum corresponds well to the real spectrum.

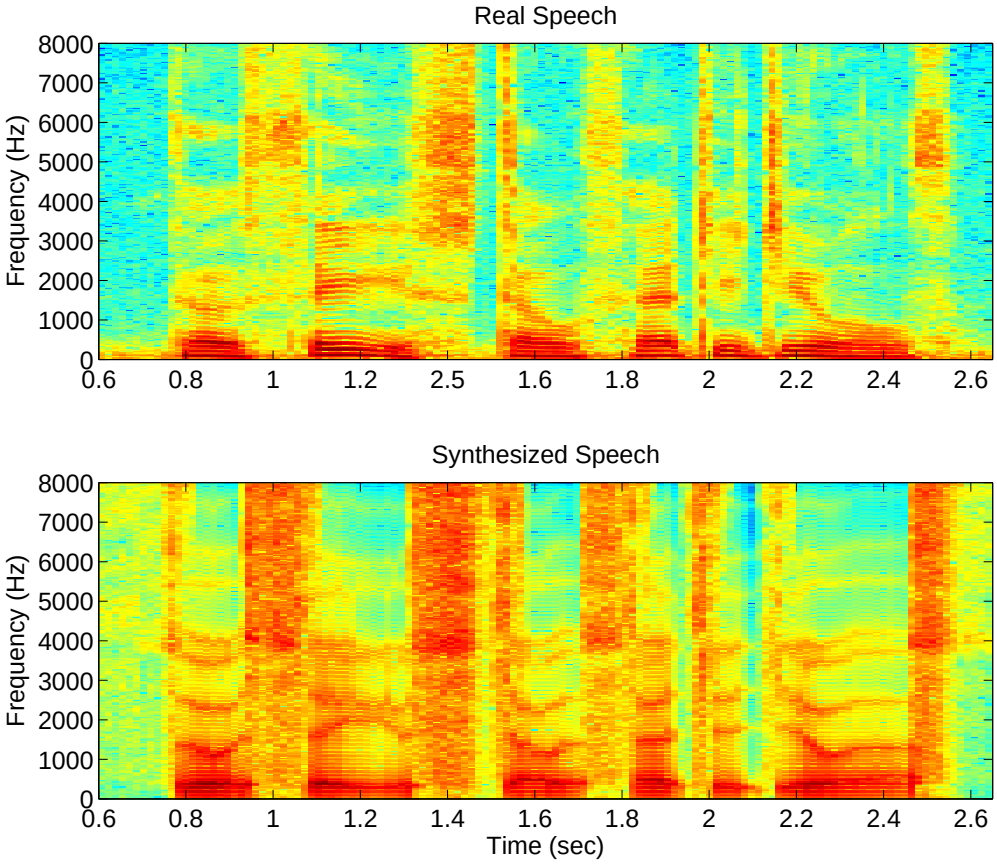


Figure 4.16: *Spectrogram of the real and synthesized speech for the utterance: "Those thieves stole thirty jewels".*

4.6 Incorporating Search and Synthesis Errors in the Adaptation Technique

The preceding work on adaptation looked at the geometric match between the estimated EMA upper palate for speaker "msak0" and the upper palate of Maeda's model. Relying only on geometric matching can be tricky. The solution of adaptation parameters for the best geometric fit is non-unique. There are many degrees of freedom to match the

upper palates. Even with the constraints we impose on the set of values of adaptation parameters, different solutions lead to similar average geometric distances.

In this section, we apply two additional measures to decide which set of adaptation parameters is most optimal. For each value of k, the grid stretch or compression parameter, we find the values of β, dx, dy that yields the best geometric match between the palates, *i.e.* the least average geometric distance GD. Next, we use the first 10 utterances from each speaker to derive two additional measures: the search error SE and the mean formant difference FD. SE is due to the distance between the EMA points from each frame and the best-fitted vocal tract contour. Note that in Figure 4.12, the EMA points did not fall completely on the green contours. The offset is what we denote as the average search error SE computed for all the mid-vowel frames in the 10 utterances. Similarly, looking at Figure 4.13, there is a mismatch of formants between the real spectra and the synthesized one for the two frames. We compute the difference between the first three formants for all the mid-vowel frames in the 10 utterances and denote it as the average formant difference FD. The overall error is the weighted sum of the normalized errors. We normalize each error, GD, SE, and FD over the different sets of adaptation parameters using the minimum and standard deviation of each error.

$$overall\_error = a_1 * SE_n + a_2 * GD_n + a_3 * FD_n \tag{4.5}$$

Equation 4.5 shows the computation of the overall error where n denotes normalization. The values of a_1, a_2, a_3 are set to $0.2, 0.3, 0.5$ respectively by experimentation. Starting with a fixed k, we find the values for the rest of adaptation parameters that yield the least GD. So the basics of the adaptation are still geometric but we add the acoustic measure FD as a second step. In addition, we only use the mid-vowel frames from 10 utterances to choose the adaptation set with minimum $overall\_error$. The search error SE is an indication of how well the adaptation set yields vocal tract contours that match the EMA data for these frames. The set of 10 utterances for each speaker can be considered as a development set. Using the formants in the adaptation process has been done before in [29, 30]. Matching the formants of the real and the synthesized speech is

supposed to improve the synthesis quality.

4.7 Experimental Results on Vocal Tract Adaptation and Synthesis for the Available Speakers in MOCHA

The MOCHA database contains EMA measurements and the corresponding acoustic speech signals for 10 speakers reading 460 TIMIT sentences. In this chapter, we use the EMA data from nine speakers: "msak0", "maps0", "mjjn0", "ss2404", "fsew0", "ffes0", "faet0", "falh0", and "fjmw0". We use all the EMA data available from each speaker to geometrically adapt Maeda's model and derive the average geometric distance GD. The number of utterances available for each speaker is shown in Tables 4.2 and 4.3.

We use EMA data from the mid-vowel frames of the first 10 utterances of each speaker to derive the search error SE. We use the first three formant frequencies from the mid-vowel frames of the first 10 utterances of each speaker to derive the formants difference measure FD. The first 10 utterances of each speaker are considered as the development set. They are not used in evaluating the synthesis quality. We use the EMA data from the remaining utterances as test data to perform the synthesis and compare it to the corresponding real speech.

Figure 4.17 summarizes the basic operations that we follow to synthesize speech from the EMA measurements. For each speaker, we preprocess the EMA measurements to center them around a common reference. Next we estimate the upper palate of the EMA data and the mouth opening and find the adaptation parameters that produce the best matching Maeda model grid. The second row of blocks describes the search for the best-matching Maeda vocal tract contours on a frame-by-frame basis. The areas and lengths of the corresponding acoustic tubes between the contours are also computed. The third row of blocks describes the synthesis of speech from the area functions and computation of the MFCC of the synthesized speech. The last row describes the computation of the MCD between the synthesized and real speech.

In the overall process, we have derived a set of realistic vocal tract shapes for each EMA frame for each speaker.

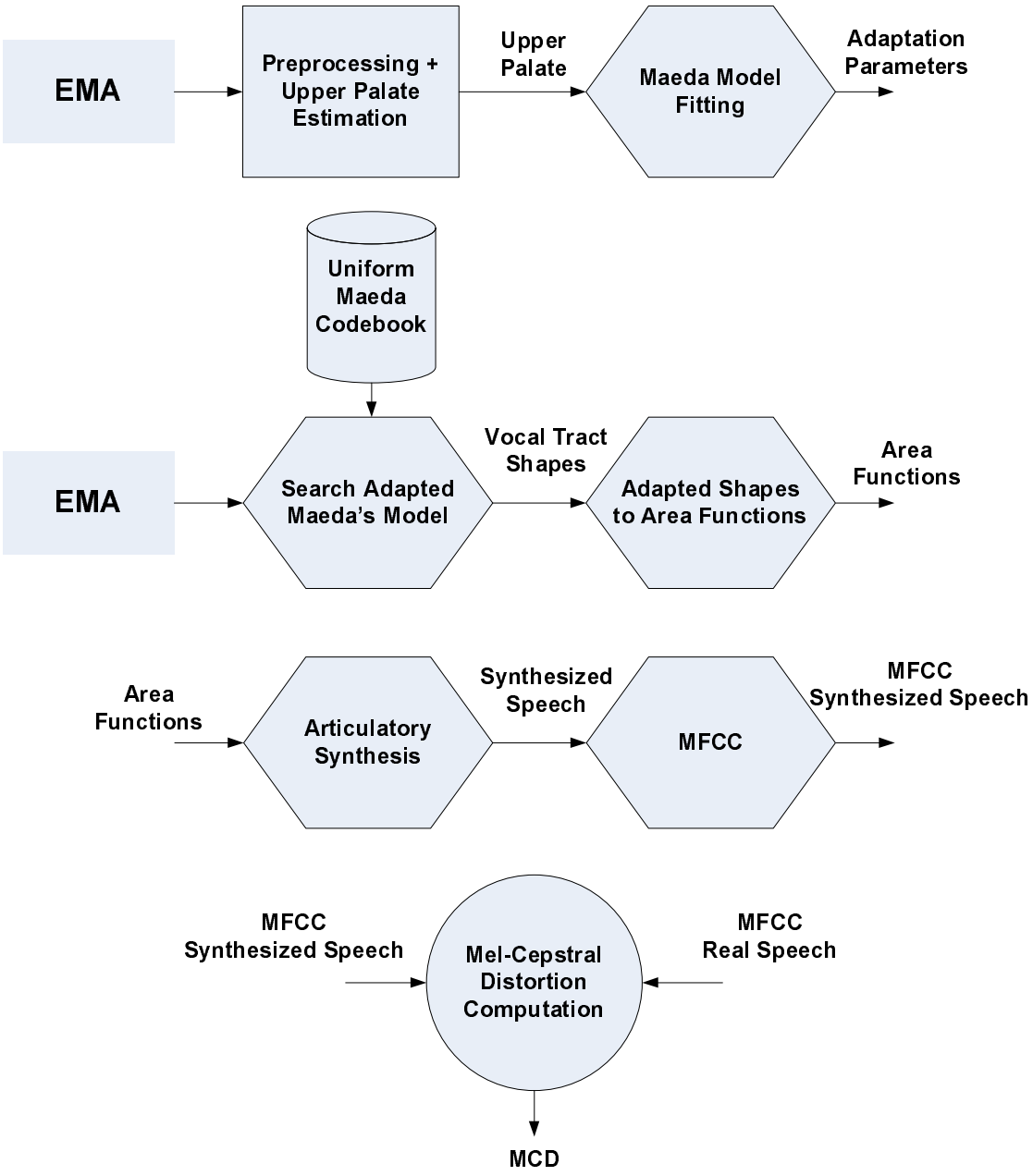


Figure 4.17: *Basic building blocks for the adaptation, search, and synthesis procedures. The diagram summarizes how we synthesize speech from EMA measurements and compare it to real speech.*

4.7.1 Adaptation Results

Tables 4.2 and 4.3 show the resulting adaptation parameters for the male and female speakers respectively. We also show the computed Lip_{sep} value. GD is the geometric distance between the estimated EMA upper palate and the Maeda model upper palate. Note that adaptation reduces significantly this value for almost all the speakers. The exception is speaker "mjjn0", where the baseline model matches the EMA palate better, but it exhibits geometric violations. We exclude setups with such violations from our search. Examples of such violations are situations where the EMA points fall above the upper palate given by Maeda's model. In the adaptation we follow here, we relax this constraint in the oral region but kept it in the polar region.

After finding the set of adaptation parameters for each speaker, we remove the vocal tract shapes that are uncommon in English from the generic codebook as described in Subsection 4.4.1. Different speakers have different codebooks according to their vocal tract dimensions. In general, the shorter the vocal tract length is (higher values of k), the smaller the codebook size.

Table 4.2: *Number of utterances, adaptation parameters, average geometric distance, and codebook size for male speakers.*

Speaker	msak0	maps0	mjjn0	ss2404
Number of Utterances	458	450	460	328
k	-0.12	-0.12	-0.02	-0.04
β (rad)	-0.0087	-0.0175	0.0262	0.0209
dx (cm)	-0.4	-0.6	-0.6	-0.8
dy (cm)	0.6	0.8	-0.6	0.2
Lip_{sep} (cm)	2.00	1.08	2.68	1.10
GD (cm)	0.67	0.70	0.29	0.77
GD + Adaptation (cm)	0.20	0.14	0.39	0.45
Codebook Size	18875	18875	18050	18200

Figures 4.18 and 4.19 show the adaptation results for the male and female speakers

Table 4.3: *Number of utterances, adaptation parameters, average geometric distance, and codebook size for female speakers.*

Speaker	fsew0	ffes0	faet0	falh0	fjmw0
Number of Utterances	460	459	456	130	281
k	0.06	0.14	-0.14	-0.1	0.2
β (rad)	0.0175	-0.0087	-0.0175	-0.0175	-0.007
dx (cm)	-0.4	-0.2	0	-0.6	0.4
dy (cm)	0.2	1.2	1.4	1.6	0.8
Lip_{sep} (cm)	1.43	1.20	1.45	1.23	1.25
GD (cm)	0.56	0.96	1.13	1.24	0.71
GD + Adaptation (cm)	0.27	0.23	0.25	0.24	0.27
Codebook Size	17325	16400	18050	18700	16225

respectively. The smoothed scatter of the EMA data is shown in yellow, black, and red. The adapted Maeda grid is superimposed on the EMA scatter (red grid lines). Note the fit of Maeda's model upper palate to the EMA data and the length and tilt of the vocal tract in each figure. In general, it is shown in the plots that males have longer vocal tract than females reflecting the estimated k parameter.

4.7.2 Synthesis Results

We compare the MCD results derived using the model adaptation and search techniques described in this chapter with the MCD results derived following the approach in Section 3.1. The latter approach, which we consider as our baseline, maps EMA data to Maeda parameters using a heuristic mapping from EMA directly to Maeda parameters rather than using the model itself. For example, p_5 is simply the normalized distance between the UL and LL. It relies on the meaning Maeda attributed to the vocal tract shape parameters as described in [9] and summarized in Table 2.2. There is no geometric search involved in the mapping. The consequential difference between the two approaches is in the values for the areas and lengths of the acoustic tubes and in the nasal tract opening

area estimation.

We estimate the nasal tract opening area from the EMA measurements of the velum sensor VL as described in Subsection 4.3.6. Reliable velum sensor measurements are available only for speakers "msak0", "fsew0", "ffes0", "falh0", and "fjmw0". We set the area to zero for the rest of the speakers. We also set the nasal tract area to zero for the all the speakers in the baseline MCD computations.

For each frame in the test utterances we synthesize speech following the approach described in Sections 4.4 and 4.5. Then we extract Mel-cepstral coefficients (MFCC) from the synthesized and the real speech respectively and compute the MCD between them for each frame. We compute the average MCD for frames from vowels, fricatives, nasals, and all the phones. When computing the average MCD, we include the MCD of the frames at the onset, middle, and offset of phones. Phonetic segmentation has been automatically extracted beforehand.

For the adapted vocal tract experiment, we achieve 6.54% relative reduction over baseline in MCD for vowels, 9.10% for nasals, and 4.75% in total for speaker "msak0". The MCD for fricatives becomes worse with adaptation. Table 4.4 presents these results.

Table 4.4: *MCD results: the absolute and relative differences are between the baseline experiment without adaptation and the adapted vocal tract approach developed in this chapter. Detailed results are presented for speaker "msak0".*

MCD Results	Vowels	Fricatives	Nasals	All
Frame Count	14517	6660	4215	42294
No Adaptation	6.95	8.66	7.85	8.07
Adapted Vocal Tract	6.50	8.78	7.13	7.68
Absolute Difference	0.46	-0.11	0.71	0.38
Relative Difference	6.54%	-1.30%	9.10%	4.75%

For speaker "fsew0", we achieve 1.33% relative reduction over baseline in MCD for vowels, 20.04% for nasals, and 2.18% in total. The MCD for fricatives also got worse with adaptation here. Table 4.5 presents these results.

Table 4.5: *MCD results: the absolute and relative differences are between the baseline experiment without adaptation and the adapted vocal tract approach developed in this chapter. Detailed results are presented for speaker "fsew0".*

MCD Results	Vowels	Fricatives	Nasals	All
Frame Count	14514	6672	4208	42929
No Adaptation	7.28	7.76	9.31	8.11
Adapted Vocal Tract	7.18	8.50	7.45	7.93
Absolute Difference	0.10	-0.75	1.87	0.18
Relative Difference	1.33%	-9.64%	20.04%	2.18%

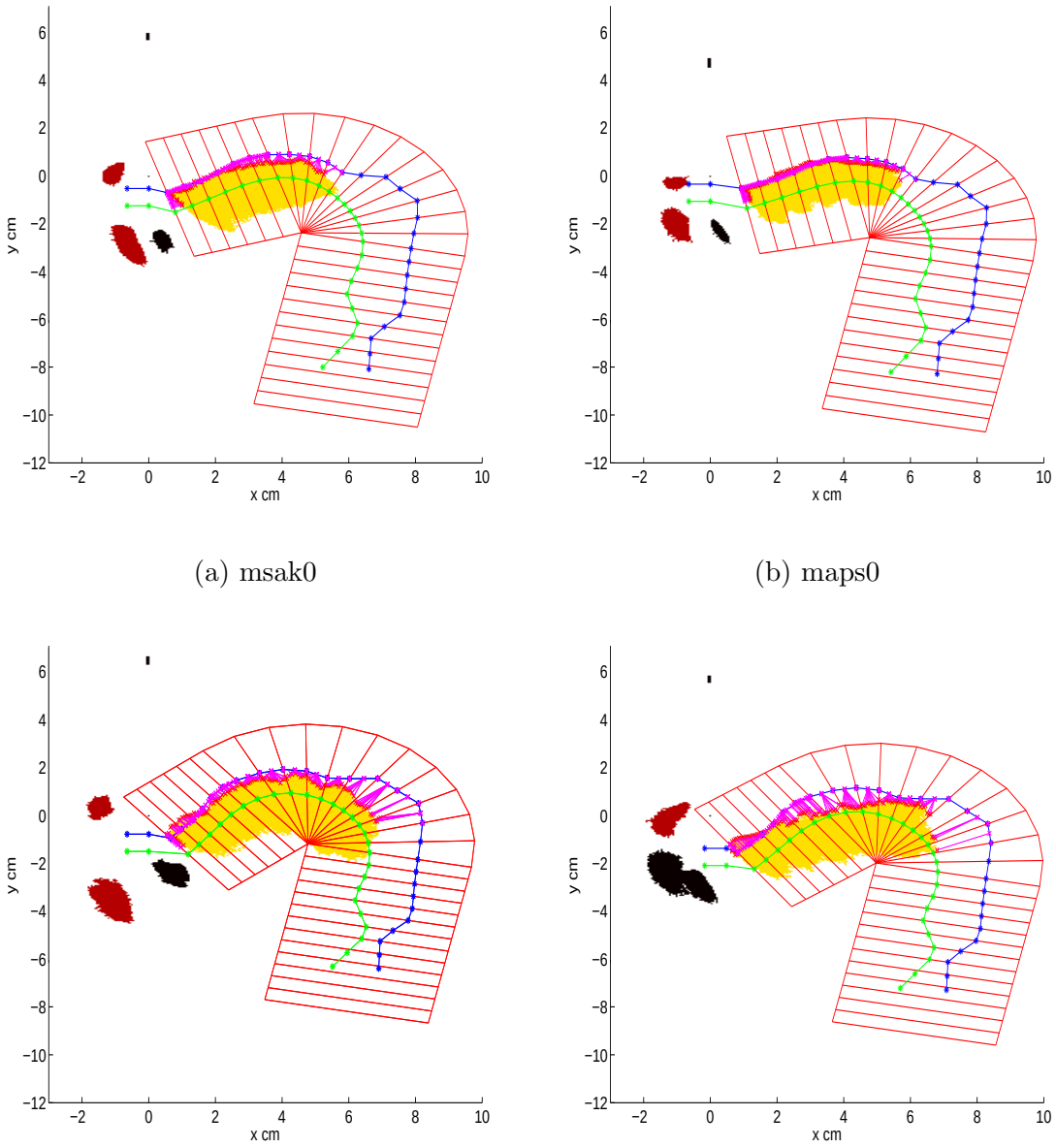
Figure 4.20 shows the average MCD results from all the test utterance for the male speakers. It is clear in the plot that reduction in MCD is achieved in adaptation for vowel frames for most speakers. In addition, reduction in MCD is achieved for all frames (including vowel frames) for all the male speakers. Similar results are achieved for the female speakers as shown in Figure 4.21. In the female cases, the MCD is in general higher than in the male cases.

Figure 4.22 shows the average MCD for the nasal frames from speakers with reliable velum sensor measurements. Compared to baseline, the adaptation improves the MCD considerably for all the speakers.

4.8 Conclusions and Future Work

We presented a principled approach for mapping EMA data to vocal tract shapes for the task of speech synthesis by a physical model of the vocal tract. We used the EMA data to adapt Maeda's vocal tract model to all available speakers in the MOCHA database. We presented a way for searching for the best fitting vocal tract contours. Experiments showed improvement in synthesis over the baseline approach we adopted without adaptation. We also showed how to estimate the nasal tract opening area from the velum sensor. In all experiments, we synthesized continuous speech waveforms solely from EMA. To our knowledge, this is the first work that synthesizes continuous speech utterances from

EMA data. In the future, we would like to use the ElectroPalatoGraph (EPG) data provided in MOCHA to improve modeling of fricatives and the constriction location.



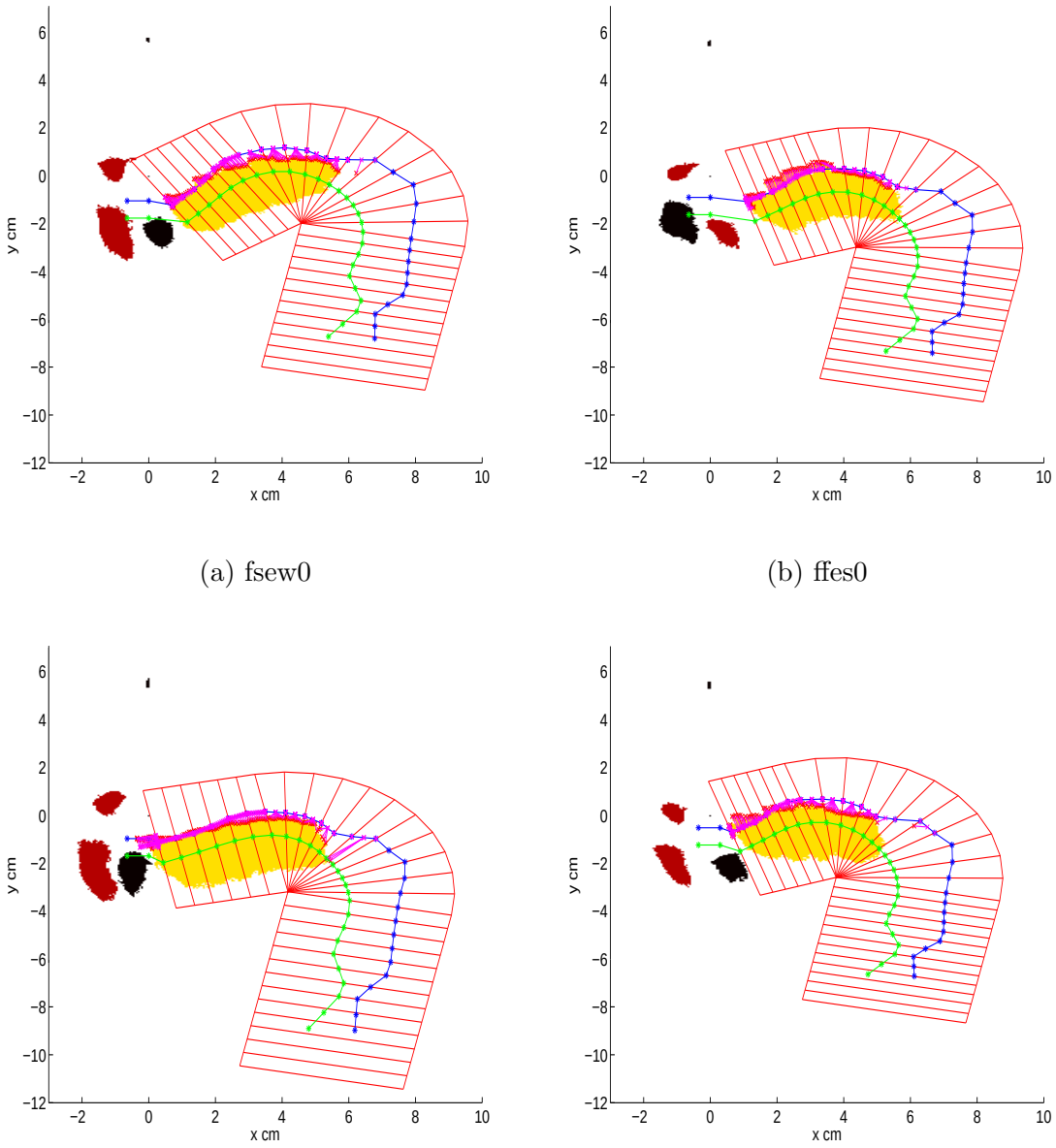
(a) msak0

(b) maps0

(c) mjjn0

(d) ss2404

Figure 4.18: *Adaptation results for the male speakers. The smoothed scatter of the EMA data is shown in yellow, black, and red. The adapted Maeda model grid is superimposed on the EMA scatter (red grid lines). The green contour is for the steady state Maeda lower contour and the blue contour is the adapted Maeda upper contour. The top-most point in each figure corresponds to the bridge of the nose cluster.*



(a) fsew0

(b) ffes0

(c) faet0

(d) fjmw0

Figure 4.19: *Adaptation results for the female speakers. The smoothed scatter of the EMA data is shown in yellow, black, and red. The adapted Maeda model grid is superimposed on the EMA scatter (red grid lines). The green contour is for the steady state Maeda lower contour and the blue contour is the adapted Maeda upper contour. The top-most point in each figure corresponds to the bridge of the nose cluster.*

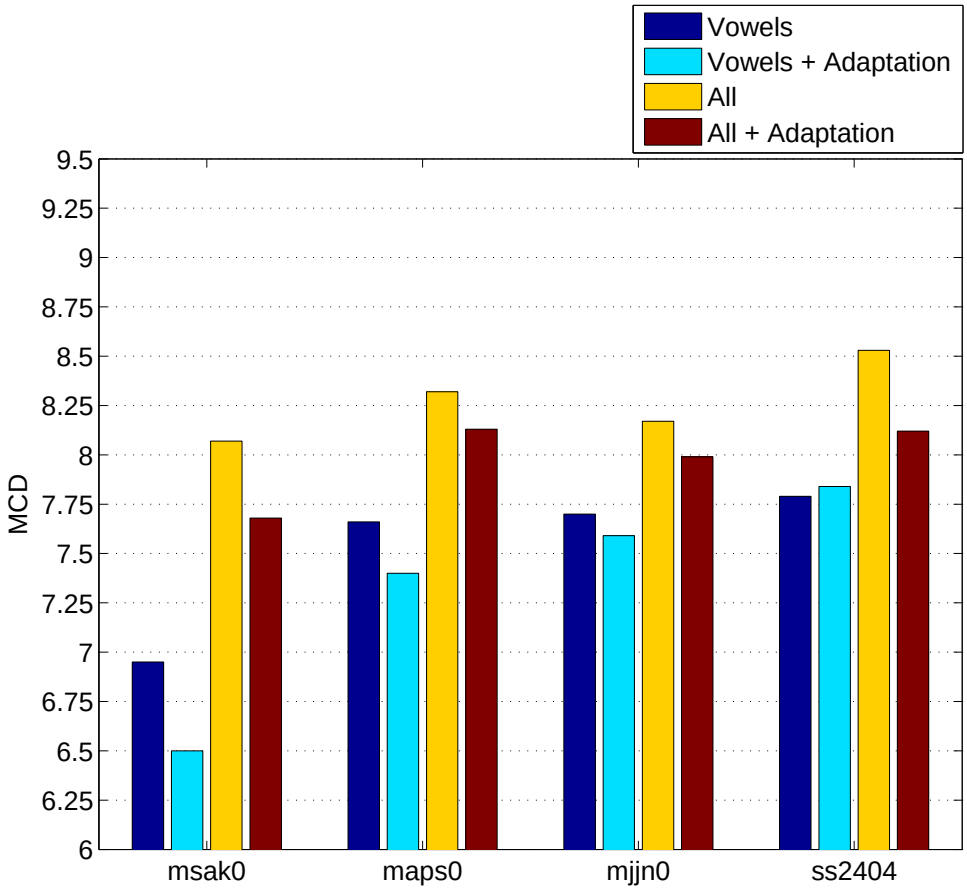


Figure 4.20: *Average MCD results for all test utterances of the male speakers, showing results for vowel frames and all frames.*

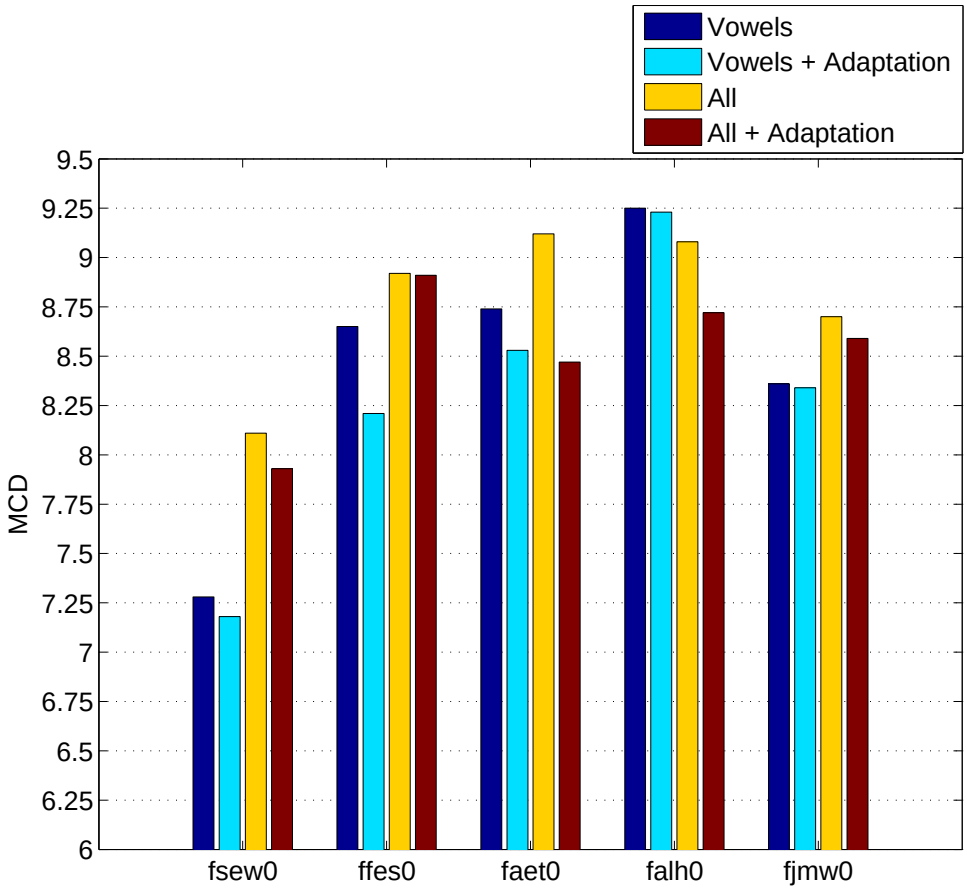


Figure 4.21: *Average MCD results for all test utterances of the female speakers, showing results for vowel frames and all frames.*

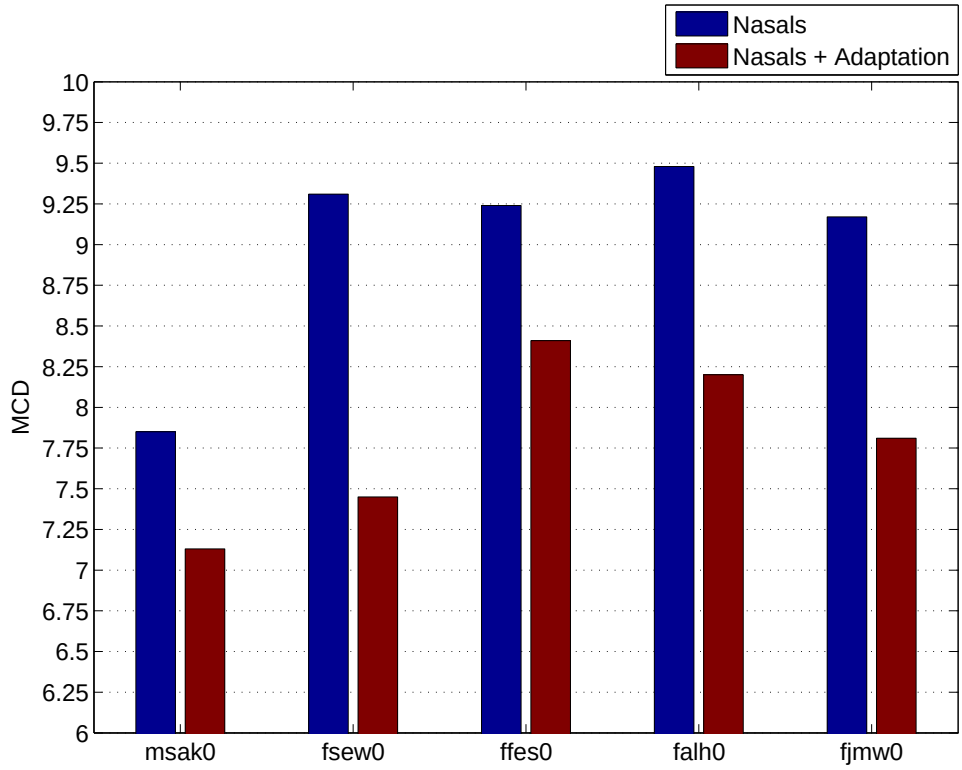


Figure 4.22: *Average MCD results for speakers with reliable velum sensor measurements, showing results for nasal frames only.*

Chapter 5

Dynamic Framework Using the Analysis-by-Synthesis Distortion Features

5.1 Introduction

Articulatory modeling [12, 32] is used to incorporate speech production information into automatic speech recognition (ASR) systems. It is believed that solutions to the problems of co-articulation, pronunciation variations, and other speaking style related phenomena reside in how accurately we capture the production process.

Our goal in this chapter is a dynamic articulatory framework for speech recognition where the model states are collections of possible vocal tract shapes. In previous work we have presented two key components that enable us to address this goal. In Chapter 3 we propose new features that convey articulatory information. Using a physically-motivated codebook of vocal tract shapes to derive analysis-by-synthesis distortion features is shown to provide improvements in phone classification accuracy. In Chapter 4 we show how to derive realistic vocal tract shapes from the EMA data in the MOCHA database. We rely solely on the EMA data to perform the synthesis, in contrast to the more common approach of learning a statistical mapping between the EMA and acoustic recordings

from parallel recordings of the two [33, 34]. The combination of the adapted vocal tract models and a physiologically-based articulatory synthesizer, *e.g.* the Sondhi and Schroeter synthesizer [8], models the physical speech production process for a speaker.

Previous and current approaches to the incorporation of articulatory models into speech recognition [32, 35, 16] have used representations of phonological features derived from the transcript through linguistic expert knowledge. This representation may not represent the actual underlying articulatory phenomena that produced the speech signal. The same speech may be produced differently. In the work reported in this chapter we use EMA measurements as a means for capturing the ground truth articulatory phenomena. EMA provides exact information about the articulators' movements rather than abstract information derived from text.

Our aim here is to build upon our previous work [28, 36] and incorporate the distortion features in a dynamic framework whose inner states are vocal tract shapes. These vocal tract shapes are derived in a principled geometric fashion as described in [36]. We then synthesize speech using the adapted vocal tract models for each speaker to closely mimic the incoming speech signal. The distortion between the incoming speech and the speech synthesized from the articulatory states is used to dynamically traverse the articulatory space. This framework not only constrains the set of possible vocal tract shapes for each phone, but is also capable of modeling the articulatory dynamics and imposing further constraints in a probabilistic fashion.

The set of all possible vocal tract shapes is quantized into a codebook of shapes, each represented by a vector of Maeda parameters. For a given phone, only a restricted region in the space of vocal tract shapes represented by a subset of the codewords is active. Hence we would only need the distortion features associated with these codewords. In this chapter, we show how we can learn this subset. We use two approaches, one that uses the EMA data (*i.e.* ground truth) and another that is data driven. Both approaches yield a solution that zeros out the weights associated with codewords not relevant to the phone in study.

In order to incorporate the distortion into the probabilistic framework, we need to

convert it to a form of probability. The key point here is to apply a density function that penalizes higher distortions (*e.g.* exponential density). The lower the distortion from a given codeword, the more likely it is to be the codeword that has generated the incoming frame of speech. Another way of looking at this is saying that we only care about the codewords that reflect the true articulatory dynamics of the phone in study. We refer to this as the "OR" approach. In the "AND" approach in Chapter 3 we included the distortion from all the codewords, whether relevant to the phone or not, and that helped provide better discrimination and classification accuracy.

Using a subset of distortion features for a particular phone is a means of applying articulatory knowledge to constrain the recognition problem. It also reduces the amount of computation involved in using all the distortion features as we did in Chapter 3. The sparsity in the estimated weights of the codewords for each phone reflects the reductions in computations. Since each state is a collection of articulatory states, then the state itself has an articulatory meaning reflected in the weights attributed to the codewords. The transition from one state to another then reflects articulatory movements. This framework can be easily expanded to incorporate articulatory dynamics in different ways.

In this chapter we present our design of the dynamic framework and the observation probability model used. We present several different methods for model training and initialization. We analyze the sparsity of the solutions to which the algorithms converge and we present preliminary phone classification results.

5.1.1 Analysis of the Features

Figure 3.2 shows the distortion computed from all the codewords for each frame in an utterance. In reality, using the EMA measurements we can tell which codeword is active for each frame. Hence, we can get an exact estimate of the distortion distribution if we take the codeword identity into consideration. In this section we show the distribution of the distortion for two phones 'O' an 'II' for the speaker "msak0" from MOCHA. Phonetic segmentation has been pre-computed and we show the distribution from all the frames that belong to segments of these phones. The distortion from each codeword

is computed from all the utterances. Figure 5.1 shows the distortion distribution for codewords 45 through 48 for the two phones, where count reflects the number of frames. Each plot contains two histograms of $distortion(codeword = j|phone = C)$, where $j = \{45, 46, 47, 48\}$ and $C = \{$ 'O','II' $\}$. From the figure we can see that Codeword 45 synthesizes speech that is closer to 'O' than 'II' due to the overall lower distortions for most available examples of these phones. Recall that we are following a context-independent analysis of phonemes. In addition to Codeword 45 synthesizing speech closer to 'O' than the other codewords, the other codewords synthesize speech closer to 'II' following the same reasoning. In the right hand side of the figure we plot the square of the distortion for further insights that help us choose the correct probability density function to model the distortion.

Figure 5.2 shows the distortion distribution for codewords 45 through 48 for the two phones. In the figure we use the information from the EMA that provides ground truth information on which frames each of these codes has occurred at. We only plot the distribution for these frames. Each plot contains two histograms of $distortion(codeword = j|phone = C, truecode = j)$, where $j = \{45, 46, 47, 48\}$ and $C = \{$ 'O','II' $\}$.

We first see that the count of the distortion is less than in Figure 5.1 since we are selecting the frames where each of these codeword truly occurs. From Figure 5.2 we can also see that codeword 45 is more active for phone 'O' than 'II' because the overall count of the number of frames where this codeword occurs is more for 'O' than 'II'. This also explains the overall lower distortion for 'O' than 'II' as shown for this codeword in Figure 5.1. Similarly, codewords 46 and 47 are more active for phone 'II' and yielded a lower overall distortion in Figure 5.1. Codeword 48 is seldom active for both phones, more for phone 'O' than 'II', yet overall the synthesis seems to be closer to 'II'.

5.1.2 Feature Normalization Techniques

We have tried three techniques for features normalization to help with choosing the probability density function to model the distortion. The top plot in Figure 5.3 shows

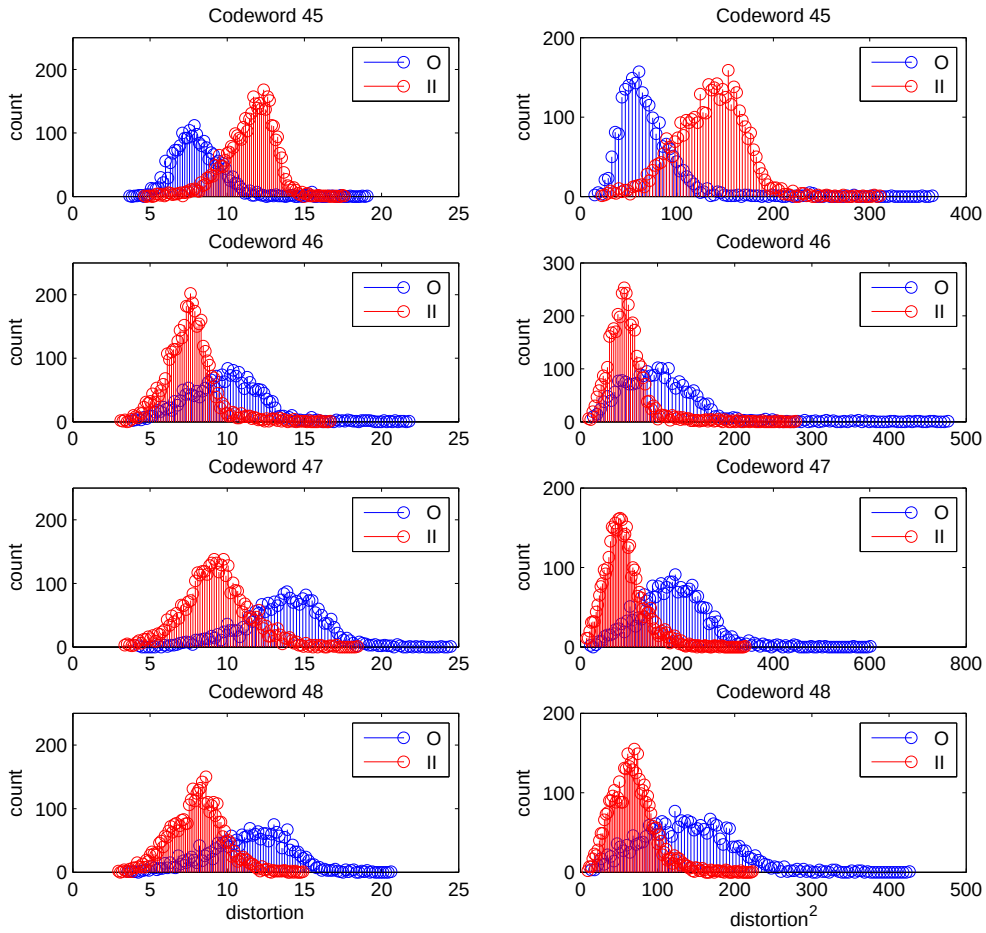


Figure 5.1: *Distortion histogram for codewords 45-48 of phones 'O' and 'II' for speaker "msak0". The histogram of the square of the distortion is shown on the right. The count reflects the number of frames.*

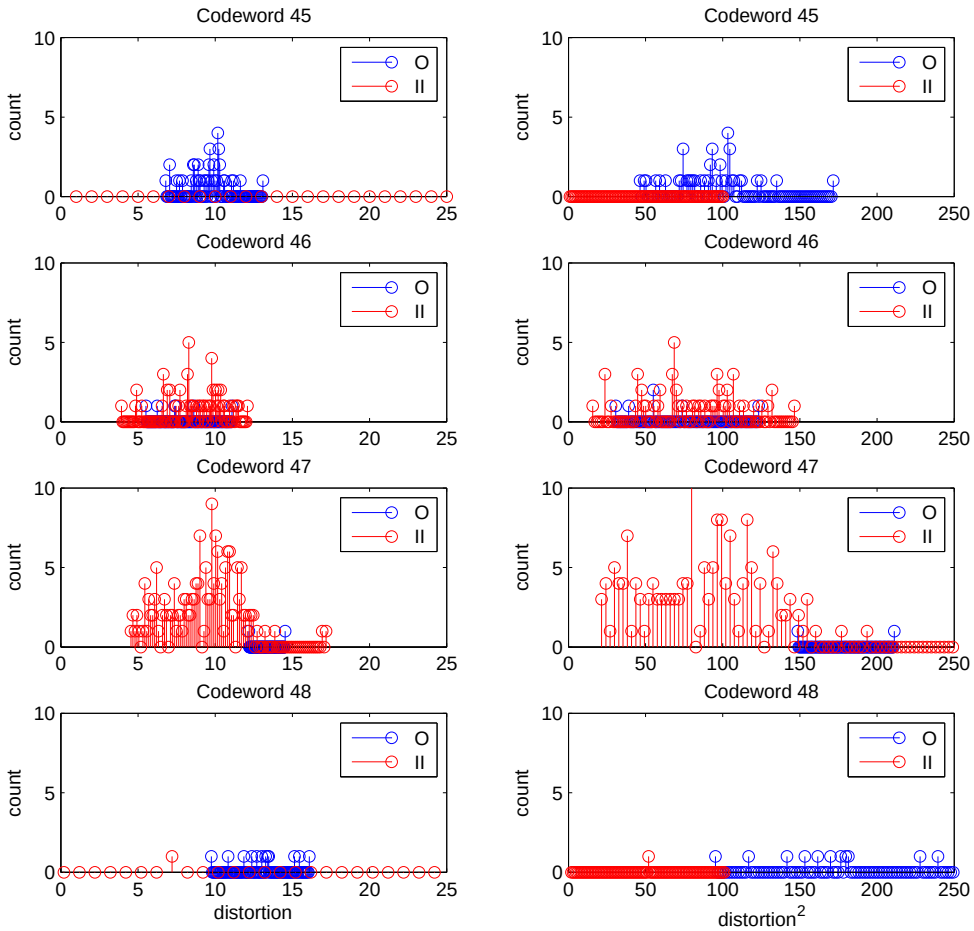


Figure 5.2: *Distortion histogram for codewords 45-48 of phones 'O' and 'II' for speaker "msak0", knowing ground-truth from EMA. The histogram of the square of the distortion is shown on the right. The count reflects the number of frames.*

the distortion from codeword 47 for the same data used above without using ground-truth information.

Minimum-Frame Codeword Distortion Normalization: minC

This normalization, *minC*, works on a frame-by-frame basis. At a given time instant it finds the minimum distortion from all the codewords. This is then subtracted from the rest of the distortions for that time instant. The motivation is that if the minimum distortion is from the correct codeword, then this distortion is mainly due to synthesis inaccuracy. Subtracting this value will make the distortion from the correct codeword zero. We add a small number (0.01) to avoid numerical instabilities especially when we apply different probability distributions to the features. The result of this normalization is shown in the second plot from the top in Figure 5.3.

Mean-Utterance Codeword Distortion Normalization: meanU

This normalization, *meanU*, works on each feature separately over the whole utterance. It is analogous to Cepstral Mean Normalization (CMN [25]) that works on the distortion from each codeword. It computes the mean of each feature over the utterance and subtracts it out. The result of this normalization is shown in the third plot from top in Figure 5.3.

Minimum-Utterance Codeword Distortion Normalization: minU

This normalization, *minU*, is similar to *meanU*, except that it subtracts the minimum of each feature over the utterance instead of the mean. The result of this normalization is shown in the bottom plot in Figure 5.3.

5.2 Dynamic Framework

As mentioned in the introduction of this chapter, we desire a dynamic framework that reduces computations and is capable of modeling the dynamic constraints of articulators, which we believe helps in improving speech recognition accuracy. In addition, this

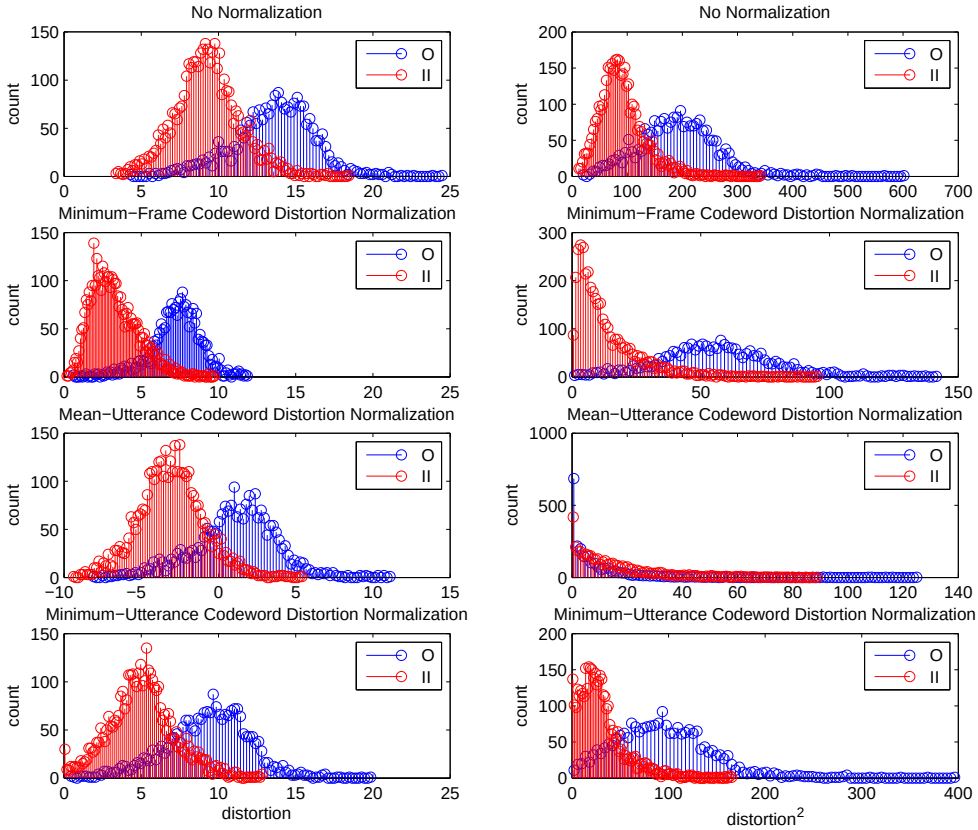


Figure 5.3: *Different normalization techniques for the distortion from codeword 47 for speaker "msak0".*

framework continues to impose the same set of vocal tract shapes on the phones as the analysis-by-synthesis distortion framework. In fact, we use the same features but without compression with LDA.

Figure 6.3 shows the ultimate dynamic framework where the states are shared by all the phones. The states are the codewords of vocal tract shapes. Weighting for each state is based on the phone identity. The transition from one state to another reflects articulatory movements. The output of each state is the distortion between the incoming speech and the speech synthesized from this state (codeword). We seek the path in the articulatory space that minimizes the error due to two measures, an articulatory and a distortion measure; $error = acoustic\_distortion + articulatory\_distance$. The articulatory measure is derived from the transition from one codeword to another. So, we seek smooth paths that are more likely for a given phone sequence. The distortion measure is the total distortion between the speech synthesized from the codewords and the incoming speech. We seek paths that generate speech that is closer to the incoming speech. These paths are directly related to the phone identities since the model parameters are functions of phones.

5.3 Mixture density function for modeling the state output observation probability

In order to incorporate distortion into the probabilistic framework, we need to convert it to a form of probability. The key point here is to apply a density function that will penalize higher distortions. The lower the distortion from a given codeword, the more likely it is to be the codeword that has generated the incoming frame of speech. Another way of looking at this is saying that we only care about the likely codewords that reflect the true articulatory dynamics of the phone in study. In this framework, we do not care about the distortion from the incorrect codewords since it will not reflect correct dynamic information. We refer to this as the "OR" approach as mentioned in Section 5.1. In the "AND" approach, we included the distortion from all the codewords, whether relevant to the phone or not, and that helped provide better discrimination and classification

accuracy. We hope that looking at the likely codewords alone for a given phone will maintain this performance, provided that we take advantage of dynamical constraints and present a less computationally-costly framework.

In this section we model the set of codewords as a mixture probability density function. We show how we can learn the subset of relevant codewords for a given phone. We use two techniques, one that uses the EMA data (*i.e.* ground-truth) and the other that is data driven. Both techniques yield a solution that will zero out the weights associated with codewords not relevant to the phone in question.

The acoustic distortion between the speech synthesized from each of the codewords and the incoming speech is $D = \{d_1, d_2, \ldots, d_M\}$, where M is the number of codewords $CD = \{cd_1, cd_2, \ldots, cd_M\}$ for state S_1 and observation x. We follow a soft decision approach in which we estimate a set of weights for each phone $\{w_1, w_2, \ldots, w_M\}$ that defines the contribution of each codeword as follows:

$$\begin{aligned} P(x|S_1) &= \sum_{j=1}^{M} P(x, cd_j|S_1) \\ &= \sum_{j=1}^{M} P(cd_j|S_1)P(x|cd_j, S_1) \\ &= \sum_{j=1}^{M} w_{1j}P(x|cd_j, S_1) \end{aligned} \tag{5.1}$$

We use the expectation-maximization algorithm (EM, Dempster *et al.* [37]) to estimate the weights and the probability distribution parameters used to model the observation probability of the distortion of each codeword.

5.3.1 Weights Estimation from Audio

We use the EM algorithm [37] to derive the mixture weights for Equation 5.1 for a given *phone C*. Refer to Bilmes [38] for the derivation of EM for a mixture of Gaussian densities. In our setup, the set of model parameters to estimate ϕ is $\{w_1, w_2, \ldots, w_M, \theta_1, \theta_2, \ldots, \theta_M\}$. The exact set of $\{\theta_j, j = 1 : M\}$ depends on the observation probability used. We assume that these parameters are derived from a set of

$\{x_i, i = 1 : N\}$ audio data points belonging to *phone* C and drop the phone identity from the equations. From Bilmes [38], the maximum likelihood solution is:

$$\begin{aligned} w_j^t &= \frac{1}{N}\sum_{i=1}^{N} P(cd_j|x_i, \theta_j^{t-1}) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{w_j^{t-1} P(x_i|cd_j, \theta_j^{t-1})}{\sum_{k=1}^{M} w_k^{t-1} P(x_i|cd_k, \theta_k^{t-1})} \\ &= \frac{w_j^{t-1}}{N}\sum_{i=1}^{N} \frac{P(x_i|cd_j, \theta_j^{t-1})}{\sum_{k=1}^{M} w_k^{t-1} P(x_i|cd_k, \theta_k^{t-1})} \end{aligned} \tag{5.2}$$

Starting with a flat initialization as described in Equation 5.3, we iterate until the values of $\{w_j, j = 1 : M\}$ converge.

$$w_j = \frac{1}{M}, (j = 1 : M) \tag{5.3}$$

5.3.2 Weights Estimation using EMA

A forced-alignment of the audio and the transcript provides the phonetic segmentation for the MOCHA database. For each frame we know which *phone* C it corresponds to. From the EMA data we can also know which codeword it corresponds to. Hence we can count the codewords for each phone and estimate the probability of being in a given codeword for this phone. This estimate can be used as a prior to estimating the weights from audio and for initialization purposes in other databases where the EMA data are not available.

$$w_j = \frac{count\_frames(phone = C, truecode = cd_j)}{total\_frames(phone = C)} \tag{5.4}$$

5.3.3 Output Distortion Probability

For each frame of speech $\{x_i, i = 1 : N\}$, we compute the distortion $\{d_{ij}, j = 1 : M\}$ for each codeword $\{cd_j, j = 1 : M\}$. We consider three different observation probability densities: exponential, Rayleigh, and Gaussian. As distortion decreases the codeword

becomes more likely to have produced the speech. This is reflected in our choice of the probability densities, except for the Gaussian, which is the traditional density used in HMMs.

Exponential Probability Density

$$P(x_i|cd_j) = \lambda_j \exp^{-\lambda_j d_{ij}^2} \tag{5.5}$$

Rayleigh Probability Density

$$P(x_i|cd_j) = \frac{d_{ij}}{\sigma_j^2} \exp^{-\frac{d_{ij}^2}{2\sigma_j^2}} \tag{5.6}$$

Gaussian Probability Density

$$P(x_i|cd_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp^{-\frac{(d_{ij}-\mu_j)^2}{2\sigma_j^2}} \tag{5.7}$$

The top left plot of Figure 5.4 shows the histogram of the distribution of the squared distortion from codeword 47 for speaker "msak0" and for phones 'O' and 'II'. The bottom left shows the exponential densities fitted into the two distributions. The top right plot shows the distribution of the $minC$ normalized distortion. The bottom right shows the Rayleigh density fitted to this distribution.

A perfect fit to the distribution of the distortion is not a guarantee of optimal performance due to the way we set up this problem. The requirement is that the smaller the distortion is the higher the likelihood of the codeword should be and vice versa. Using the model we describe in this chapter, experiments have shown that the exponential density does slightly better than the Rayleigh density in phone classification. In the following, we will only show the equations for the exponential distribution.

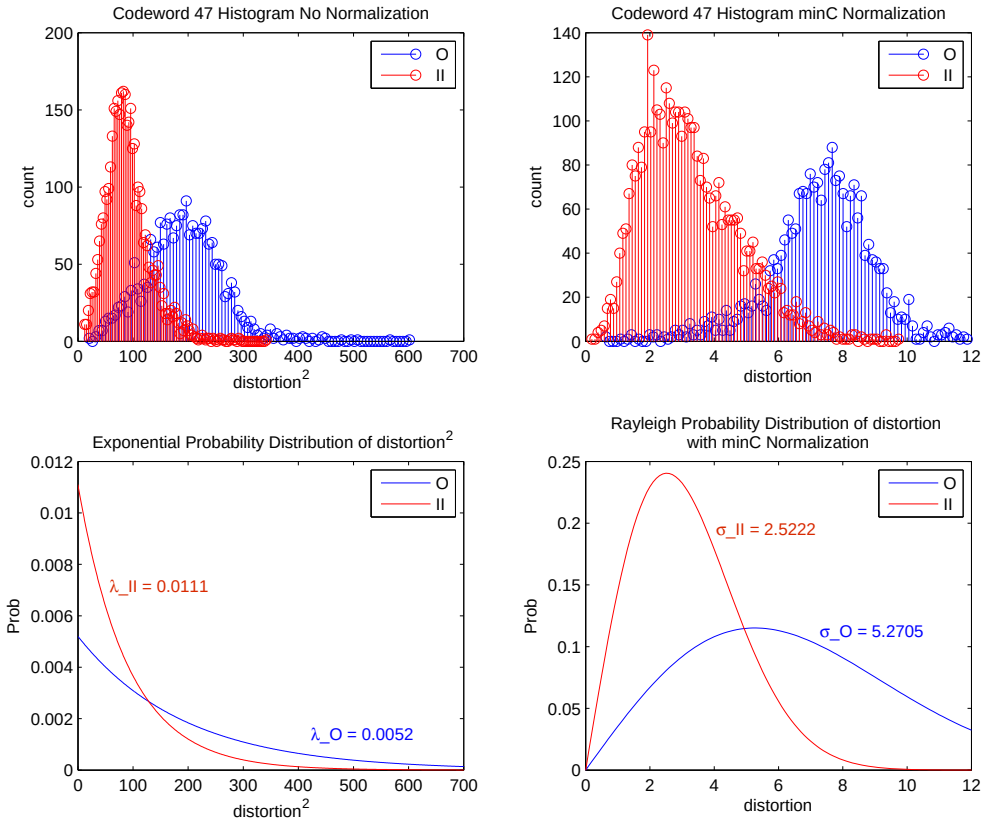


Figure 5.4: *Exponential density function for modeling the square of the distortion from codeword 47 for speaker "msak0" on the left. Rayleigh density function for modeling the minC normalized distortion from codeword 47 for speaker "msak0" on the right. The count reflects the number of frames.*

5.3.4 Estimating the Lambdas of the Exponential Distribution from Audio

The set of parameters to estimate is $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ for a given *phone* C. Including the estimation of lambdas in the EM, the update equation is:

$$\begin{aligned} \lambda_j^t &= \frac{\sum_{i=1}^{N} P(cd_j|x_i, \theta_j^{t-1})}{\sum_{i=1}^{N} d_{ij}^2 P(cd_j|x_i, \theta_j^{t-1})} \\ &= \frac{\sum_{i=1}^{N} \frac{w_j^{t-1} P(x_i|\lambda_j^{t-1})}{\sum_{k=1}^{M} w_k^{t-1} P(x_i|\lambda_k^{t-1})}}{\sum_{i=1}^{N} d_{ij}^2 \frac{w_j^{t-1} P(x_i|\lambda_j^{t-1})}{\sum_{k=1}^{M} w_k^{t-1} P(x_i|\lambda_k^{t-1})}} \end{aligned} \tag{5.8}$$

Starting with a flat initialization as described by Equation 5.9, we iterate until the values of $\{\lambda_j, j = 1 : M\}$ converge.

$$\lambda_j = \frac{1}{mean(d_{ij}^2|phone = C)}, (i = 1 : N) \tag{5.9}$$

5.3.5 Estimating the Lambdas of the Exponential Distribution from EMA

As mentioned in Subsection 5.3.2, using forced-alignment and EMA we get the codewords identities. EM has been used before to overcome this missing information from the audio. Hence, we can estimate $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ directly for each codeword without corrupting the estimation by data generated from other codewords. This estimate can be used as a prior to estimating $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ from audio and for initialization purposes in other databases where the EMA data are not available.

$$\lambda_j = \frac{1}{mean(d_{ij}^2|phone = C, truecode = cd_j)} \tag{5.10}$$

5.3.6 Classification using Estimated Parameters

To classify each segment of speech $\mathbf{X}$ into the most likely phone $\hat{C}$, we compute:

$$\begin{aligned} \hat{C} &= \text{argmax}_C P(C|\mathbf{X}) \\ P(C|\mathbf{X}) &= (\prod_{i=1}^{L} P(x_i|C))P(C) \\ logP(C|\mathbf{X}) &= \sum_{i=1}^{L} log(\sum_{j=1}^{M} w_j P(x_i|cd_j, \theta_j)) + log(P(C)) \end{aligned} \tag{5.11}$$

where $P(C)$ is the prior probability of phone C.

5.3.7 HMM Formulation for Exponential Observation Probabilities

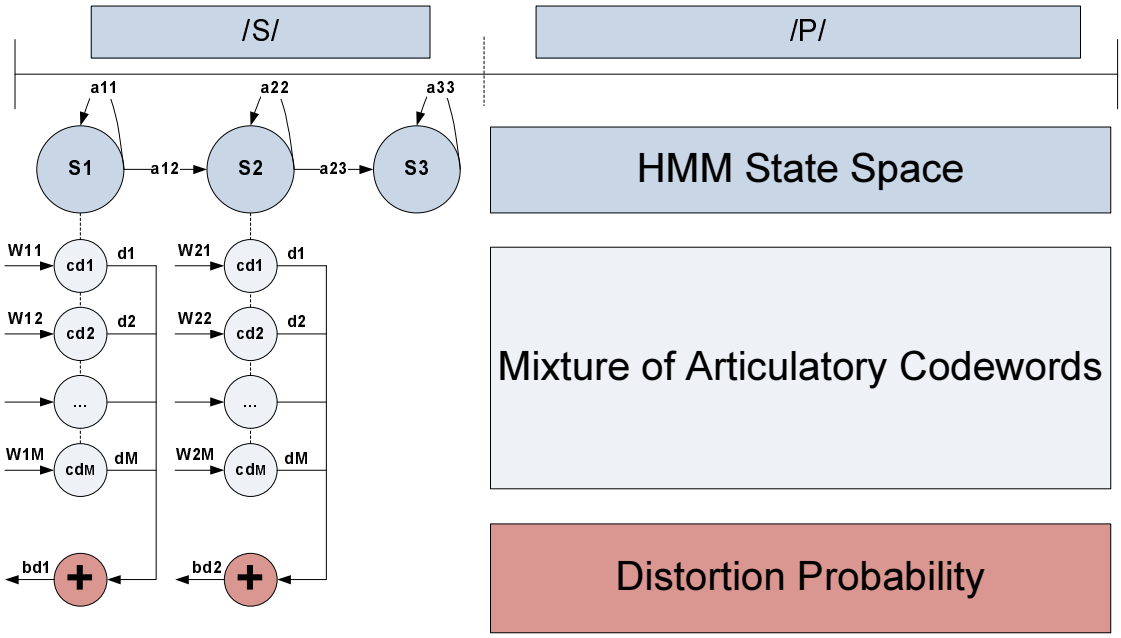


Figure 5.5: *Mixture probability density for the distortion features in a dynamic framework.*

To model dynamics, we use a three-state left-to-right HMM for each phone $S = \{S_1, S_2, S_3\}$. Each state will model part of the phone and hence converge to different values for the parameters used to model the distribution of the features associated with it. HMMs are defined by three main parameters $\phi = \pi, A, B$, where π is the vector of state initial probabilities, A is the transition matrix, and B is the matrix containing the

values of the parameters describing the output observation density. Refer to [39, 38] for a detailed description of HMMs.

We use the basic formulation of HMM parameters in [39, 38], with some modification to reflect the observation densities used to model the distortion features. To enforce the left-right three-state topology we initialize the vector of initial state probabilities $\pi_i, (i = 1 : N)$, where N is the number of states $N = 3$, as $\pi = [1, 0, 0]$. We also initialize the transition probability matrix $a_{ij}, i, j = 1 : N$ to $a_{ij} = [0.5, 0.5, 0; 0, 0.5, 0.5; 0, 0, 1]$. The EM parameters we define for each segment are $bd_i(t), \gamma_i(t)$, and $\gamma_{il}(t)$. The HMM model parameters computed from all the segments are w_{il}, and λ_{il}. Each segment $\mathbf{X}$ of a given phone is made of observations $x(t), t = 1 : T$, where T is the number of frames in each segment. Assuming M mixtures, the output observation probability from each state is given by Equation 5.12.

$$\begin{aligned} bd_i(t) &= P(x(t)|S(t) = i) \\ &= \sum_{l=1}^{M} w_{il}\lambda_{il} \exp^{-\lambda_{il} d_l^2(t)} \end{aligned} \tag{5.12}$$

Another parameter we modified is $\gamma_{il}(t)$ as shown in Equation 5.13. $m(t)$ is the mixture at time t. $\gamma_i(t)$ is as described in [38].

$$\begin{aligned} \gamma_i(t) &= P(S(t) = i|\mathbf{X}, \phi) \\ \gamma_{il}(t) &= P(S(t) = i, m(t) = l|\mathbf{X}, \phi) \\ &= \gamma_i(t) \frac{w_{il}\lambda_{il} \exp^{-\lambda_{il} d_l^2(t)}}{bd_i(t)} \end{aligned} \tag{5.13}$$

Given E segments of the phone, the HMM model parameters are updated using the formulation in Equation 5.14. Te is the length of each segment e.

$$\begin{aligned} w_{il} &= \frac{\sum_{e=1}^{E} \sum_{t=1}^{Te} \gamma_{il}^e(t)}{\sum_{e=1}^{E} \sum_{t=1}^{Te} \gamma_i^e(t)} \\ \lambda_{il} &= \frac{\sum_{e=1}^{E} \sum_{t=1}^{Te} \gamma_{il}^e(t)}{\sum_{e=1}^{E} \sum_{t=1}^{Te} d_l^{2e}(t) \gamma_i^e(t)} \end{aligned} \tag{5.14}$$

This formulation is integrated in the forward-backward code for HMM model estimation.

5.3.8 HMM Classification using Estimated Parameters

For scoring each segment we calculate the log-likelihood probability using the sum of αs of the forward-backward algorithm as in Equation 5.15. The sum of αs over all states at the end Te of segment e is the likelihood of the segment as shown in [38].

$$\begin{aligned} \hat{C} &= \text{argmax}_C P(C|\mathbf{X}) \\ P(C|\mathbf{X}) &= P(\mathbf{X}|C)P(C) \\ P(\mathbf{X}|C,\phi)) &= \sum_{i=1}^{N} \alpha_i(Te) \\ logP(C|\mathbf{X}) &= log(P(\mathbf{X}|C,\phi)) + log(P(C)) \end{aligned} \tag{5.15}$$

5.3.9 Alternative Training Approaches for the Exponential Distribution

We considered several different approaches for model training. These approaches vary in initialization and update strategies.

Estimating Lambdas and Weights Solely from Data

Using the framework described above, we use flat-initialization of the lambdas and weights from data as shown in Equations 5.9 and 5.3. The EM algorithm converges to the most likely solution given the distortion data for each phone separately. EM has the tendency to arrive at local, rather than global optima. Hence initialization is important.

Initializing Lambdas and Weights from EMA and Updating them from Data

As we mentioned in the introduction of this chapter, looking only at the distortion data for a given codeword regardless of the true codeword identity corrupts the distribution of

the distortion of that codeword. It is corrupted by distortions arising from data actually generated from other codewords. Using EMA we can tell at which frames each codeword occurs in the data and learn the weights of the codewords for a given phone as described in Equation 5.4. We can also use the distortion associated with these frames to learn the true lambdas of the exponential distribution as described in Equation 5.10. The EM will start from the solution provided by EMA and converge to the most likely solution given the distortion data for each phone.

Initializing Lambdas and Weights from EMA and Updating Weights Only from Data

We can also update just the weights from the distortion data associated with the phone and keep the lambdas fixed as estimated from EMA. This way the true distribution of codewords distortion is preserved.

Phone-Independent Initialization of Lambdas from EMA: Analogue to Semi-Continuous HMMs

Semi-continuous HMMs use a set of distributions that are shared among different phones. The differences in the models are only in the weights associated with the distribution. In our problem, we also try an analog to semi-continuous HMMs where the distribution of the distortion at the output of each codeword is estimated from EMA. Knowing the codeword's identity, and irrespective of the phone identity, we estimate the lambda for the distortions from that codeword. We fix the set of lambdas for each phone and only update the weights from the distortion data associated with the phone.

5.4 Generating a Realistic Articulatory Codebook and Articulatory Transfer Functions

In Chapter 4 we adapt Maeda's geometric model of the vocal tract to the EMA data of each speaker in the MOCHA database. We then search, on a frame-by-frame basis, a uniform codebook of Maeda parameters for vocal tract shapes that fit each frame of

EMA data. In this chapter, we sample each phone at five positions: the beginning, middle, end, between beginning and middle, and between middle and end, and read the corresponding Maeda parameter vectors found in the geometric search process. We also add the nasal tract opening area as an additional parameter to the Maeda vector to account for nasal sounds as described in Subsection 4.3.6. Table 5.1 shows the new codeword description.

Table 5.1: *Codeword made of the seven Maeda parameters derived from the uniform codebook and appending the velum opening area (VA).*

Codeword	p1	p2	p3	p4	p5	p6	p7	VA

We then perform k-means clustering over the set of parameter vectors obtained in this manner. We designate the vector closest to the mean of each cluster as codeword representing the cluster. This is done to guarantee that the codeword is a legitimate articulatory configuration. The set of codewords obtained in this manner is expected to span the space of realistic articulatory configurations which also accounts for information about the velum.

Once we compute the codebook, we convert it to articulatory transfer functions to be used to derive the analysis-by-synthesis distortion features described in Subsection 3.4.2. We have the option of using the adapted Maeda model to map the codewords to area functions and then to transfer functions or to use the unadapted model. Figure 5.6 shows the basic blocks for deriving the codebook of realistic vocal tract shapes.

5.4.1 Viewing the Phones in the Condensed Maeda Space

In order to visualize some of the results of our work, we use multi-dimensional scaling (MDS) [40] for high-dimensional data. MDS applies a technique similar to principal component analysis (PCA) to compute the eigenvectors of the data. Using the highest two eigenvectors we can project the data to a two-dimensional plane and visualize it.

We apply MDS on the mean vector of the Maeda parameters for each phone computed as described above. We have a vector of eight parameters for each phone. Projecting

the vectors into a two-dimensional plane helps us visualize the phones in a compressed Maeda space. Figure 5.7 shows this projection. It is interesting to note how the phones cluster together. For example, 'B', 'P', and 'M' cluster together and so do 'D', 'T', and 'N'. Each set of phones has a very similar vocal tract configuration. For example 'S' and 'Z' have the same constriction location and vocal tract shape and differ only in the voicing, as do 'F' and 'V'. The Maeda parameters for these two sets of phones are very similar. This supports the validity of our speaker-independent mapping from the EMA data to the Maeda parameters. The means of the Maeda parameters have been computed over all the speakers. These codewords are used as the basis of the dynamic framework described in this chapter.

5.5 Experimental Analysis and Results using the Analysis-by-Synthesis Frameworks

We conduct a number of experiments to evaluate the usefulness of the proposed articulatory framework for speech recognition. In order to avoid obfuscating our results with the effect of lexical and linguistic constraints that are inherent in a continuous speech recognition system, we evaluate our features on a simple phone classification task, where

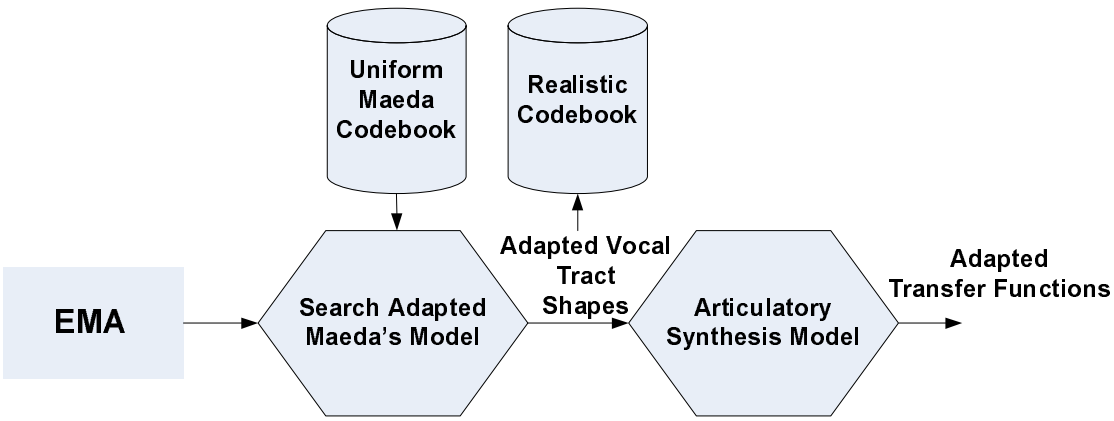


Figure 5.6: *Figure shows our approach of deriving a codebook of realistic vocal tract shapes from the uniform Maeda codebook.*

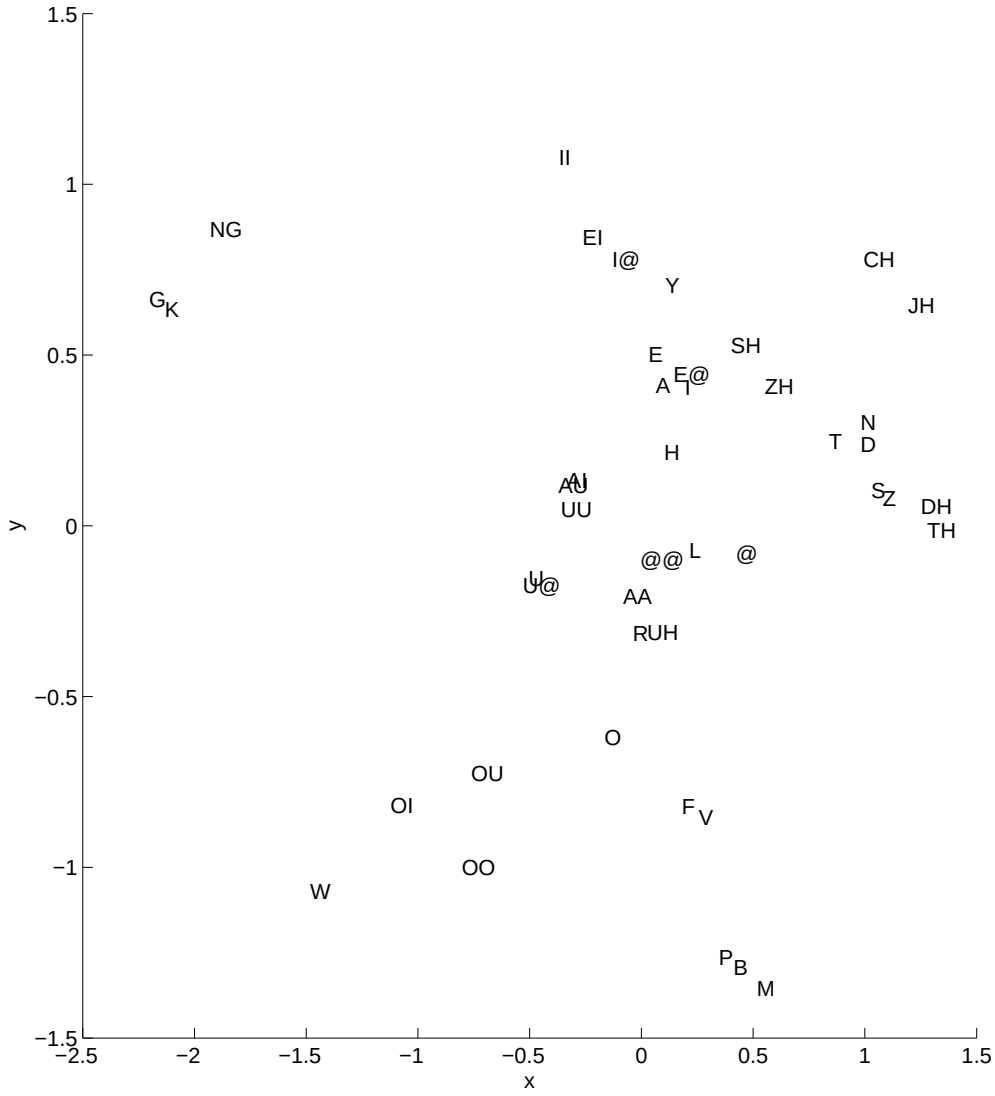


Figure 5.7: *Projecting the means of Maeda vectors of each phone into a compressed space by MDS. x and y are the first and second parameters of the MDS decomposition.*

the boundaries of phones are assumed to be known.

We choose as our data set the audio recordings from the MOCHA database itself, since it permits us to use the exact articulatory configurations for any segment of sound. We use the data from nine speakers for our work: "faet0", "falh0", "ffes0", "fjmw0", "fsew0", "maps0", "mjjn0", "msak0", and "ss2404". Five of the speakers are females and four are males. We choose to test on the female speaker "fsew0" and the male speaker "msak0" and train on the rest. All experiments are speaker independent. The number of utterances used for training is 2569 and for testing is 918. The test utterances are composed of 14352 phone segments from speaker "fsew0", 14302 segments from speaker "msak0", and 28654 segments in total. Only EMA data from the training speakers are used to compute the articulatory codebook and to initialize the model parameters. The codebook consists of 1024 codewords after clustering 394247 articulatory vectors sampled from all the phones. The articulatory data of the test speakers have not been used.

The phone $\hat{C}$ for each segment is estimated as:

$$\hat{C} = \text{argmax}_C P(C)P(MFCC|C)^{\alpha}P(FastDist|C)^{(1-\alpha)} \tag{5.16}$$

where C represents an arbitrary phone, and $MFCC$ and $FastDist$ represent the set of MFCC features and fast analysis-by-synthesis distortion features for the segment respectively. α is a positive number between 0 and 1 that specifies the relative contributions of the two features to classification. We vary the value of α between 0 and 1.0 in steps of 0.05, and choose the value that results in the best classification in the form of phone error rate (PER).

We have tried all of the described training approaches, feature normalization techniques, and probability density functions. We report below the best results achieved thus far. Most of the best results followed from using the exponential density function and without feature normalization.

5.5.1 EXP HMM 1: PER with Flat Initialization from Audio

The *Baseline* experiment reports phone error rates for the two speakers using 13-dimensional MFCC features with cepstral mean normalization (CMN). We use a three-state HMM with left-to-right topology with observation probabilities that are mixtures of 128 gaussian densities. We use vector quantization (VQ) to initialize the means of the mixtures.

In experiment *EXP HMM 1* we use the distortion features derived as discussed in Section 3.4.2 but using the realistic codebook described in Section 5.4. We use the articulatory synthesis model without adaptation to derive these features. We apply a three-state HMM and mixtures of 1024 exponential densities functions for the output probabilities. We initialize the weights and lambdas of the exponential distribution from the distortion features as described in Subsections 5.3.1 and 5.3.4. Using $\alpha = 0.2$ and combining the probability of the baseline system with this system yields a reduction of 5.3% in PER. This shows that our new framework does indeed improve the classification performance. The sparsity of the weights is defined as the percent of weights that are zeros for a given codeword over the three states, computed over all the codewords and phones. The codewords that have zero weights over the three HMM states do not need to be considered during classification, *i.e.* there is no need to synthesize speech from these codewords when considering a particular phone. Initializing from the distortion features (audio only) causes 21% of the weights to become zero. This is the "OR" approach we described before.

Table 5.2: *EXP HMM 1 PER using MFCC and a combination with the fast analysis-by-synthesis distortion features with parameters initialized from audio.*

Features (dimension)	Topology	Obser Prob	Sparsity	α	fsew0	msak0	Both
MFCC + CMN (13)	3S-128M-HMM	Gaussian	0%	1	61.6%	55.9%	58.8%
Fast Dist (1024)	3S-1024M-HMM	Exponential	21%	0.2	57.6%	53.7%	55.7%
Relative Improvement					6.5%	4.0%	5.3%

5.5.2 EXP HMM 2: PER with Initialization from EMA

In experiment *EXP HMM 2* we follow the same approach as *EXP HMM 1* except that now we initialize the weights and lambdas from the EMA data as described in Subsections 5.3.2 and 5.3.5. The EM algorithm starts from the solution provided by EMA and converges to the most likely solution given the distortion data for each phone. This increases the sparsity to 51% with small degradation in phone accuracy, which reduces the computation required considerably.

Table 5.3: *EXP HMM 2 PER using MFCC and a combination with the fast analysis-by-synthesis distortion features with parameters initialized from EMA.*

Features (dimension)	Topology	Obser Prob	Sparsity	α	fsew0	msak0	Both
MFCC + CMN (13)	3S-128M-HMM	Gaussian	0%	1	61.6%	55.9%	58.8%
Fast Dist (1024)	3S-1024M-HMM	Exponential	51%	0.2	58.3%	53.9%	56.1%
Relative Improvement					5.4%	3.9%	4.6%

5.5.3 EXP HMM 3: PER with Initialization from EMA using Adapted Transfer Functions

In experiment *EXP HMM 3* we follow the same approach as *EXP HMM 2* except that now we use the articulatory synthesis model with adaptation to derive the distortion features. Table 5.6 shows the effect of adaptation on phone classification. Note that especially for speaker "msak0", the adaptation has provided a small improvement in classification accuracy. The overall classification accuracy is the same as in *EXP HMM 1* but with the same sparsity as in *EXP HMM 2*. This shows that when the system focuses on a subset of articulatory configurations related to each phone and closely mimics the incoming speech through adaptation, it is most effective in classification. This is more evident in the "msak0" speaker case whose geometric adaptation is more effective on the synthesis quality than the adaptation of speaker "fsew0".

Table 5.4: *EXP HMM 3 PER using MFCC and a combination with the adapted fast analysis-by-synthesis distortion features with parameters initialized from EMA.*

Features (dimension)	Topology	Obser Prob	Sparsity	α	fsew0	msak0	Both
MFCC + CMN (13)	3S-128M-HMM	Gaussian	0%	1	61.6%	55.9%	58.8%
Adapted Fast Dist (1024)	3S-1024M-HMM	Exponential	51%	0.25	58.4%	53.1%	55.7%
Relative Improvement					5.2%	5.3%	5.3%

5.5.4 Viewing the Weights Estimated from Audio, from EMA, and from EMA with Adaptation

We apply MDS on the eight dimensional codewords derived in Section 5.4. We have computed 1024 codewords from clustering all the Maeda parameters of the phones. This way, we can visualize the location of each codeword on a two-dimensional grid. We use the pixel intensity to show the magnitude of the estimated weights of the codewords. We plot the weights estimated for phone 'OU' after the training approaches described above.

Figure 5.8 shows these weights derived in four ways. The top left figure shows the weights estimated from the EMA data as described in Subsection 5.3.2 and before EM updating from audio. This is the ground-truth prior distribution. The top right shows the weights initialized from audio as described in Subsection 5.3.1 and after EM update. These weights are used in *EXP HMM 1*. There is some correspondence in the two plots between the regions in the compressed Maeda space where these weights are active.

The bottom left plot shows the weights initialized from EMA and updated using EM. These weights are used in *EXP HMM 2*. There is more correspondence to the ground-truth weights than in the second plot. The bottom right plot shows the distribution of the weights initialized from EMA and after EM but using analysis-by-synthesis distortion features derived using an adapted vocal tract model for all the speakers. These weights are used in *EXP HMM 3*. It is clear that adaptation preserves the ground truth distribution of the weights and is a more accurate representation of the production process.

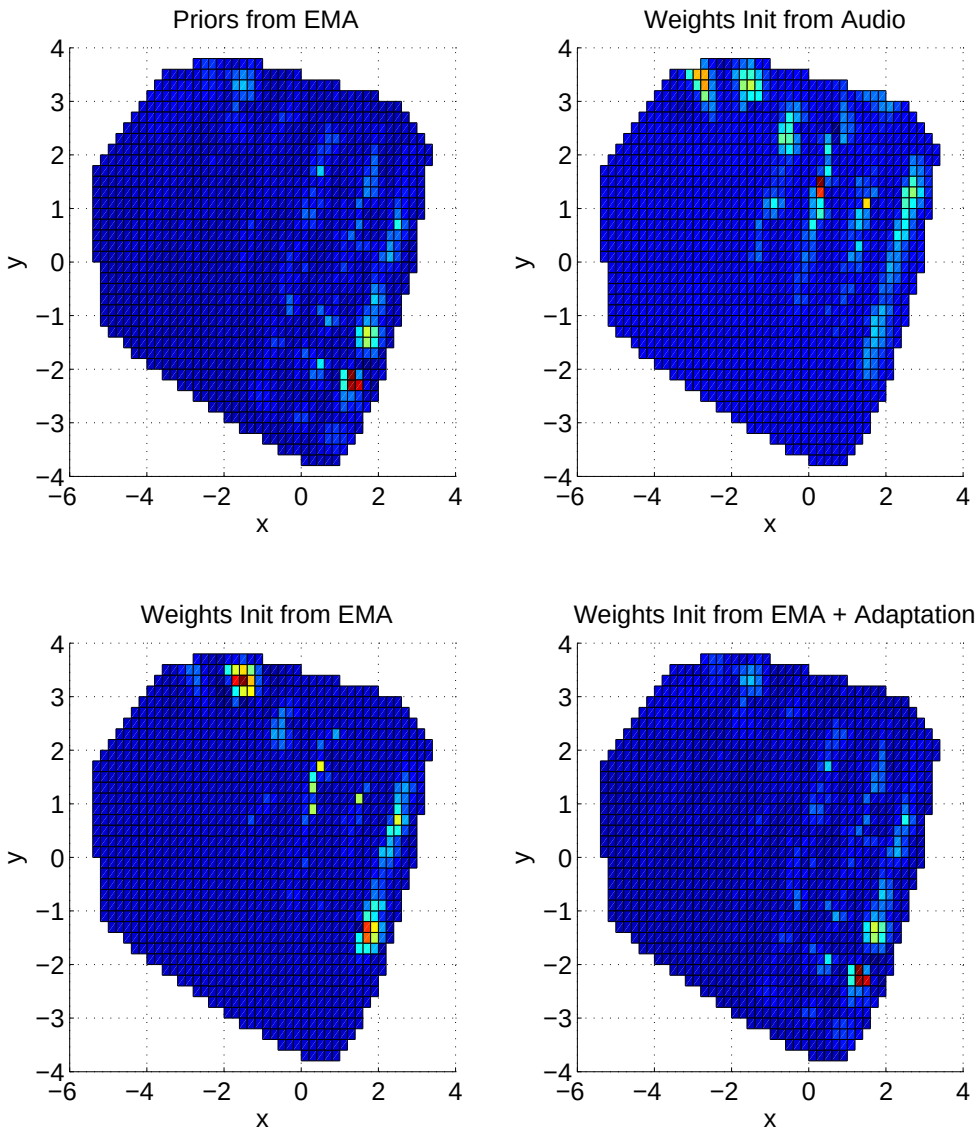


Figure 5.8: *Projection of codewords weights for phone 'OU' for the different experiments described in Section 5.5 and Table5.6. x and y are the first and second parameters of the MDS decomposition.*

5.5.5 EXP GAUS HMM: PER with LDA Compressed Features

Finally, in experiment *GAUS HMM* we use a similar setup to that in Chapter 3. Here we use all the distortion features in the "AND" approach. We apply LDA to compress the features to 20 dimensions and then apply CMN to them. We use a three-state HMM with 128 Gaussian Mixtures to model the new features. We combine the probabilities of this system with that of the baseline system. Using $\alpha = 0.6$ yields 10.9% reduction in phone error rate. This is the biggest improvement we achieved and shows that information in all the distortion features is helpful in discriminating among phones. The drawback of this topology is that there is no more an articulatory meaning to the states and hence we can not model the dynamics explicitly as we can in the previous experiments.

Table 5.5: *EXP GAUS HMM PER using MFCC and a combination with the LDA compressed fast analysis-by-synthesis distortion features.*

Features (dimension)	Topology	Obser Prob	Sparsity	α	fsew0	msak0	Both
MFCC + CMN (13)	3S-128M-HMM	Gaussian	0%	1	61.6%	55.9%	58.8%
Fast Dist + LDA + CMN (20)	3S-128M-HMM	Gaussian	0%	0.6	54.9%	49.8%	52.4%
Relative Improvement					10.8%	11.5%	10.9%

The summary of classification results and the optimal value of α are shown in Table 5.6 [41].

Table 5.6: *Phone error rates for the two speakers using different features, topologies, and initialization procedures.*

Experiment	Features (dimension)	Adaptation	Topology	Obser Prob	Initialization	Sparsity	α	fsew0	msak0	Both	Improvement
Baseline	MFCC + CMN (13)		3S-128M-HMM	Gaussian	VQ	0%	1	61.6%	55.9%	58.8%	
Exp HMM 1	Fast Dist (1024)	NO	3S-1024M-HMM	Exponential	Flat	21%	0.2	57.6%	53.7%	55.7%	5.3%
Exp HMM 2	Fast Dist (1024)	NO	3S-1024M-HMM	Exponential	EMA	51%	0.2	58.3%	53.9%	56.1%	4.6%
Exp HMM 3	Fast Dist (1024)	YES	3S-1024M-HMM	Exponential	EMA	51%	0.25	58.4%	53.1%	55.7%	5.3%
GAUS HMM	Fast Dist + LDA + CMN (20)	NO	3S-128M-HMM	Gaussian	VQ	0%	0.6	54.9%	49.8%	52.4%	10.9%

5.6 Conclusions and Future Work

We have described a dynamic articulatory model for phone classification that incorporates realistic vocal tract shapes in a statistical HMM framework. We have shown how to

incorporate analysis-by-synthesis distortion features in a probabilistic pattern recognition approach. Our new framework attributes articulatory meaning to the states through a set of weights. We have shown how to initialize these weights from ground-truth articulatory information and to update them from distortion data. Experimental results have demonstrated improvement in phone classification over the accuracy obtained using baseline MFCC features. We performed a speaker independent analysis of highly speaker-dependent phenomena. The framework we presented is a basic prototype for incorporating physical constraints in a statistical framework, and it can be expanded in the future to incorporate further dynamic constraints. Future work will integrate the trained models into a continuous speech recognition system.

Chapter 6

Suggestions for Future Work

Articulatory modeling is at the core of the speech production mechanism. Many problems in speech research can be approached from this point of view for additional insights. In our future research we would like to continue investigating articulatory phenomena and applying the knowledge we gain to various speech problems. In addition to speech recognition, our approach can be used for speech coding and speech synthesis, for speaker, age, and gender identification problems, and for pronunciation modeling, assessment, and tutoring. The synthesis approach we developed can help in speech pathology research. The articulatory approach we presented is language independent and can be easily used with other languages.

Below we outline methods for expanding our framework into new domains and into fully continuous speech recognition.

6.1 Other Domains Where Our Approach Maybe Helpful

We have evaluated our novel approach for speech recognition on the MOCHA database which is composed of data from nine British English speakers reading TIMIT utterances. We have obtained improvements for segmented phone recognition experiments.

6.1.1 Spontaneous Speech

The primary motivation for our work is spontaneous and conversational speech recognition, which remains a major challenge for current state-of-the-art systems. Our framework models the articulatory mechanism and is expected to provide even greater improvements in the recognition of spontaneous speech than recognizing read speech. Future experiments are needed to verify this claim.

6.1.2 Noisy Environments

The MOCHA database contains clean recordings. We expect to achieve similar improvements in noisy environments, especially when the nature of the noise is different from that of human speech. Since our model mimics the speech production system, we expect that it will be robust to noise. Future experiments are needed to verify this claim.

6.2 Articulatory Features

6.2.1 Further Exploration of the Features

We will continue the work described in chapters 3 and 5 by further exploration of the analysis-by-synthesis features. We will experiment with different mathematical models for the vocal tract. Besides Maeda's model, other geometric models exist in the literature including those of Mermelstein [42] and Coker [43]. These models use different sets of parameters and control over these parameters to account for different sounds. We will seek synthesis models other than the Sondhi and Schroeter model such as Steven's [7] electric transmission line analog of the vocal tract.

We will explore different methods to represent the distortion distance between synthesized and incoming speech. The distance can be thought of as a negative log-likelihood. The larger the distortion (dist) the less likely is the articulatory configuration (AC) used to generate the incoming observation (x), as is shown in Equation 6.1, with the constants α and β to be determined. For the distortion measure, in addition to Mel-cepstral distortion as defined in Equation 3.1, we can use the Itakura-Saito distance.

$$P(x|AC) \propto \beta \cdot \exp^{-\alpha \cdot dist(\mathbf{c}_{incoming}, \mathbf{c}_{synth})} \tag{6.1}$$

We will explore other dimensionality-reduction techniques besides LDA. Principal component analysis (PCA) is a candidate and a neural network can also be used to map the long-dimensional feature vector to a smaller one whose dimensions represent posterior probabilities of phone identities, as is done in Hermansky's tandem approach [44].

We will also experiment with different codebook sizes and compare with using other codebooks and mapping functions from codewords to acoustics, such as those used in speech coding.

6.2.2 Feature Combination

There are different ways to combine the acoustic features and the analysis-by-synthesis distortion features. We can simply concatenate the two feature streams or use the hypothesis combination approach of Singh [45].

6.2.3 State Probability Combination

Li in [46] also describes a state combination approach to combine the probabilities from the states of two systems, each trained using one of the two feature sets. In our framework, we can combine the probability scores from MFCC features with the scores from the analysis-by-synthesis distortion features at the output of each state as described in Equation 6.2. For state $S(t) = i$, the observation density $b_i(t)$ is a product of $bc_i(t)$ which based on the MFCC features $\{c0(t), c1(t), \ldots, c12(t)\}$ and $bd_i(t)$ which is based on the distortion features $\{d0(t), d1(t), \ldots, dM(t)\}$. A gaussian mixture density is used for $bc_i(t)$ and an exponential mixture density is used from $bd_i(t)$. Figure 6.1 describes this hybrid HMM framework.

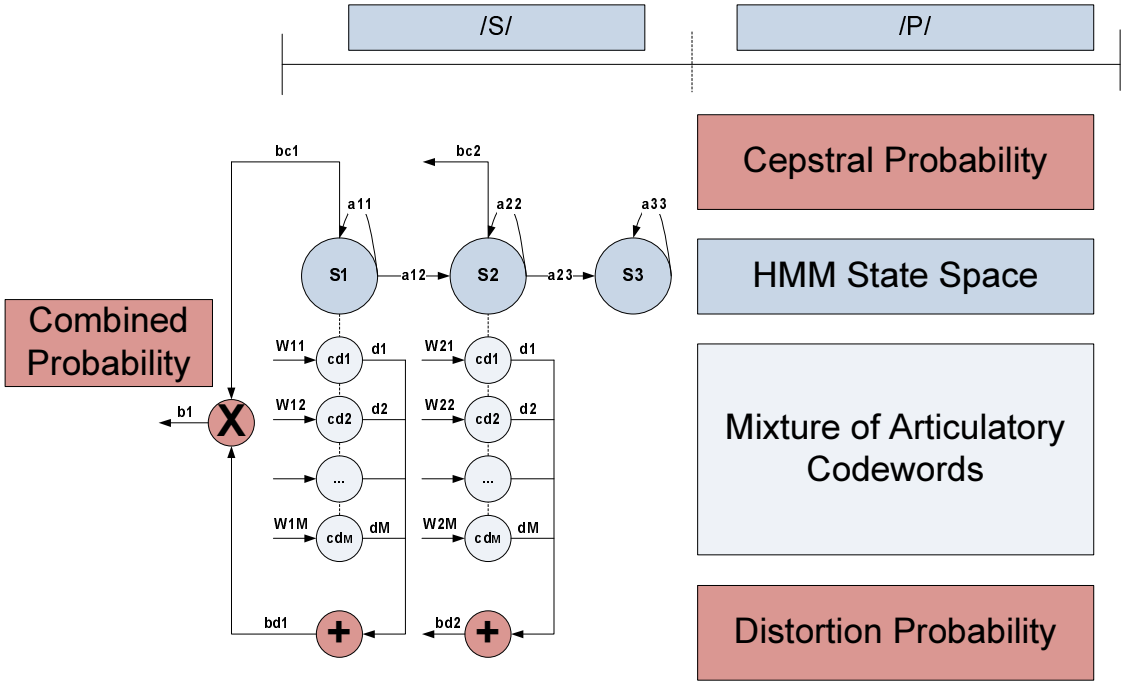


Figure 6.1: *Hybrid HMM with two streams of features: MFCCs modeled by traditional mixture of Gaussian densities and distortion features modeled by a mixture density function where different probability functions would be tried.*

$$
\begin{aligned}
b_i(t) &= P(x(t)|S(t)=i) \\
&= P(c0(t), c1(t), \ldots, c12(t), d0(t), d1(t), \ldots, dM(t)|S(t)=i) \\
&\approx P(\mathbf{c}(t)|S(t)=i) * P(\mathbf{d}(t)|S(t)=i) \\
&\approx bc_i(t) * bd_i(t) \\
&\approx \sum_{k=1}^{Mc} wc_{ik} N(\mathbf{c}(t); \mu_{ik}, \mathbf{\Sigma}_{ik}) * \sum_{l=1}^{M} w_{il}\lambda_{il} \exp^{-\lambda_{il} d_l^2(t)}
\end{aligned} \tag{6.2}
$$

Since the two streams of features and the probability densities used to model them are different, we need to weight the probability streams separately. We would use a parameter α to exponentially weight the two probability streams and to compute a combined output observation probability. This formulation would be applied during model training and during model testing.

$$
b_i(t) = bc_i(t)^{\alpha} * bd_i(t)^{(1-\alpha)} \tag{6.3}
$$

6.3 Smoothing the Estimates to Ensure Sparsity

6.3.1 Estimation with Deleted Interpolation

To obtain a smoother estimate of the weights for each *phone C*, we would follow the approach of deleted interpolation which can be summarized as follows. The data would be split into L parts and a set of weights $\{\alpha_{lj}, (j = 1 : M)\}$ would be estimated for each part l separately. We would have L different models for each phone. Each part l would then be held out and used to get an estimate of the weights $\{\rho_{ml}, (m = 1 : L, m \neq l)\}$, which would be the interpolation weights of the pre-trained $L - 1$ models.

Once the interpolation weights that maximize the likelihood of part l are found, the mixture weights would be averaged from the $L - 1$ models to get $\{\beta_{lj}, (j = 1 : M)\}$. The mixture weights computed for each of the held out parts would then be averaged to get the final set of mixture weights $\{\alpha_j, (j = 1 : M)\}$. Below is the formulation we would use.

- Step 1:

  Divide the data into L parts, each part consisting of randomly selected N_L points.

- Step 2:

  For each part l $\{l = 1 : L\}$, estimate the mixture weights for this part $\{\alpha_{lj}, (j = 1 : M)\}$:

$$\alpha_{lj}^{t} = \frac{\alpha_{lj}^{t-1}}{N_L} \sum_{i=1}^{N_L} \frac{P(x_i^l|s_j)}{\sum_{k=1}^{M} \alpha_{lk}^{t-1} P(x_i^l|s_k)} \tag{6.4}$$

- Step 3:

  For each part l designated as the deleted part, $\{m, m = 1 : L, m \neq l\}$ would be the remaining set of parts:

$$\begin{aligned} P(x_i^l|C) &= \sum_{m=1, m\neq l}^{L} \rho_{ml} P_m(x_i^l|C) \\ P_m(x_i^l|C) &= \sum_{j=1}^{M} \alpha_{mj} P(x_i^j|s_j) \end{aligned} \tag{6.5}$$

  Equation 6.5 is similar to Equation 5.1 since the α_{mj} and hence $P_m(x_i^l|C)$ are fixed and we want to find the maximum likelihood estimate of ρ_{ml}, the $L-1$ mixture weights in this case. The solution for $\rho_{ml}, \{m, m = 1 : L, m \neq l\}$, would be:

$$\rho_{ml}^{t} = \frac{\rho_{ml}^{t-1}}{N_L} \sum_{i=1}^{N_L} \frac{P_m(x_i^l|C)}{\sum_{k=1, k\neq l}^{L} \rho_{kl}^{t-1} P_k(x_i^l|C)} \tag{6.6}$$

  Once the iterations for interpolation weights ρ_{ml} converge, the average mixture weights $\{\beta_{lj}, (j = 1 : M)\}$ for part l would be computed as follows:

$$\beta_{lj} = \sum_{m=1, m\neq l}^{L} \rho_{ml} \alpha_{mj} \tag{6.7}$$

- Step 4:

  Finally, we would average the $\{\beta_{lj}\}$ for all parts:

$$\alpha_j = \frac{1}{L}\sum_{l=1}^{L}\beta_{lj} \tag{6.8}$$

6.3.2 Entropy Minimization to Ensure Sparseness in Transition Matrices

In addition to smoothing the estimates using deleted interpolation, we can apply the minimum entropy approach of Brand [47].

6.4 Dynamic Framework

6.4.1 Extending the Framework for Word Recognition

In our work so far, we have used a simple framework to test the articulatory features. We used an HMM framework for a segmented context independent (CI) phone recognition experiment. The CI phone recognition experiments will be extended to word recognition experiments by incorporating a dictionary and word transition probabilities in the search.

6.4.2 Rescoring the N-best Hypotheses

Using forced-alignment of the acoustic data to the transcript, we determine the state identity for each frame of speech. Using the EMA data recorded in parallel to speech, we determine the codeword identity for each frame as well. Following this procedure, each state will have a collection of codewords assigned to it. The upper HMM states in Figure 6.2 define acoustic distributions and may correspond to different articulatory configurations (codewords). Recall that similar acoustic observations may come from different articulatory configurations. We keep the codewords whose count for a given state is more than a certain threshold. These are the most common vocal tract shapes that occur for a given acoustic observation modeled by a state. Then we use this information in a rescoring experiment. Forced-alignment of the N-best hypotheses provides

the acoustic state sequence for each hypothesis. We compute a new score (error) for each hypothesis that is composed of two measures, an articulatory and a distortion measure. The articulatory measure is a function of the transitions from one codeword to another. The distortion measure is the total distortion between the speech synthesized from the codewords and the incoming speech. Using dynamic programming, we choose the path that leads to the minimum overall error measure. We then choose the hypothesis with the minimum error as our best hypothesis. The estimation of α would be optimized using the MOCHA database development set.

$$error = \alpha * acoustic\_distortion + (1 - \alpha) * articulatory\_distance \tag{6.9}$$

6.4.3 Articulatory Distance

From the EMA we not only find which codewords correspond to which acoustic state, but also the transition probability of the codewords and the prior probability of each codeword.

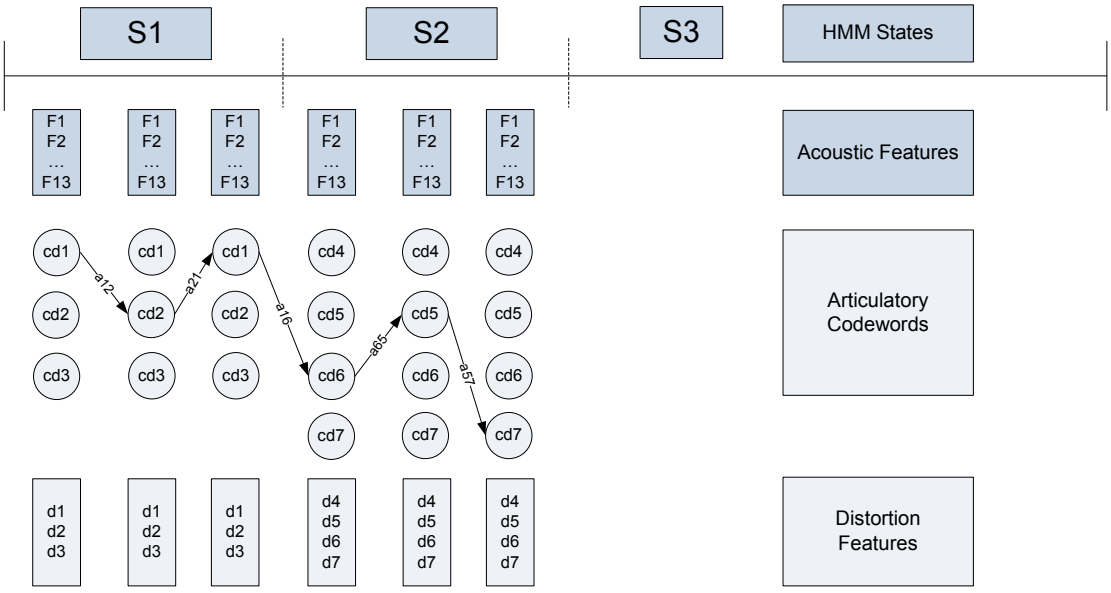


Figure 6.2: *Dynamic framework where each HMM state defines a trajectory in the articulatory space.*

$$P(cd_2|cd_1, S_1) = \frac{count\_transitions(cd_2|cd_1, S_1)}{total\_transitions(S_1)} \tag{6.10}$$

We can also model the articulatory distance as a function of a weighted difference of the seven Maeda parameters corresponding to the codewords. The weights will penalize big changes in the Maeda parameters and ensure smooth dynamic transitions. They can be estimated from the Maeda parameters pertaining to each acoustic state. Since each acoustic state is a collection of codewords, we can estimate the information gain, the variance, and the autocorrelation of each parameter in the codewords. We will run information theoretic analysis on the parameters derived from MOCHA to learn suitable weights. The relative information gain finds the contribution of each parameter in classifying each phone. The auto-correlation function of each parameter will give a measure of the rate of change with time. The variance will determine the range of the parameters at each time instant for a given phone. Parameters with low variance are associated with the primary articulators. Hence changes in these parameters would be penalized more than changes in parameters with high variance which are not important for articulation and correspond to secondary articulators. Equation 6.11 defines the articulatory distance between codewords cd_1 and cd_2 as the squared difference of the articulatory parameters of these codewords $\{p_i, i = 1:7\}$. The differences are weighted by $\{\omega_i, i = 1:7\}$.

$$articulatory\_distance(cd_1, cd_2) = \sum_{i=1}^{7} \omega_i (p_i^2 - p_i^1)^2 \tag{6.11}$$

6.4.4 Incorporating Dynamic Constraints into the Articulatory Space

So far, our approach in deriving the articulatory features can be described as being "instantaneous". For each incoming frame of speech, distortion is computed with respect to speech synthesized from all the codewords without taking the distortion of the previous frame into consideration, in a way performing an exhaustive search for all the possible configurations. Beside being computationally expensive, this approach does not

take advantage of the physical constraints on the dynamics of the articulators. Smooth articulatory dynamics are more likely than abrupt changes. Incorporating dynamic information as a constraint will reduce the computations and improve the performance by only evaluating these articulatory configurations that are possible at each stage. This limits the search space and reduces the noise that could be added by the physically unlikely transitions. In this part, we describe a technique for incorporating "dynamic" constraints on top of the already existing "instantaneous" constraints that allowed only a limited set of configurations to span the articulatory space. Dynamic modeling of the articulatory trajectories is also aimed at tackling some of the phenomena that occur during speech production like coarticulation, sloppiness in spontaneous speech, faster speech rates, etc.

In Figure 6.3, the states represent articulatory configurations, unlike conventional HMM systems where they represent abstract units as mentioned before. Transition from one state to another can be assigned a physical measure proportional to the "effort" exerted in moving the articulators from one configuration to another as shown in Equation 6.12. We will develop a flexible framework that allows for incorporating different models for effort.

$$P(S2|S1) \propto \beta \cdot \exp^{-\alpha \cdot effort(AC1,AC2)} \tag{6.12}$$

As shown in Figure 6.3, all the states are shared by all the phones. Each phone C will have its own transition matrix A_C and initial probability vector π_C that measure the likelihood of transition from one articulatory configuration to another within each phone. The vector of articulatory parameters represented by the state is mapped to acoustic parameters using the mathematical model of the vocal tract and the excitation signal of the original speech. So the output of the same state will vary with time. We will find the best path in this articulatory space that leads to the least distortion between the incoming speech and the speech synthesized from the states visited and also to the least articulatory effort, as we will describe next.

We will start with a simple dynamic programming (DP) based approach and incor-

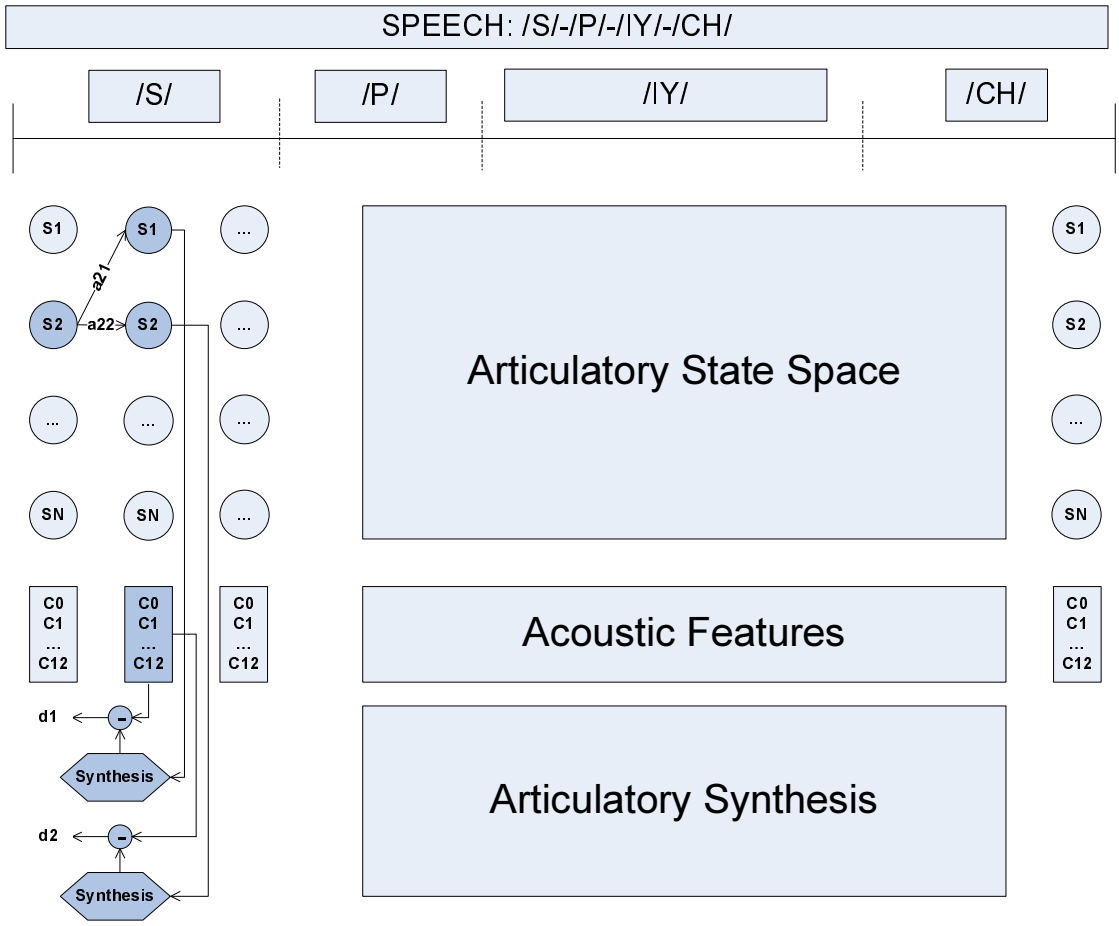


Figure 6.3: *Dynamic framework in which transition probabilities are denoted by a21, a22, etc. and distortion features by d1, d2, etc. The darker color corresponds to states and features being considered in the search process.*

porate dynamic constraints. Two penalties will guide the search process, one based on the articulatory effort in moving from the previous to the current articulatory configuration and an acoustic one based on the distance between incoming speech and speech synthesized from the current articulatory configuration. Dynamic constraints on the articulators are expected to account for the "many-to-one" problem where many articulatory configurations could generate the same acoustic observation. Paths with "minimum effort" and best acoustic matching between the synthesized and original signal are preferred. A stack decoder [48] or comparable frame-synchronous search can keep more than one trajectory hypothesis active until the future context helps resolve the ambiguity.

During training we will find these paths using prior estimates of the transition probabilities. We will present three ways of estimating theses priors: from MOCHA, using a weighted geometric distance, and using a mechanical analogue of the vocal tract. To find these paths, we will compare incoming speech to a signal synthesized by applying the vocal tract parameter values represented by the states to the mathematical model of the vocal tract to derive a "synthesis error". This error will be a measure of the "fit" of each state to the incoming data and will be used together with the prior estimate of the transition probability to find the paths of least cost.

Once the path is determined, the transition probability matrices will be re-estimated. We will enforce sparsity on the transition matrices using entropy minimization approaches [47]. This will help reduce computations since it will prune out the states that are not likely to be visited.

Finally, we will extend the dynamic framework for connected CI phone recognition and connected word recognition experiments as mentioned at the beginning of this section. We will then combine it with the conventional framework based on MFCC features following a state combination approach.

6.4.5 Finite Element Model Accounting for Physiological Constraints

In this approach, we will devise a mechanical analogue of the vocal tract. We will model the articulators using springs, masses, dampers, and other mechanical compo-

nents. Hence we propose a finite element model of the vocal tract which will provide the measure of the energy needed to move the articulators along the x and y axis. Each articulator will have a velocity and acceleration from which we will predict its movement and the movement of the dependent articulators. Equation 6.13 describes mathematically this model. The mass of each articulator is modeled by m, the damping coefficient by b, and the stiffness coefficient by k. Varying the parameters of the model accounts for variations in the vocal tract configurations (*i.e.* phone identities) and in speaker and speaking style. This model would provide an estimate of the effort described in Equation 6.12. We will use real measurements of articulatory trajectories and biologically inspired limitations to learn these parameters. This is presently an immature idea thats needs further analysis and discussion. A task dynamic articulatory model has been proposed in the literature by Saltzman [49].

$$\begin{aligned} effort_x &= \sum_x F \\ &= m\frac{d^2x}{dt^2} + b\frac{dx}{dt} + kx \\ effort_y &= \sum_y F \\ &= m\frac{d^2y}{dt^2} + b\frac{dy}{dt} + ky \end{aligned} \tag{6.13}$$

6.4.6 Factored-State Representation

We will develop a latent-variable model with each variable representing one articulatory parameter. Such a model will account for asynchrony in the articulator movements and pronunciation variations, in addition to the other high-level speech phenomena. It also allows us to model the critical articulators explicitly.

As before, statistical dependencies between the latent variables that capture dynamic relationships will be separately learned for each parameter through a maximum likelihood (or alternative) approach as in the previous section and incorporating the constraints we mentioned.

Two frameworks can be used here, a factorial HMM and a switched Kalman filter. For

the factorial HMM, the states will be quantized configurations, each articulator modeled by an independent state stream. In order to allow for dependencies between the states, the factorial HMM will be expanded into a Dynamic Bayesian Network (DBN). As for the Kalman filter, the states are continuous valued and they can model the articulatory parameters if trained from real data.

$$\begin{aligned} x_k &= F_k \cdot x_{k-1} + w_{k-1} \\ z_k &= H_k \cdot x_k + v_k \end{aligned} \tag{6.14}$$

The F_k matrix in Equation 6.14 will learn the dynamic propagation constraints while the H_k matrix models the mapping function from articulatory states to acoustic observations, which is the articulatory synthesis model. x_k is the vector of articulatory states, z_k is the acoustic observation, and w_{k-1} and v_k account for the noise. In testing, the F_k matrix can be fixed and will track the articulatory parameters imposing the physical constraints learned in training.

Finally, we would like to extend this approach to other databases where articulatory information is not available by incorporating in the learning process a Bayesian prior computed from real data like MOCHA. The EM algorithm will be used to estimate the parameters of the unobserved articulatory trajectories. Speech *mimicry* will be the objective function that will guide the parameter estimation procedure.

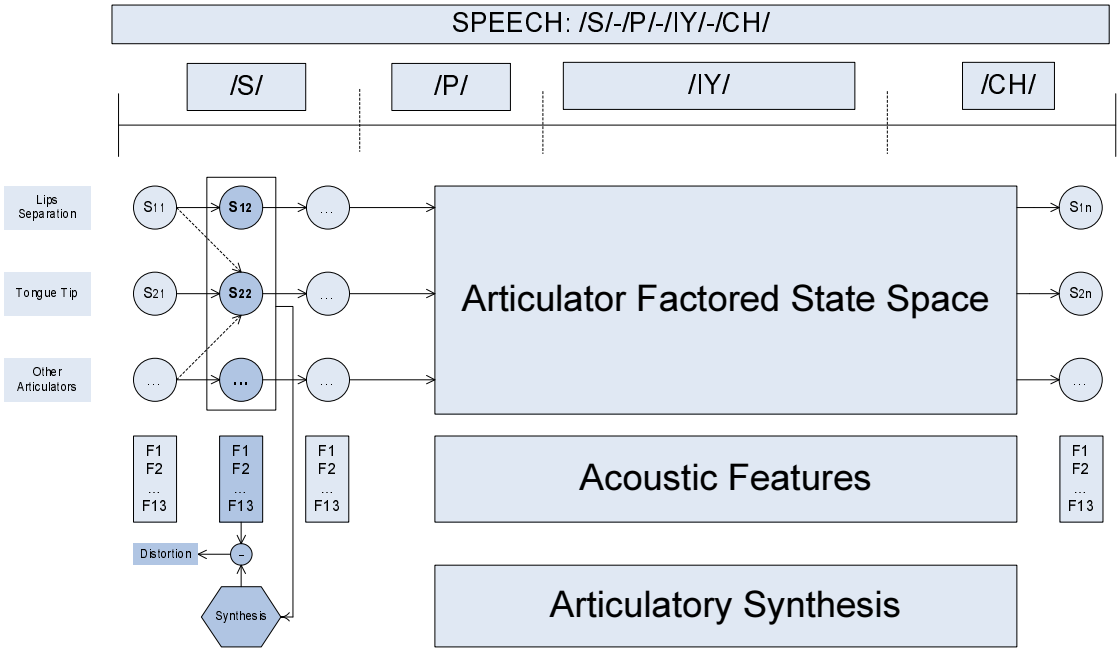


Figure 6.4: *Factored state-space framework where articulators propagate independently. Some dependencies could be included as shown by the dotted arrows.*

Chapter 7

Conclusions and Contributions

7.1 Summary of Major Results

We presented a novel approach for speech recognition that incorporates knowledge of the speech production process. We discussed our contributions in going from a purely statistical speech recognizer to one that is motivated by the physical generative process of speech. We followed an analysis-by-synthesis approach. We conclude that a model that mimics the actual physics of the vocal tract results in better classification accuracy. This work would not be possible without the recent availability of databases that open new horizons to better understand the articulatory phenomena. In addition, current advancements in computation and machine learning algorithms facilitate the integration of physical models in large scale systems.

7.1.1 Analysis-by-synthesis features

In Chapter 3 we explained how we derived the analysis-by-synthesis distortion features for speech recognition. Improvements on a segmented phoneme classification experiment using nine speakers in the MOCHA database were achieved using these features. This shows that the proposed features do indeed convey extra information about the articulatory process that are not present in the acoustic features alone.

7.1.2 Deriving Realistic Vocal Tract Shapes from EMA Measurements

In Chapter 4 we showed how we derive realistic vocal tract shapes from crude measurements that do not sufficiently describe the overall contour of the vocal tract. We presented a method for adapting Maeda's model to the EMA data in the MOCHA database. This approach is not limited to Maeda's model or the MOCHA database. The articulatory synthesis approach of Sondhi and Schroeter was then applied to synthesize speech from these vocal tract shapes. Our research has thus presented a technique for synthesizing speech solely from EMA data without any statistical mapping from EMA to acoustic parameters. This would not have been possible without our knowledge of the physics of the speech generation process, which was heavily exploited by the models used in this work. Reductions in Mel-cepstral distortion between the real speech and the synthesized speech confirmed the effectiveness of the adaptation procedure followed.

7.1.3 Dynamic Framework Using the Analysis-by-Synthesis Distortion Features

In Chapter 5 we described a dynamic articulatory model for phone classification that incorporates realistic vocal tract shapes in a statistical HMM framework. We showed how to incorporate analysis-by-synthesis distortion features in a probabilistic pattern recognition approach. Our new framework attributed articulatory meaning to the states through a set of weights. We showed how to initialize these weights from ground-truth articulatory information and how to update them from distortion data. Experimental results demonstrated improvements in phone classification over baseline MFCC features.

7.2 Summary of Contributions

We described the basic steps needed to incorporate knowledge of a physical phenomenon into a statistical patten recognition system to improve its performance. The main contributions of this thesis can be divided into three parts. First, a physical meaning is attributed to the inner states of the recognition system pertaining to the articulatory

configurations the human vocal tract takes over time. Second, the mapping from the states to the observations is based on a biologically-inspired model of speech production. Third, the distortion between the speech synthesized from the vocal tract configurations and the incoming speech is used in an analysis-by-synthesis framework to measure the likelihood of each state (the vocal tract shape generating the sound).

7.2.1 A Knowledge-Based Approach to the Speech Recognition Problem

State-of-the-art speech recognition systems use Hidden Markov Models (HMMs) which are composed of states and observations. Each word is represented by a string of phones which in turn are modeled using a concatenation of non-overlapping states. During search, the HMM is used to find the best sequence of states that is most likely to have generated the given acoustic signal. While the HMM is a hypothetical generative model, the vocal tract is the actual generative model. In this thesis, we devised a technique for incorporating a mathematical model of the physics of the vocal tract into speech recognition. Table 7.1 summarizes the main differences between our production-based HMM and the conventional HMM framework.

Table 7.1: *Production-based HMM versus conventional HMM.*

	Production-Based HMM	Conventional HMM
States	Real articulatory configurations	Abstract, no physical meaning
Output Observ Prob	Exponential prob using fast distortion feat	Gauss prob using MFCC
Adaptation	Vocal tract explicit geometric adaptation	VTLN, MLLR, MAP
Transition Probability	Learned from articulatory dynamics	Based on acoustic observation

Defining the Inner Components of the Model by a Set of Meaningful Units

We defined the inner states of the HMM by a set of codewords. The codewords were derived using Maeda's geometric vocal tract model and ElectroMagnetic Articulography data (EMA) in the MOCHA database. The EMA measurements were obtained from seven speakers and represented all the different phonemes of British English. Each

codeword represented a distinct vocal tract shape. This setup constrains the search process to these vocal tract shapes only.

Utilizing a Realistic Mapping from the Units Space to the Observation Space

To go from the abstract units space to the observation space, we utilized a mapping based on the physical generative process. Such mappings are essential in order to generalize among different speakers. We used Sondhi and Schroeter's articulatory synthesis model that mimics the physics of the sound generation mechanism to map from the vocal tract space to the acoustic space. We derived analysis-by-synthesis distortion features between the incoming speech and the speech synthesized from the vocal tract shapes.

Accounting for Speaker Differences by Adaptation

We adapted Maeda's model to the geometric distribution of the EMA measurements of each speaker separately. This accomplished a more accurate mapping from the articulatory space to the acoustic one. This accomplished in turn a better estimation of the likelihoods of the articulatory configurations given the acoustic signals, which led to improved phone classification.

Initializing the Model's Free Parameters from Ground-Truth Information Used as a Prior Distribution

Since the set of codewords that defined the states of the model had articulatory meaning, we were able to use the articulatory data to initialize the model free parameters (*e.g.* the mixture weights and the lambdas of the exponential distribution) associated with the states. Hence we used ground-truth articulatory knowledge to learn these parameters and use them as a prior distribution. The estimates of these parameters were updated in a maximum-likelihood sense from the observation data (*i.e.* the distortion between the synthesized speech and the incoming speech). The advantage of this prior was that the estimation algorithm started from a solution obtained from measurements of the real phenomenon. It also converged to a similar solution. Hence, this is an additional way

to constrain the system parameters to the physical phenomena of speech production.

7.2.2 Novel Aspects of Our Work

To the best of our knowledge, we are the first to apply the analysis-by-synthesis paradigm in a statistical fashion to phone classification. We are the first to integrate realistic and speaker-adapted vocal tract shapes in a dynamic framework and to incorporate a physiologically-motivated articulatory synthesis model in a pattern recognition framework. Our work is the first to synthesize continuous speech waveforms solely from EMA and to perform a speaker-independent analysis of a highly speaker-dependent phenomena.

Bibliography

[1] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *J. Acoust. Soc. Am.*, vol. 118(1), pp. 444–460, July 2005.

[2] A. Wrench, "A new resource for production modeling in speech technology," in *Workshop on Innovations in speech processing*, Stratford-upon-Avon, UK, 2001.

[3] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences," *IEEE Transac. ASSP*, vol. ASSP-28, No. 4, pp. 357–366, August 1980.

[4] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *ASRU*, Keystone, CO., December 1999.

[5] E. Fosler-Lussier, "Underspecified feature models for pronunciation variation in asr," in *ISCA Tutorial and Research Workshop on Speech Recognition and Intrinsic Variation*, Toulouse, France, 2006.

[6] J. L. Flanagan, *Speech analysis, synthesis, and perception*, Berlin: Springer-Verlag, 2nd edition, 1972.

[7] K. N. Stevens, *Acoustic phonetics*, MIT Press, 1998.

[8] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Transac. ASSP*, vol. 35, pp. 955–967, July 1987.

[9] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in *Speech Production and Modellling*. W.J. Hardcastle and A. Marchal (eds.), 1990, pp. 131–149, Kluwer.

[10] P. Hawkins, *Introducing Phonology*, Hutchinson, London, 1984.

[11] A. Liberman and I. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, pp. 1–36, 1985.

[12] K. Erler and G. H. Freeman, "An hmm-based speech recognizer using overlapping articulatory features," *J. Acoust. Soc. Am.*, vol. 100(4), pp. 2500–2513, 1996.

[13] C. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[14] V. Singampalli and P. Jackson, "Statistical identification of critical, dependent and redundant articulators," in *Interspeech*, Antwerp, Belgium, 2007.

[15] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy., "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *J. Acoust. Soc. Am.*, vol. 92(2), pp. 688–700, 1992.

[16] K. Livescu, *Feature-Based Pronunciation Modeling for Automatic Speech Recognition*, Ph.D. thesis, MIT Department of Electrical Engineering and Computer Science, 2005.

[17] C. S. Blackburn, *Articulatory Methods for Speech Production and Recognition*, Ph.D. thesis, University of Cambridge, 1996.

[18] L. Deng, *DYNAMIC SPEECH MODELS — Theory, Algorithms, and Applications*, Morgan and Claypool Publishers, May 2006.

[19] T. Stephenson, H. Bourlard, S. Bengio, and A. Morris, "Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory vari-

ables," in *International Conference on Spoken Language Processing*, Beijing, China, 2000.

[20] K. Markov, J. Dang, and S. Nakamura, "Integration of articulatory and spectrum features based on the hybrid hmm/bn modeling framework," *Speech Communications*, vol. 48(2), October 2006.

[21] J. Frankel, *Linear dynamic models for automatic speech recognition*, Ph.D. thesis, University of Edinburgh, Edinburgh, UK., 2003.

[22] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 121, pp. 723–742, February 2007.

[23] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J .*, vol. 51, no. 6, pp. 1233–1268, 1972.

[24] E. L. Riegelsberger, *The Acoustic-to-Articulatory Mapping of Voiced and Fricated Speech*, Ph.D. thesis, The Ohio State University, 1997.

[25] B. S. Atal, "Effectiveness of linear prediction charactersitics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55(6), pp. 1304–1312, June 1974.

[26] T. Toda, A. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA. USA, 2004, pp. 31–36.

[27] A. Toth and A. Black, "Using articulatory position data in voice transformation," in *ISCA SSW6*, Bonn, Germany, 2007.

[28] Z. Al Bawab, B. Raj, and R.M. Stern, "Analysis-by-synthesis features for speech recognition," in *ICASSP*, Las Vegas, Nevada. USA, April 2008.

[29] R. McGowan and S. Cushing, "Vocal tract normalization for midsagittal articulatory recovery with analysis-by-synthesis," *J. Acoust. Soc. Am.*, vol. 106, Issue 2, pp. 1090–1105, August 1999.

[30] B. Mathieu and Y. Laprie, "Adaptation of maeda's model for acoustic to articulatory inversion," in *Eurospeech*, Rhodes, Greece, 1997, pp. 2015–2018.

[31] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am*, vol. 49 2, pp. 583590, 1971.

[32] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator markov models for speech recognition," *Speech Communications*, vol. 41(2), October 2003.

[33] C. S. Blackburn and S. J. Young, "Pseudo-articulatory speech synthesis for recognition using automatic feature extraction from X-ray data," in *Proc. ICSLP '96*, Philadelphia, PA, 1996, vol. 2, pp. 969–972.

[34] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *IEEE Transactions on Speech and Audio Processing*, vol. 12(2), pp. 175–185, 2004.

[35] V. Mitra, H. Nam, C. Espy-Wilson, and E. Saltzman L. Goldstein, "Noise robustness of tract variables and their application to speech recognition," in *Interspeech*, Brighton, UK, September 2009.

[36] Z. Al Bawab, L. Turicchia, R. M. Stern, and B. Raj, "Deriving vocal tract shapes from electromagnetic articulograph data via geometric adaptation and matching," in *Interspeech*, Brighton, UK, September 2009.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39(1), pp. 1–38, 1977.

[38] J. Bilmes, "A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," in *Technical Report TR-97-021, ICSI*, 1997.

[39] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall PTR, Englewood Cliffs, NJ, 1993.

[40] J. B. Kruskal and M. Wish, *Multidimensional Scaling*, Sage University Paper series on Quantitative Application in the Social Sciences, 07-011, Beverly Hills and London, 1978.

[41] Z. Al Bawab, B. Raj, and R.M. Stern, "A hybrid physical and statistical dynamic articulatory framework incorporating analysis-by-synthesis for improved phone classification," in *ICASSP*, Dallas, Texas. USA, March 2010.

[42] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, vol. 53, pp. 1070–1082, 1973.

[43] C.H. Coker, "A model for articulatory dynamics and control," *Proceedings of the IEEE*, vol. 64(4), pp. 452–460, 1976.

[44] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *ICASSP*, Istanbul, Turkey, June 2000.

[45] R. Singh, M.L. Seltzer, B. Raj, and R.M. Stern, "Speech in noisy environments: Robust automatic segmentation,feature extraction, and hypothesis combination," in *ICASSP*, Salt Lake City, Utah. USA, May 2001.

[46] X. Li, *Combination and Generation of Parallel Feature Streams for Improved Speech Recognition*, Ph.D. thesis, Carnegie Mellon University ECE Department, Pittsburgh, PA., February 2005.

[47] M. Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction," *Neural Computation*, vol. 11, pp. 1155–1182, 1999.

[48] K.F. Lee and F.A. Alleva, "Continuous speech recognition," in *Recent Progress in Speech Signal Processing*. S. Furui and M. Sondhi, 1990, Marcel Dekker Inc.

[49] E. Saltzman, "Task dynamic coordination of the speech articulators: A preliminary model," in *Experimental Brain Research Series 15*, New York, NY., 1986, H. Heuer and C. Fromm (Eds.), pp. 129–144, Springer-Verlag.