Doctoral Thesis Proposal Draft:
Using Articulatory Position Data to Improve Voice Transformation

Arthur R. Toth

November 3, 2005
# Contents

1 Introduction .............................................................. 1  
1.1 Speech Synthesis ................................................... 2  
1.1.1 Articulatory Speech Synthesis .............................. 2  
1.1.2 Formant-Based Speech Synthesis ............................ 2  
1.1.3 Concatenative Speech Synthesis ............................. 2  
1.2 Voice Transformation .............................................. 3  
1.2.1 Single-Speaker Model Goals ................................. 3  
1.3 Modeling Speech with Articulatory Positions ............... 5  
1.4 What is Good Voice Transformation? ........................... 8  
1.4.1 Metrics .......................................................... 8  
1.5 Thesis ............................................................... 10  
1.6 Document Structure ............................................... 10  

2 Related Work ............................................................ 11  
2.1 Voice Transformation .............................................. 11  
2.1.1 VQ Codebook-Based Techniques ............................ 11  
2.1.2 GMM Mapping-Based Techniques ............................ 12  
2.2 Articulatory Position Data ....................................... 15  
2.2.1 Mappings to/from Spectral Features ...................... 15  
2.2.2 Speech Recognition .......................................... 17  

3 Preliminary Work .......................................................... 18  
3.1 Data ................................................................. 19  
3.1.1 Training and Test Sets ...................................... 19  
3.2 Machine Learning Techniques .................................... 21  
3.2.1 Naïve Baselines .............................................. 21  
3.2.2 Linear Regression .......................................... 21  
3.2.3 CART .......................................................... 22  
3.3 Mappings to MFCCs ................................................ 22  
3.3.1 Cross-Speaker EMA ......................................... 23  
3.3.2 Mappings Between msak0 MFCCs and EMA ............. 23  
3.3.3 msak0 Mapping Results Summary ....................... 25  
3.4 Cross-speaker Mappings .......................................... 26  
3.4.1 Overview ...................................................... 26
3.4.2 Baseline Approach ........................................ 26
3.4.3 Improving the Mappings between FAF MCEP and MSAK0 MCEP ........................................ 31
3.4.4 Directly Mapping Between FAF MFCCs and MSAK0 EMA after DTW ........................................ 35
3.4.5 Discussion of Results ........................................ 37
3.5 Using EMA to Improve Speech Synthesis Models ................. 39
  3.5.1 Duration Models ........................................ 40
  3.5.2 $F_0$ Models ........................................ 41
  3.5.3 Power Model ........................................ 43
  3.5.4 Phonetic Feature Mapping ................................ 44
3.6 Producing Speech from Articulatory Positions ................. 49
  3.6.1 Synthesis using Articulatory Positions ................. 49
  3.6.2 Voice Transformation using Articulatory Positions ....... 50
3.7 Evaluation ........................................ 50
  3.7.1 Transformation Triangle Diagrams ...................... 53
  3.7.2 Evaluating Voice Transformation with TTDs .............. 55
4 Proposed Experiments ........................................ 58
  4.1 Experiments ........................................ 58
    4.1.1 Articulatory Positions in Speech Models .............. 59
    4.1.2 Parameterization for Voice Transformation ............ 60
    4.1.3 Evaluation of Voice Transformation Quality ............ 62
  4.2 Expected Contributions ................................... 63
5 Timeline ........................................ 64
Chapter 1

Introduction

Voice Transformation (also known as Voice Conversion or Voice Morphing) is a name given to techniques which take speech from one speaker as input and attempt to produce speech that sounds like it came from another speaker. One particularly compelling argument for good Voice Transformation is that it reduces the difficulty in creating synthetic voices with new identities and styles. Once a full-sized corpus has been collected from a source speaker, the amount of additional data needed to produce a new voice is much smaller than what is necessary to produce a new voice based on concatenative synthesis alone.

Although current Voice Transformation techniques appear to perform well in the sense that humans typically judge transformed speech to sound more like the target speaker than the source speaker, there is still room for improvement.

We propose the use of articulatory position data to improve Voice Transformation. When a person speaks, motions of the articulators affect the shape of the vocal tract, which affects the produced sound. Recently, data that includes measurements of the positions of various articulators along with recordings of the produced speech has been made publicly available. This articulatory position data gives us new information about the production of speech and has already been used successfully to predict quantities such as mel-frequency cepstral coefficients and acoustic-phonetic features. Such data gives us a different source of information from typical features derived from speech signals and enables promising new approaches to Voice Transformation.

As we are attempting to improve Voice Transformation, we will need to consider what this actually means. Although a number of objective and subjective tests have been used to judge Voice Transformation quality, the best way to evaluate it is still an open question. For this reason, we will also be investigating methods of Voice Transformation evaluation.

The rest of this chapter provides introductory material leading to a thesis statement and concludes with a description of the organization of the rest of this document. Section 1.1 gives a brief overview of speech synthesis to provide a context for our Voice Transformation work. Section 1.2 goes into a little more detail about Voice Transformation and related considerations that arise
concerning speech modeling. Section 1.3 discusses the possibility of using articulatory positions to model speech for Voice Transformation. This will be the main focus of this thesis. Section 1.4 discusses qualities that are important in the evaluation of Voice Transformation, and some of the metrics that have been used to measure them. Section 1.5 contains the thesis statement, and Section 1.6 describes the content of the remaining chapters and the bibliography.

1.1 Speech Synthesis

Speech synthesis is the name given to techniques used to create artificial human speech. It has a rich history going back to at least the 1939 World Fair, where a mechanical device called a Voder could be played like a keyboard instrument to produce synthetic speech [11]. Today, we are able to use computers to produce synthetic speech from entered text, and keyboard virtuosity (beyond typing) is less of an issue. There have been a number of computer-based approaches for producing synthetic speech, including articulatory approaches, formant-based approaches, and concatenative approaches.

1.1.1 Articulatory Speech Synthesis

Articulatory approaches have traditionally been based on vocal tract models that treat it as the concatenation of tubes of differing diameters. A notation consisting of electrical circuit components is often used to represent such models due to the similarity of the underlying mathematics and the ability to build the corresponding electrical circuits [10]. Eventually, computers could be used to simulate these circuits to perform speech synthesis [12]. Such approaches were popular many years ago, but they were very time-consuming in part due to the fact that computers were much slower then. There has been some renewed interest in this approach now that computers are much faster.

1.1.2 Formant-Based Speech Synthesis

Formant-based approaches focus on the resonant frequencies in the sound wave instead of the underlying physical activity that produces the sound. These approaches produce intelligible, though quite unnatural sounding, speech. Due to their greater tractability, these approaches have been used to create practical applications of speech synthesis. One system that exemplifies this approach is MITalk [3].

1.1.3 Concatenative Speech Synthesis

Concatenative speech synthesis techniques are currently very popular. They involve taking recordings of speakers, analyzing the recordings into units that are separated, stored, and concatenated (*i.e.* joined sequentially) to create new utterances. These techniques have the advantage of capturing some of the
character of the original speaker’s style, and thus tend to sound more natural than articulatory or formant-based approaches. However, it may be necessary to record a great number of units to cover all the phenomena that may appear in the desired synthetic speech. The quality of such a synthetic voice will depend on how well the recorded units cover the desired ones and also on the consistency of the recorded speaker over what may be a fatigu ing recording session or sessions.

Now that concatenative speech synthesis techniques allow the creation of voices that have more distinct identities, there is a desire to create many different voices. One difficulty with creating new voices is that not everyone has the skill or time to complete a full recording session.

1.2 Voice Transformation

One approach that simplifies the data collection process is Voice Transformation. The idea behind Voice Transformation is that a mapping between two speakers can be learned using fewer recordings than are required by concatena tive synthesis. If this mapping is of good quality, then creating a new synthetic voice would require making only a few recordings of a new speaker and applying the mapping to a previously recorded database.

But how would one go about learning a mapping between two speakers? There are numerous methods in the Voice Transformation literature that use statistical machine learning techniques to learn such mappings, and this appears to be a promising path to follow. Such methods require a model, which relates two speakers, and the corresponding data.

The model used for mapping between two speakers, in turn, depends on the model used to represent a single speaker. The data used in the single-speaker models is collected and used in the mapping between them. When processing speech from a single speaker, it is typical to use a model and perform further analysis on a time-varying amplitude signal of a spoken utterance. Not doing such an analysis and working directly from the signal can be problematic for a number of reasons. Signals that differ may only differ in ways that are irrelevant to the task at hand. If the goal involves human perception of synthesized speech, modeling differences that are not noticed or not valued differently by humans is wasteful. It may also lead to modeling unwanted noise and degrading the final output.

1.2.1 Single-Speaker Model Goals

What is important for such a single-speaker model? Such a model needs to represent quantities that are both derivable from the signal and related to the ultimate goals. There is a wealth of information on signal processing techniques used to derive information from speech signals. Much of this work has been performed for the sake of speech recognition, which has goals that are related to, but differ from those of speech synthesis. Furthermore, the goals of Voice Transformation typically include the goals of speech synthesis, but have additional
ones. The typical goal of speech recognition is to have a low Word Error Rate (WER). This means that words need to be correctly identified, but the idiosyncrasies of particular speakers are not important. In fact, it is considered better to have a representation that minimizes the effects of such information. The goals of speech synthesis, on the other hand, include producing speech that is intelligible to humans and natural. One would expect a good speech recognition model to be able to distinguish between various sounds used for words. Looking at the term “intelligibility” from a broader perspective, one could say that models for speech recognition are concerned with a representation that is intelligible to computers. It is not unreasonable to expect some connection between intelligibility for computers and intelligibility for humans. However, typical speech recognition is only concerned with the classification of these sounds and not the related effects that come from the human production of them which lead to its naturalness. In speech synthesis, prosodic quantities such as power, pitch, and duration are important, whereas in speech recognition, it is typical to minimize their effects in the model.

One very useful model of speech that has arisen from speech recognition research is the source-filter model depicted in Figure 1.2. It is claimed that treating the source and system response of the vocal tract as linearly separable, as in the source-filter model, is an assumption that has been used since the
beginning of speech synthesis theory [12]. In the source-filter model, the speech signal is represented as a source signal, or excitation, which is multiplied by a gain factor, and then convolved with a filter. This model has some physical justification. Speech can be seen as a process which starts from the diaphragm and lungs, which provide the power of respiration, continues through the vocal folds, which may impart quasi-periodic vibrations, and then passes through the vocal tract, which can alter the spectral properties of the signal including changes in resonance and frication (friction-based noise). In a source-filter model, the source represents the power from the diaphragm and lungs, plus either a pulse train to represent vibrations from the vocal folds or noise to represent frication. The filter represents the spectral characteristics of the vocal tract.

In speech recognition, it is typical to use some representation of the filter and to ignore the source. The filter carries most of the phonetic information, which is important for correctly identifying words and minimizing WER, and the source contains much prosodic information that may not be relevant for the task of minimizing WER. However, the prosodic information in the source is important for speech synthesis because it carries much of the naturalness of the speech. Furthermore, this prosodic information carries meaning and emotion, which may be desirable in synthesized speech. Additionally, it may be desirable to modify the prosody of synthetic speech without changing the phonetic information. A source-filter representation can facilitate such modifications by (mostly) separating phonetic and prosodic information into different components. Interestingly, for at least one type of source-filter model (pitch-synchronous LPC analysis), the residual, or amount of error the model makes, also appears to be related to naturalness and identity. Using the residual as the excitation, instead of a pulse train or noise, improves naturalness and identity.

Where does Voice Transformation fit into this? Because it shares the goals of intelligibility and naturalness with speech synthesis and is also concerned with identity, a source-filter model of speech appears to be a useful single-speaker model, and indeed has been used by numerous Voice Transformation researchers. For a summary of various speech tasks and related goals, see Figure 1.1.

We believe that using articulatory position information can improve the models used for Voice Transformation by improving the portion of the single-speaker model that is concerned with the vocal tract. In the source-filter model of speech, this would be the filter.

### 1.3 Modeling Speech with Articulatory Positions

The primary parameterizations of speech used in computer speech research are abstractions based on Digital Signal Processing (DSP) techniques. Linear Predictive Coding (LPC) coefficients, Mel-Frequency Cepstral Coefficients (MFCC), and figures derived from them appear to be the dominant features used in machine learning algorithms in automatic speech recognition. In speech synthesis, the most commonly used parameters include LPC, MFCC, and derived figures along with formants. Although these features allow the construction of
systems with relatively high performance and are relatively convenient for the
analysis and synthesis of speech waveforms, they are more based on the propo-
ties of the acoustic signal and are a bit removed from the actual physical process
of speaking.

While a person speaks, the produced sound is the result of respiration and
voicing combined with the motions of articulators, which control the shape of the
channel through which speech is produced. The locations of these articulators
is an additional, (literally) tangible knowledge source that should help in the
parameterization of speech and enable the construction of new models.

One problem with using articulatory position information, however, is that
it is difficult to collect. Unlike LPC and MFCC, which can be derived directly
from an acoustic waveform, which is easily collected through a microphone,
articulatory position information is typically collected through intrusive means.
Two of the better known technologies for gathering articulatory position data
are the Microbeam machine, and the Electromagnetic Articulograph (EMA).
Both machines require the attachment of sensors inside the mouth. Due to the
need for expensive, specialized machinery for data collection and the increased
burden on the test subjects, not much articulatory position data is available.

Given these circumstances, a couple natural questions to ask are:

1. How can articulatory position data be used to model speech?
2. How can the small amount of available articulatory position data best be
   leveraged?

For the first question, the answer is there are many possible ways to use ar-
ticulatory position data to model speech. Many mathematical approaches can
be divided into two basic categories. One set of techniques is based on trying
to use equations (e.g., Navier-Stokes or transmission line theory) from the fields
of fluid dynamics and acoustics to construct a system that predicts the output
sound wave based on the articulator positions and perhaps a few more param-
eters. These equations would be solved either analytically or approximately.
Such a system would be satisfying in the sense that it would truly model the
physical situation. Unfortunately, although some researchers have succeeded in
constructing speech synthesizers based on this approach, their success has been
limited, and in practice there are numerous problems. In order to make the
equations tractable, many simplifications and concessions must be made. Fur-
thermore, the types of parameters used by the equations may not be the same
as the ones that are measurable by the equipment.

Another set of mathematical approaches that can be applied to the first
question is machine learning techniques. Such techniques attempt to generalize
from whatever data is available. For that reason, to an extent (based on model
assumptions), these approaches may not be as hindered by the specific types
of parameters that are available or even if not all of the parameters necessary
for a complete model are available. However, there are other problems with
using machine learning. A general-purpose machine learning algorithm will not
directly model the interaction of articulatory positions and acoustic waveforms,
but will have its own set of simplifications and concessions based on its model assumptions.

Both approaches have advantages and disadvantages. As for mathematical modeling techniques, this thesis will only use machine learning approaches and will not use fluid dynamic or acoustic equations.

The next question then is: What relationship will these machine learning techniques be used to learn? As mentioned before, the speech signal is produced from a combination of articulatory positions and other factors. One way of simplifying the problem is to try to analyze the speech signal to determine some features that are more directly related to articulatory positions. It would also be helpful if the resulting parameterization of the speech signal allowed resynthesis of the speech signal (possibly with the addition of other features not as closely related to articulatory positions). LPC and MFCC seem to be reasonable choices according to these criteria. They both minimize the effects of voicing and power, and they both allow resynthesis of the speech signal. Furthermore, there is literature on reasonably successful attempts to use machine learning techniques to map between articulatory positions and MFCCs [26] [27] [15] [35] [34] [28].

Thus one way to address the first question is to state that articulatory position data can be used to model speech by using machine learning techniques to construct mappings between articulatory positions and Mel-Frequency Cepstral Coefficients, which along with other parameters, can be used in turn to construct a speech signal.

Now let us turn our attentions to the second question. We have a strategy for using articulatory position data to model speech, but so far it only applies to the speakers for which articulatory position data has been recorded, and there aren’t very many. If there were a way to map from the MFCCs of one speaker to those of another, then we could compose these maps to construct mappings between the MFCCs of one speaker and the articulatory positions of another speaker. Alternatively, we could treat the speech waveforms of two people speaking the same text as being equivalent, and we could use machine learning techniques to map directly between the MFCCs of one speaker and the articulatory positions of another speaker (after taking differences in duration into account). Again, machine learning techniques can be used to implement the maps.

Now that we have strategies for both questions, the next natural question is: “What are these mappings good for?”

We believe that articulatory positions are influenced by and influence a number of other factors in the production of speech that affect the acoustic waveform. We choose to focus on areas where we feel articulatory position information could be used to improve voice transformation. We believe the use of articulatory position data will improve the modeling of spectral characteristics in speech.

The specific articulatory data used in this thesis will be from the freely available MOCHA data [39], and is described in further detail in Section 3.1.
1.4 What is Good Voice Transformation?

In order to evaluate the results of various Voice Transformation techniques, it is necessary to ask what constitutes good voice transformation in general. As Voice Transformation is a technique for producing synthetic speech, it is useful to first ask what constitutes good synthetic speech. The goal of synthetic speech is to provide sounds that are easily understandable by a human listener. The speech should be intelligible in the sense that a person can recognize the exact words that are being synthesized. Furthermore, the speech should sound as human as possible.

It is also necessary to ask what additional considerations should be added for Voice Transformation. Because Voice Transformation has the additional goal of creating synthetic speech that sounds like it was produced by a particular speaker, speaker identity is an important component of Voice Transformation quality.

It should also be noted that speaker identity is an area that needs to be investigated more broadly with respect to Voice Transformation. In some applications, it may be important for transformed speech to be identifiable as coming from a specific individual. In other cases, it may be sufficient for transformed speech to come from distinct, though not necessarily recognizable, speakers. In yet other cases, for example when a person reads a children’s story where different characters speak, it may be sufficient to have the different voices sound like they come from the same speaker, and thus have the same identity, yet appear to represent different characters, and thus be distinct. Thus, specificity and distinctness are two possible dimensions of speaker identity which we may further investigate.

1.4.1 Metrics

How does one measure quantities such as intelligibility, naturalness, and identity? The literature has examples both of objective metrics based on calculations using quantities derived from the signal and of subjective metrics based on the judgments of human subjects.

Objective Metrics

For evaluating predicted articulatory positions, root-mean-square error (RMSE) seems appropriate as the articulatory positions actually are in Euclidean space. The formula for RMSE is:

\[
RMSE = \sqrt{\frac{1}{V} \sum_{v=1}^{V} \sum_{d=1}^{D} (t_{v,d} - e_{v,d})^2}
\]

where \( V \) is the number of vectors, \( D \) is the number of dimensions in the vectors, \( t_{v,d} \) is the \( d \)-th coefficient of the \( v \)-th target vector, and \( e_{v,d} \) is the \( d \)-th coefficient of the \( v \)-th estimate vector.
RMSE can also be used to evaluate predicted MFCCs. Another error measure for MFCCs that appears in the literature is called Mel-Cepstral Distortion (MCD) [35]. The formula for the MCD between two 24-dimensional vectors from [35] (corrected after correspondence with the first author) is:

\[ MCD = \frac{10}{\ln 10} \sqrt{2 \sum_{d=1}^{24} (mc_d^{(t)} - mc_d^{(e)})^2} \]

where \( mc_d^{(t)} \) is the \( d \)-th coefficient of the target and \( mc_d^{(e)} \) is the \( d \)-th coefficient of the estimate.

The above MCD formula is for a single vector pair. MCD means and standard deviations were calculated for MFCC prediction experiments to be more easily comparable with results from [35]. One thing to note is that MCD mean, while somewhat similar to RMSE scaled by a constant, is different in that the mean over all vectors is taken after the step involving the square root.

**Subjective Metrics**

Although objective metrics are often simpler to provide and give some sense of how close synthetic speech is to a target, the true judge is the human ear. Although some researchers have demonstrated correlations between objective metrics such as MCD and subjective metrics such as listener preferences [18] [36], there are still cases when synthetic speech that scores better according to some objective metric sounds worse to a human listener. This is because no known objective metric has performance that is close enough to human perception. Furthermore, if we use human judgment as our metric, we also encounter the problem of different people having different preferences. For this reason, we call such metrics subjective. The typical strategy for handling differences in human judgment is to collect judgments from a large enough number of people in the hopes that the average scores from the population will minimize the biases of a few idiosyncratic individuals.

Given the circumstances, it is natural to ask what kind of subjective metrics should be used for judging the intelligibility of speech, the naturalness of speech, and what kind of subjective metrics should be used to judge the identities of the speakers. For these questions, we turn to the existing literature and find a number of answers. Intelligibility and naturalness of speech can be measured similarly to the ways they were measured in the 2005 Blizzard Speech Synthesis Evaluation [6]. Intelligibility can be measured by asking subjects to listen to synthetic speech and write or type what they heard. Naturalness of synthetic speech can be measured by asking subjects to score the naturalness on a scale, for example from 1 to 5, where 1 means unnatural and 5 means natural. For the judgment of speaker identity, we would favor using pair comparison tests where for each trial, the subjects are given two utterances, which may be recordings and/or transformed speech, and ask them to score the similarity of the two speakers on a scale. Although another possibility would be to perform “ABX”
tests in which each trial consists of having the listener decide whether the transformed speech sounds more like a recording of the source speaker or the target speaker, we agree with [18], in that such a test does not measure the possibility that the transformed speech may sound dissimilar to both.

1.5 Thesis

Over time, the dominant speech synthesis techniques have progressed from physical and acoustic models to data-driven models. This has enabled the creation of synthetic voices that sound more natural and have more recognizable identities, but the flexibility of such systems is limited by the amount and type of collected data. Now that there has been some success with data-driven techniques, there is a desire to reincorporate some of the flexibility of model-based techniques in an effort to reduce the amount of necessary data. Articulatory positions present a tangible way to help parameterize speech data that differs from traditional methods that are based on analysis techniques from Digital Signal Processing (DSP). In addition to allowing the modification of parameters in a different space, articulatory positions are also subject to static and dynamic physical constraints that can be incorporated into models of speech to make them more realistic. New methods for analyzing speech based on modeling of articulatory positions can be used to improve Voice Transformation. Voice transformation can be improved in terms of intelligibility, naturalness of speech, and identity of speaker.

1.6 Document Structure

The rest of this document is organized as follows. Chapter 2 surveys work done by other people that is related to our thesis. Chapter 3 describes preliminary experiments we conducted to examine the feasibility of our approaches. Chapter 4 motivates and lists the experiments we propose to perform to investigate our thesis. Chapter 5 is a timeline for the proposed work. This document concludes with a bibliography that lists the sources referenced in this proposal.
Chapter 2

Related Work

This chapter describes work done by other researchers that is related to our thesis. As there doesn’t appear to be any literature on using articulatory position data with Voice Transformation, we will cover the areas individually. Section 2.1 will cover work on Voice Transformation and Section 2.2 will cover work on articulatory position data.

2.1 Voice Transformation

The amount of literature on Voice Transformation is sizable and goes back to at least 1985 [8]. Earlier work focused more on Vector Quantization (VQ) codebook-based techniques [2]. One of the problems with codebook-based approaches is that the output vectors often had discontinuities that led to artifacts in the synthesized speech. VQ-based techniques have been superseded by techniques that produce continuous output. Some techniques are still based on clustering techniques, but additional measures are taken to provide some smoothing of discontinuities. Other techniques, such as Gaussian Mixture Model (GMM) mapping-based techniques, are based on continuous models from the beginning.

2.1.1 VQ Codebook-Based Techniques

Abe et al.

A Voice Transformation technique based on VQ codebooks was introduced in [1]. The process for learning a mapping between the codebooks for two different speakers went as follows:

1. Extract LPC coefficients from frames from words recorded by source and target speakers.
2. Use VQ on the resulting LPC coefficient vectors.
3. Align each word between speakers using Dynamic Time Warping (DTW).
4. Create histograms for the vector correspondence between speakers.

5. Create a mapping codebook that uses the histograms to create linear combinations of the target speaker’s vectors.

6. Repeat the DTW, histogram creation, and mapping codebook creation until the mapping codebook is considered to be good enough.

Mapping codebooks were also created for pitch and power, but the procedures were slightly different as the values were scalars, and the maximum histogram occurrence was used in creating the pitch mapping codebook.

Transformation was performed by extracting the LPC coefficients, pitch, and power from a new recording by the source speaker, decoding them using the previously learned mapping codebooks, and synthesizing from the resulting values.

In the analysis of their experimental results, the authors concluded that both pitch and spectrum were necessary for the individuality of the speech, and that neither alone was sufficient. All of the male-to-female transformed speech was judged female by listeners, and 65% of the male-to-male transformed speech was judged more similar to the target speaker than the source speaker by listeners.

### 2.1.2 GMM Mapping-Based Techniques

Some of the more recent Voice Transformation work has been based on Gaussian Mixture Model (GMM) mapping techniques. These techniques first use aligned data from source and target speakers to learn parameters for a GMM over the joint source/target feature space. Then, various techniques can use this GMM to map a source speaker utterance to the target speaker utterance space.

The probability density for a vector, \( x \), under a \( p \)-dimensional Gaussian distribution, \( \mathcal{N} \) with a mean vector, \( \mu \), and a covariance matrix, \( \Sigma \), is:

\[
\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}} \left| \Sigma \right|^{-1/2} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]
\]

A GMM is a weighted sum of Gaussian probability density functions subject to a number of constraints and defines a probability distribution function, \( p(x) \), for a vector, \( x \), as follows:

\[
p(x) = \sum_{i=1}^{M} \alpha_i \mathcal{N}(x; \mu_i, \Sigma_i)
\]

where \( M \) is the number of Gaussian components, \( \mu_i \) is the mean of the \( i \)th Gaussian, \( \Sigma_i \) is the covariance matrix of the \( i \)th Gaussian, and the \( \alpha_i \) weights are subject to the constraints that

\[ \forall i, \alpha_i \geq 0 \]
The parameters of a GMM are typically learned by taking a number of data vectors that are assumed to have been generated by the GMM and using the EM algorithm [9] to search for parameters that will maximize the likelihood of the data vectors. In the following approaches, these data vectors consist of features derived both from source and target speakers.

Different researchers have devised various ways to use GMMs to perform transformations from source speakers to target speakers. These methods are described below in more detail.

**Yannis Stylianou et al.**

The approach of using GMMs for Voice Transformation appears to have originated with Yannis Stylianou and his coauthors [31] [30]. They trained GMMs for voice conversion as described above.

In the later work [30], acoustic features were derived from a Harmonic + Noise Model (HNM) [32]. In this model, speech is divided into a low band modeled by harmonically-related sine waves, and a high band modeled by noise modulated by a time-domain amplitude envelope. This representation was chosen because modifications to duration and pitch seemed to be relatively straightforward to perform and of high-quality. In this work, the low and high bands were converted separately, and the cutoff was fixed at 4 kHz. The low band was converted using a GMM mapping, and the high band was converted using two different filters. One filter was for voiced frames, and the other was for unvoiced frames.

Features were extracted from the low band by converting the harmonic frequencies to a Bark frequency scale and then extracting cepstral coefficients every 10ms. The 1st through 20th coefficients from the source and target speaker were then aligned using a Dynamic Time Warping (DTW) algorithm and used to train a GMM with full covariance matrices.

The mapping function was constructed as a generalization of the minimum means square estimation function for a single Gaussian distribution. It has the form:

\[ F(x) = \sum_{i=1}^{M} P(C_i|x)[\nu_i + \Gamma_i\Sigma_i^{-1}(x - \mu_i)] \]

where \( \nu_i \) is the target mean vector for class \( i \), \( \Gamma \) is the cross-covariance matrix of the source and target vectors: \( \Gamma = E[(y - \nu)(x - \mu)^T] \), and \( P(C_i|x) \) is the conditional probability density of the \( i \)th Gaussian class given data vector \( x \):

\[ P(C_i|x) = \frac{\alpha_i(N)(x; \mu_i, \Sigma_i)}{\sum_{j=1}^{M} \alpha_jN(x; \mu_j, \Sigma_j)} \]

A couple tests were performed for evaluation using 20 listeners. One test was an “XAB” test (called an “ABX” test in other literature [18]) which provided
an example “X” of converted speech and asked listeners to compare it with examples “A” and “B” of source and target speech and decide which was more similar. Listeners overwhelmingly found the voice conversion technique with 16 Gaussians in the mixture model to perform better than a technique which only modified prosody, and further improvements occurred when 64 Gaussians were used and the same sentence was used for all three examples. In this last example, 97% of the converted utterances were considered to be closer to the target speech. This demonstrates that converting spectral characteristics in addition to prosodic characteristics improves the quality of Voice Transformation in terms of speaker identity. Another test called an “opinion” test (called a pair comparison test in [18]) involved presenting listeners with pairs of utterances where they could be actual speech, converted speech, or prosodically-modified speech. The listeners were then asked to rate the similarities of the pairs on a scale from 0 meaning “identical” to 9 meaning “very different”. The results of this test corroborated the results from the “XAB” test.

**Alex Kain’s Ph.D. Dissertation**

Alex Kain’s Ph.D. dissertation [18] extended the idea of using GMM mappings in Voice Transformation in a few more directions. He tried different acoustic features, used a different alignment strategy, and also attempted to predict residuals based on his model.

In this work, speech was parameterized in a different way. Instead of spacing frames every 10ms, they were spaced pitch synchronously. Unvoiced speech was assumed to have a constant pitch of 125 Hz. Speech was represented using a harmonic sinusoidal model. Then, the frequencies were warped according to the Bark scale, and Linear Predictive Coding (LPC) coefficients were calculated. Because LPC coefficients are difficult to modify in a stable manner, Line Spectral Frequencies (LSF) [17] were then computed from the LPC coefficients. These Line Spectral Frequencies were used as features in the Gaussian Mixture Model.

Alignment of source and target features proceeded differently from Stylianou’s DTW approach. “Time marks” were created by performing HMM-based forced alignment on the source and target speech using phonetic transcriptions. The HMM states were aligned, and frames were repeated in or deleted from the target speech to match the length of the source speech.

Although the transformation function, like Stylianou’s, is based on the expectation of the target features conditioned on the source features, synthesis proceeds differently due to differences in the speech models and frame selection.

One of this dissertation’s major contributions is the idea of attempting to predict the LPC residual for the target speaker. This is done by deriving cepstral coefficients from the LPC coefficients, and training a GMM classifier for deciding class membership for vectors of such cepstral coefficients. Classifications from the GMM are used to create a codebook of residuals, where residuals are represented as 100-point samples. In order to predict a residual for a vector of LPC cepstral coefficients, the GMM parameters and residual codebook are used to produce a weighted sum of residual codebook vectors, which is the estimate.
Residual prediction was found to improve the speaker identity of transformed speech in subjective evaluations. This supported the thesis that LPC residuals include information that is related to speaker identity.

**Tomoki Toda’s Ph.D. Thesis**

Part of Tomoki Toda’s Ph.D. Thesis [36] was concerned with using GMM mapping techniques with Voice Transformation. He created a Voice Transformation technique that used a combination of GMM mapping and Dynamic Frequency Warping (DFW). Subjective evaluations supported the conclusion that the quality of the synthesized speech was improved while the speaker identity remained just as good, when compared to a conventional GMM mapping technique that did not use DFW.

The acoustic features, analysis, and synthesis in this thesis differed from the ones in the other approaches. For the acoustic features, STRAIGHT [19] analysis was used to produce a smoothed spectrum, and MFCCs were calculated from this spectrum at a fixed interval. Alignment of source and feature vectors was performed by removing silence frames and then using a Dynamic Time Warping algorithm on the remaining frames. Then an iterative process is used to improve the alignment. First, a GMM mapping is trained on the source and target speakers’ features. Then alignment is performed again using the transformed version of the source speaker’s features and the target speaker’s features. This process is repeated iteratively until the change in the Mel-Cepstral Distortion is less than a threshold.

Later innovations included using the EM [9] algorithm to produce maximum likelihood estimates of the target features instead of just using expectations, using additional “dynamic features” (weighted windows of features), and changing the optimization function from maximum likelihood to a weighted mixture of maximum likelihood and global variance.

### 2.2 Articulatory Position Data

Although we are unaware of other attempts to incorporate articulatory position data into Voice Transformation, there have been numerous attempts to use it in speech models. A number of researchers have investigated mappings between articulatory features and spectral features. A couple groups have gone a step further and used these mappings to synthesize speech from articulatory positions. Others have attempted to use it to improve speech recognition performance.

#### 2.2.1 Mappings to/from Spectral Features

We are unaware of any other work which attempts to map between the spectral characteristics of one speaker and the articulatory positions of another. However, there have been a few attempts to map between the spectral characteristics and articulatory positions for a single speaker. The most current attempts at
single-speaker mappings were in Korin Richmond’s Ph.D. thesis [26] (and related journal article with additional coauthors [27]), a conference paper by Hiroya and Honda [15], a couple conference papers by Tomoki Toda et al. [35] [34], and a conference paper by Yoshinori Shiga and Simon King [28].

**Korin Richmond’s Ph.D. Thesis**

In Korin Richmond’s work, a Mixture Density Network (MDN), which is a special type of Artificial Neural Network (ANN) used to learn Gaussian Mixture Model (GMM) parameters, was used to model the “inverse mapping” from MFCCs to articulatory positions. This approach was taken because it was felt that a MDN would be more able than a Multi-Layer Perceptron (MLP), which is a more typical ANN, to represent the many-to-one aspect of the mapping in this direction. The results of this work indicated that the MDN did in fact perform better than the baseline MLP.

**Tomoki Toda et al.**

The results of this work were improved upon in work by Tomoki Toda et al. [35] and [34]. The first paper investigated mappings from articulatory positions to MFCCs, and the second paper investigated mappings from MFCCs to articulatory positions. In addition to baseline results using only articulatory positions and MFCCs, other features, including power and the logarithm of the fundamental frequency, were shown to improve performance. For both mapping directions, the strategy was to treat the joint distribution of all of the features (both predictors and predictees) as a Gaussian Mixture Model. Prediction was then performed by setting the predictor features to the observed values and using the Expectation-Maximization (EM) algorithm to search for predictee feature values that would maximize the likelihood. The results from these papers were used for comparison with our preliminary experiments. It is interesting to note that these papers used the same mapping techniques that were used in Voice Transformation work by Tomoki Toda [36]. In one of these papers [35], the mapping from articulatory positions to MFCCs was used to synthesize speech, which was then used to evaluate the quality of various articulatory-to-acoustic mappings. The synthesis was performed by converting the MFCCs to linear spectra, controlling power, extracting $F_0$, and producing a signal using STRAIGHT synthesis [19]. The evaluations demonstrated that the improvements in Mel-Cepstral Distortion measures from using source information, maximum likelihood estimation, and dynamic features corresponded to improvements in approval rating for the synthesized speech.

**Shiga and King**

Shiga and King used another approach to map from articulatory positions to cepstral coefficients [28]. Their approach used LBG clustering of the articulatory feature space followed by learning linear maps from each of the articulatory
feature clusters to the associated cepstral feature vectors. In a later paper [29], sinusoidal synthesis [22] was used with cepstra predicted from articulatory positions. It was reported that the resulting synthesis was informally confirmed to be intelligible and of high quality.

**Hiroya and Honda**

Hiroya and Honda used what they called an HMM-based Speech Production Model to perform a mapping from MFCCs to articulatory positions [15]. Their HMM-based Speech Production Model consisted of a Hidden Markov Model (HMM) that predicted articulatory position feature vectors from phone sequences and linear mappings from articulatory positions to MFCCs that were conditioned on the state from the Hidden Markov Model.

### 2.2.2 Speech Recognition

Going a little farther afield, a number of researchers have tried to use articulatory position data to improve speech recognition. One thing to note is that these experiments tended to use far less data than typical speech recognition experiments because they required articulatory position data which is much less available than the typical MFCCs used for speech recognition.

**Frankel and King**

Frankel and King used linear dynamic models with MFCCs and articulatory positions as input [13]. They found that using a combination of MFCCs and articulatory positions worked better than using either one alone.

**Markov et al.**

Markov et al. used articulatory positions and MFCCs in what they called a hybrid HMM/BN speech recognition system [21]. Their model was a Dynamic Bayesian Network (DBN) that was equivalent to a HMM with GMM observations. The articulatory positions were essentially used as indexes for the various Gaussians. They also experimented with various groupings of features with their velocities and accelerations. They found that speech recognition could be improved when articulatory position features were used in addition to MFCCs.
Chapter 3

Preliminary Work

An amount of preliminary work was performed to investigate the plausibility of using articulatory position features with Voice Transformation. In Sections 3.1 and 3.2, we describe the data and machine learning techniques used in our experiments.

Our first experiments involved learning mappings between MFCCs and articulatory position data (which we often refer to as EMA for Electromagnetic Articulograph). The results of these experiments are in Section 3.3. Due to the previously mentioned sparsity of available articulatory position data, we then experimented with cross-speaker mappings where we tried to predict the articulatory positions for one speaker from the MFCCs of another and vice-versa. These techniques are described in Section 3.4 and form the basis of a number of later experiments.

After completing these basic experiments and obtaining results that seemed reasonable in comparison to similar experiments in the available literature, we then experimented with using articulatory position data to predict quantities that were useful for speech synthesis. First we tried to predict prosodic quantities, such as duration, $F_0$, and power. This was mostly unsuccessful, which was not very surprising as these quantities are more closely related to the source, and articulatory positions are more closely related to the filter in the source-filter model of speech. After attempting to predict prosodic quantities, we tried to predict acoustic-phonetic features using articulatory positions and cross-speaker articulatory positions. When actual articulatory positions were used for a speaker, the accuracy of a number of acoustic-phonetic feature predictors improved beyond a baseline of only using MFCCs, thus demonstrating that there was additional useful information in the articulatory position data, at least in the context of our modeling strategy. For the cross-speaker experiments, only a few acoustic-phonetic feature predictions were improved with the predicted articulatory features. All of these experiments which attempted to predict quantities useful for speech synthesis are described in Section 3.5.

After trying to predict prosodic quantities and acoustic-phonetic features, we tried to produce speech from articulatory positions. Our initial attempts...
to synthesize speech from actual articulatory positions (using additional source information) and from articulatory positions transformed from one speaker to another are described in Section 3.6.

Finally, we explored the area of Voice Transformation evaluation. An evaluation was conducted to investigate the effects of speaker familiarity with perception of speaker identity in Voice Transformation. This is described in Section 3.7. For the types of listening experiments we conducted, it appeared that knowledge of the speaker did not significantly affect the perception of closeness of speaker identity. This will be kept in mind as we design further listening experiments. Also, we created a new type of diagram, called a Transformation Triangle Diagram (TTD) to help visually summarize the results of a specific type of listening test. We found these diagrams convenient for quickly seeing the relationships between a number of relevant quantities from the evaluation.

3.1 Data

3.1.1 Training and Test Sets

The corpora used in the following experiments are the MOCHA msak0 and fsew0 databases [39], and the Facts and Fables cmuacusartfaf database [41].

We decided to use a tenth of the data as a held-out test set. Due to possibilities of drift in the recorded data, we decided to start with the basic strategy of holding out every tenth utterance for the test set. Drift can arise both due to electrical fluctuations in the recording equipment, and also as a result of speaker fatigue. Furthermore, for the MOCHA data, which was collected intrusively, there may be drift due to the speakers adapting to the various probes in their mouths during the recording session. By choosing roughly every tenth utterance for the held-out test sets, we hope to have collections that are representative of the data as a whole.

For certain types of training, such as stepwise training for CART, it was necessary to have an additional held-out set. For experiments needing an additional held-out set, the msak0 training set was split again. The files put in the additional held-out set are the ones in the training set whose numbers are congruent to 9 mod 10. For example, for files msak0_001 through msak0_010:

- training set: msak0_001 through msak0_009
  - reduced training set: msak0_001 through msak0_008
  - additional held-out set: msak0_009
- test set: msak0_010

The fsew0 data were split into training, test, and held-out sets in the same manner.
MOCHA Files

According to the README.txt file packaged with the msak0 database, it consists of 460 single-sentence utterances from the British TIMIT corpus. They are labeled from msak0001 through msak0460. Four types of data files were recorded for each utterance:

- Acoustic files recorded at a sample rate of 16kHz with 16 bits per sample with NIST headers
- Laryngograph files record at a sample rate of 16kHz with 16 bits per sample with NIST headers
- Electromagnetic Articulograph (EMA) files of 10 (x,y) coordinate pairs recorded at a sample rate of 500Hz with 16 bits per value (stored as 4 byte floats)
- Electropalatograph files recorded at 200 samples per second of 62 bits of tongue contact data padded to 64 bits.

The README.msak0.v1.1 file packaged with the msak0 data mentions the following problems that occurred while collecting data:

- msak0040: reattached tongue dorsum coil
- msak0118: corrupted tongue tip signal
- msak0251: reattached velar coil
- msak0268: corrupted audio and laryngograph files
- msak0299: reattached tongue dorsum coil
- msak0426: reattached tongue body coil

Visual inspection of EMA plots discovered the following additional problem:

- msak0050: some tongue tip locations are implausible - the data appear to be corrupted

These utterances were discarded from the training and test sets. If the number of an msak0 utterance was not divisible by 10 and it was not in the above list of problematic files, it was placed in the training set. If the number of an msak0 utterance was divisible by 10 and it was not in the above list of problematic files, it was placed in the test set. For experiments that required an additional held-out set, the reduced training set and additional held-out set were selected as previously described and the problematic utterances were discarded.

The lengths of the msak0 acoustic files were compared to their corresponding EMA files. The longer ones were truncated to match the length of the shorter ones.

Processing for calibration and alignment had already been performed on the released files. No further calibration and alignment were performed.
The fslew0 data had the same types of files as the msak0 data, but did not have problems with data corruption, so when we were able to base our fslew0 training and test sets on all 460 utterances.

**cmu_us_art_faf Files**

The *Facts and Fables* database consists of 107 utterances, but they are longer than the MOCHA utterances, and can consist of multiple paragraphs of text. In all, there are over 14,000 words in the *Facts and Fables* utterances. cmu_us_art_faf is a freely-available set of recordings of the *Facts and Fables* database by a male with a Midwestern American accent [41].

Of the file types mentioned for the msak0 database, the cmu_us_art_faf database only has acoustic files recorded at a sample rate of 16kHz with 16 bits per sample.

For trials involving only cmu_us_art_faf data, the files were split into training, test, reduced training and additional held-out sets using the process previously described after numbering the sorted files names consecutively. (The utterance names contain numbers which are not consecutive.)

### 3.2 Machine Learning Techniques

In the following experiments, naive baselines were evaluated in some cases, and the machine learning techniques of linear regression and Classification and Regression Trees (CART) were used to predict quantities in some cases.

**3.2.1 Naive Baselines**

As a sanity check to compare against various mapping approaches, in cases where numerical values (as opposed to categorial values) were predicted, a couple naive strategies were attempted:

- Always predict values of 0.
- Always predict values equal to the training set means.

For predictive value, these approaches both give constant results, and are obviously useless for modeling dynamic quantities such as EMA and mel-cepstral coefficients. However, calculating results for these approaches on the test set provides naive baselines to roughly compare against other mapping approaches.

**3.2.2 Linear Regression**

Linear regression is one common statistical model used to perform mappings. The underlying idea is that each datum, $y$, arises from a process where a predictor datum, $x$, undergoes a linear transformation and has zero-mean Gaussian noise added to it. It can be shown that the maximum likelihood estimate (MLE) of the linear transformation matrix, $B$, can be determined using the method of
least squares. This linear transformation matrix can then be used to map \( x \) values to \( y \) values.

Matlab was used to perform linear regression experiments from mel-cepstral coefficients to EMA values and from EMA values to mel-cepstral coefficients. In both cases, the predictor data was augmented with a column of ones. This is standard practice when performing linear regression because it allows the possibility of a non-zero y-axis intercept.

### 3.2.3 CART

Although CART is often used for classification, it can be used for regression as well (it’s what the “R” stands for, after all). Multiple trees were built and results were combined to produce mappings from msak0 mel-cepstral coefficients to msak0 EMA values.

### 3.3 Mappings to MFCCs

The first experiments were to investigate what articulatory positions could be used for. It appeared that they should be useful for models of speech as they contribute so heavily to the physical production of speech, but it was unclear whether they would provide anything beyond the quantities that were available through conventional analysis of speech.

One of the first areas we investigated was mappings between articulatory positions and MFCCs.

Preliminary experiments were performed on single-speaker and cross-speaker mappings between articulatory positions and MFCCs in order to investigate the feasibility of this work. These experiments were intended to provide performance baselines. In order to conduct such experiments, it was necessary to construct training and test sets from the available data and to choose which machine learning techniques would first be applied.

A number of criteria were considered when selecting machine learning techniques for the preliminary experiments:

1. ability to handle continuous data
2. reasonable model assumptions
3. relatively quick performance
4. reasonable quality of performance

Based on these criteria, linear regression and Classification and Regression Trees (CART) were selected for machine learning techniques. Both techniques are able to handle continuous data. Their model assumptions are fairly general as they have been successfully applied to data across many domains. Trying two different types of machine learning techniques with different assumptions increases the chance of avoiding pathological assumptions. Also, these techniques
are relatively quick. Due to the large number of combinations of preliminary experiments, it was necessary to find methods that would take hours at most for each trial, instead of days. The quality of the performance in the following experiments seemed reasonable when compared to naive baselines and also reasonable when compared to other baselines in the literature, where available.

3.3.1 Cross-Speaker EMA

For some trials involving cross-speaker evaluation, it was helpful to have utterances for the same text from both databases. Because there was different text in the two databases, something had to be done to produce the missing utterances. The first step was to use a unit selection synthesis voice based on the cmu_us_art_faf data to synthesize the utterances in msak0. Other possibilities include synthesizing the utterances with the cmu_us_art_faf HTS (Hidden Markov Model based synthesis) voice, and having the speaker record the msak0 utterances. (This depends on speaker availability.)

The synthesized utterances using text from the msak0 training set were treated as the training set for cross-speaker evaluation, and the synthesized utterances using text from the msak0 test set were treated as the test set for cross-speaker evaluation. Reduced training and additional held-out sets were similarly based on which utterances had text in the msak0 reduced training and additional held-out sets.

Due to differences in phonetic coverage, not all of the msak0 utterances could be synthesized by the cmu_us_art_faf voice. However, all of the utterances in the msak0 test set were successfully synthesized. The text from the following msak0 utterances could not be synthesized by the cmu_us_art_faf voice due to a lack of phonetic coverage:

- msak0_279
- msak0_353
- msak0_361
- msak0_372

For trials that used DTW based on the Itakura rule [17], the following utterance had to be discarded due to failure of DTW (the synthesized utterance was less than half as long as the msak0 utterance and the Itakura rule doesn’t allow alignment under such circumstances):

- msak0_198

3.3.2 Mappings Between msak0 MFCCs and EMA

EMA Features

The msak0 articulatory data files consist of 10 (x,y) coordinate pairs in the mid-sagittal plane for each sample. Two of these pairs are for the bridge of
the nose and upper incisor and serve as calibration points. Another pair is labeled “xxxx” in the data and is not used for anything. (The corresponding probe was lying outside the EMA helmet during the experiment). The seven remaining coordinate pairs are used for predictions in this work and consist of measurements for:

- lower incisor
- upper lip
- lower lip
- tongue tip
- tongue body
- tongue dorsum
- velum

The EMA data are sampled at 500Hz, and the scale is thousandths of centimeters. The origin is roughly the location of the upper incisor calibration point, the x-coordinates increase from the front to the back of the head, and the y-coordinates increase from the bottom to the top of the head. In other words, when you graph the data, it looks like a profile of the speaker facing left.

**MFCC Features**

Mel-cepstral features were generated a few different ways before settling on a final working version. For completeness and comparison, the results of various abandoned approaches are included.

For the first experiments, the mgcep program from the SPTK toolkit (freely available from http://kt-lab.ics.nitech.ac.jp/ tokuda/SPTK/) was used to extract mel-generalized cepstral coefficients. They were used first, because there was another program available that could generate an acoustic waveform from them, and this appeared useful for future experiments. The 0th through 24th mel-generalized cepstral coefficients were extracted using a 25ms Blackman window, and frame offsets of 2ms from the acoustic files in the msak0 data using the SPTK toolkit (freely available from http://kt-lab.ics.nitech.ac.jp/ tokuda/SPTK/). The number of coefficients and window size were chosen to be consistent with other work in this area [35]. The frame offset was chosen to match the EMA sample rate. An alpha value of 0.42 and a gamma value of -0.5 were used to be consistent with the example in the SPTK documentation. It was confirmed that a reasonably understandable audio file could be resynthesized from a mel-cepstral file extracted in this way.

A few experiments were conducted using this data, but the size of the data was too large to allow running certain CART experiments in the amount of time that was allotted. For this reason, a second set of experiments was created that used the same mel-generalized cepstral parameters, but were instead extracted
with a frame advance of 10ms. This reduced the data to roughly one-fifth of the size, and it was then possible to continue with more involved experiments. Also, it did not appear to hurt the results on the simpler experiments after switching to a smaller data set.

After running a number of experiments and comparing the results to the results from [35] and our previous experiments, we noticed that the error results for the mel-generalized cepstral coefficients were almost an order of magnitude smaller than expected. Looking at some data files, and reading more of the SPTK documentation led to the realization that there was a separate mcep program that extracted (non-generalized) mel-cepstral coefficients, and that they could also be used for resynthesis. The mcep program was used to extract mel-cepstral coefficients with the same parameters as for the mgcep program, except for the gamma value, which isn’t valid for mcep. It was also confirmed that a reasonably understandable audio file could be resynthesized from a mel-cepstral file extracted in this way. Then, the mel-cepstral coefficients were compared with the coefficients that were used in [35] and found to be much closer in range. Further experiments with these (non-generalized) mel-cepstral coefficients gave results that were more consistent with previous results, so the experiments using these values are considered to be the main ones in this report. Results for the other types of extracted coefficients are given for completeness and comparison.

The 1st through 24th coefficients were considered to be the features for prediction. The 0th mel-cepstral coefficient is different in nature from the other mel-cepstral coefficients in that it is related to power in the signal. The baseline mel-cepstral to EMA experiment in [35] uses the 1st through 24th coefficients, so they were used as the baseline in this work. The baseline EMA to mel-cepstral experiment in [34] did add the 0th coefficient as well, but the following work did not at first add it in order to have consistent features in the two mappings, especially because they would be composed. It was unclear whether the 0th mel-generalized cepstral coefficient was also different in nature from the other mel-generalized cepstral coefficients, so an extra round of experiments was performed which used the 0th through 24th mel-generalized cepstral coefficients.

Round-Trip Mappings

After constructing mappings from msak0 mel-cepstral values to EMA values and back, it is possible to compose the mappings to construct mappings from msak0 mel-cepstral values to msak0 mel-cepstral values.

3.3.3 msak0 Mapping Results Summary

The results of the attempted mappings between the mel-cepstral coefficients and the EMA values using only msak0 data are summarized in Table 3.1.

In every case, except for the round-trip mapping using mel-generalized cepstral coefficients including the 0th, CART had the best performance among the
baseline mapping approaches. In the one remaining case, linear regression performed the best, however the results in this case were so close that they may not be significant.

The EMA RMSE averages are further broken down in Table 3.2 to allow further comparisons among themselves and to [34].

3.4 Cross-speaker Mappings

3.4.1 Overview

Next, we investigated mappings from mel-cepstral coefficients from the cmu_us_art_faf database to EMA values from the msak0 database and, inversely, from EMA values from the msak0 database to mel-cepstral coefficients from the cmu_us_art_faf database. A mapping from cmu_us_art_faf mel-cepstral coefficients to msak0 mel-cepstral coefficients can be applied, followed by one of the previously mentioned msak0 mel-cepstral to msak0 EMA value mappings to construct the first type of mapping. A mapping from msak0 mel-cepstral coefficients to cmu_us_art_faf mel-cepstral coefficients can be applied after applying one of the previously mentioned msak0 EMA value to msak0 mel-cepstral coefficient mappings to construct the second type of mapping.

Evaluating mappings from cmu_us_art_faf mel-cepstral coefficients to msak0 EMA values and from msak0 EMA values to cmu_us_art_faf mel-cepstral coefficients individually is difficult because in some sense, there are no measured correct values for the locations of one person’s articulators while another person speaks, and it is not clear what should be considered correct. If there are utterances for the same text from both speakers, we could consider the utterances equivalent in some sense and compare the outputs from the mappings from one speaker to the results of the other speaker. One complication here is that the utterance lengths for the different speakers can vary. For our purposes, we will use a Dynamic Time Warping (DTW) approach using the Itakura rule [17] on an utterance-by-utterance basis and RMSE and DTW on the resulting paths to compare different length utterances.

If both maps are composed to create a “round-trip”, the result should be a mapping from cmu_us_art_faf mel-cepstral coefficients to cmu_us_art_faf mel-cepstral coefficients. This mapping is more straightforward to evaluate because the number of the predicted mel-cepstral coefficients will match the number of the actual coefficients. Then RMSE and MCD can be calculated without DTW.

3.4.2 Baseline Approach

One approach is to notice that the cmu_us_art_faf mel-cepstral coefficients and the msak0 mel-cepstral coefficients are in the same domain and the identity mapping can be used between them. Then the mapping from cmu_us_art_faf mel-cepstral coefficients to msak0 EMA values would be the same as the mapping from msak0 mel-cepstral coefficients to msak0 EMA values. Similarly, the map-
Table 3.1: Single-speaker Baseline Mapping Results

<table>
<thead>
<tr>
<th>EMA Prediction Method</th>
<th>RMSE</th>
<th>Ave. Artic. RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive (0)</td>
<td>93.042</td>
<td>18.001</td>
</tr>
<tr>
<td>Naive (means)</td>
<td>9.931</td>
<td>2.404</td>
</tr>
<tr>
<td>Linear Regression (MGCEP)</td>
<td>8.674</td>
<td>2.088</td>
</tr>
<tr>
<td>Linear Regression (MGCEP+0th)</td>
<td>8.662</td>
<td>2.084</td>
</tr>
<tr>
<td>Linear Regression (MCEP)</td>
<td>8.549</td>
<td>2.067</td>
</tr>
<tr>
<td>CART (MGCEP)</td>
<td>8.103</td>
<td>1.967</td>
</tr>
<tr>
<td>CART (MGCEP+0th)</td>
<td>8.033</td>
<td>1.950</td>
</tr>
<tr>
<td>CART (MCEP)</td>
<td>7.999</td>
<td>1.945</td>
</tr>
<tr>
<td>MGCEP Prediction Method</td>
<td>RMSE</td>
<td>MCD mean and std</td>
</tr>
<tr>
<td>Naive (0)</td>
<td>0.307</td>
<td>1.769 ± 0.655</td>
</tr>
<tr>
<td>Naive (means)</td>
<td>0.192</td>
<td>1.139 ± 0.294</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.159</td>
<td>0.931 ± 0.300</td>
</tr>
<tr>
<td>CART</td>
<td>0.152</td>
<td>0.881 ± 0.310</td>
</tr>
<tr>
<td>MGCEP+0th Prediction Method</td>
<td>RMSE</td>
<td>MCD mean and std</td>
</tr>
<tr>
<td>Naive (0)</td>
<td>1.735</td>
<td>10.620 ± 0.861</td>
</tr>
<tr>
<td>Naive (means)</td>
<td>0.247</td>
<td>1.477 ± 0.353</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.188</td>
<td>1.083 ± 0.406</td>
</tr>
<tr>
<td>CART</td>
<td>0.179</td>
<td>1.012 ± 0.421</td>
</tr>
<tr>
<td>MCEP Prediction Method</td>
<td>RMSE</td>
<td>MCD mean and std</td>
</tr>
<tr>
<td>Naive (0)</td>
<td>2.300</td>
<td>13.572 ± 3.918</td>
</tr>
<tr>
<td>Naive (means)</td>
<td>1.269</td>
<td>7.042 ± 3.337</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>1.088</td>
<td>6.142 ± 2.631</td>
</tr>
<tr>
<td>CART</td>
<td>0.999</td>
<td>5.610 ± 2.485</td>
</tr>
<tr>
<td>MGCEP to EMA to MGCEP Prediction</td>
<td>RMSE</td>
<td>MCD mean and std</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.144</td>
<td>0.849 ± 0.235</td>
</tr>
<tr>
<td>CART</td>
<td>0.143</td>
<td>0.835 ± 0.267</td>
</tr>
<tr>
<td>MGCEP+0th to EMA to MGCEP+0th Prediction</td>
<td>RMSE</td>
<td>MCD mean and std</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.161</td>
<td>0.960 ± 0.248</td>
</tr>
<tr>
<td>CART</td>
<td>0.163</td>
<td>0.942 ± 0.342</td>
</tr>
<tr>
<td>MCEP to EMA to MCEP Prediction</td>
<td>RMSE</td>
<td>MCD mean and std</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.985</td>
<td>5.536 ± 2.442</td>
</tr>
<tr>
<td>CART</td>
<td>0.937</td>
<td>5.304 ± 2.231</td>
</tr>
</tbody>
</table>
Table 3.2: Single-speaker EMA RMSE Breakdowns

<table>
<thead>
<tr>
<th>Articulator</th>
<th>Naive Zero</th>
<th>Naive Means</th>
<th>Lin. Regress. MGCEP w/0th</th>
<th>MCEP</th>
<th>CART MGCEP w/0th</th>
<th>MCEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>lower incisor x</td>
<td>5.206</td>
<td>0.696</td>
<td>0.638</td>
<td>0.638</td>
<td>0.628</td>
<td>0.626</td>
</tr>
<tr>
<td>upper lip x</td>
<td>10.220</td>
<td>0.998</td>
<td>0.959</td>
<td>0.955</td>
<td>0.949</td>
<td>0.943</td>
</tr>
<tr>
<td>lower lip x</td>
<td>5.813</td>
<td>1.711</td>
<td>1.476</td>
<td>1.470</td>
<td>1.437</td>
<td>1.427</td>
</tr>
<tr>
<td>tongue tip x</td>
<td>18.218</td>
<td>3.779</td>
<td>3.276</td>
<td>3.262</td>
<td>3.413</td>
<td>3.012</td>
</tr>
<tr>
<td>tongue body x</td>
<td>33.217</td>
<td>3.272</td>
<td>2.980</td>
<td>2.973</td>
<td>3.090</td>
<td>2.764</td>
</tr>
<tr>
<td>tongue dorsum x</td>
<td>46.444</td>
<td>2.785</td>
<td>2.520</td>
<td>2.520</td>
<td>2.576</td>
<td>2.396</td>
</tr>
<tr>
<td>velum x</td>
<td>58.462</td>
<td>1.420</td>
<td>1.026</td>
<td>1.015</td>
<td>1.142</td>
<td>1.010</td>
</tr>
<tr>
<td>lower incisor y</td>
<td>24.762</td>
<td>1.429</td>
<td>1.229</td>
<td>1.229</td>
<td>1.170</td>
<td>1.149</td>
</tr>
<tr>
<td>upper lip y</td>
<td>1.992</td>
<td>1.570</td>
<td>1.277</td>
<td>1.273</td>
<td>1.300</td>
<td>1.261</td>
</tr>
<tr>
<td>lower lip y</td>
<td>27.172</td>
<td>3.208</td>
<td>2.513</td>
<td>2.513</td>
<td>2.445</td>
<td>2.338</td>
</tr>
<tr>
<td>tongue body y</td>
<td>4.405</td>
<td>3.530</td>
<td>3.017</td>
<td>3.017</td>
<td>2.783</td>
<td>2.769</td>
</tr>
<tr>
<td>tongue dorsum y</td>
<td>3.270</td>
<td>3.229</td>
<td>2.956</td>
<td>2.955</td>
<td>2.778</td>
<td>2.772</td>
</tr>
<tr>
<td>velum y</td>
<td>3.805</td>
<td>1.666</td>
<td>1.453</td>
<td>1.451</td>
<td>1.501</td>
<td>1.446</td>
</tr>
<tr>
<td>Average</td>
<td>18.001</td>
<td>2.404</td>
<td>2.088</td>
<td>2.084</td>
<td>2.067</td>
<td>1.967</td>
</tr>
</tbody>
</table>

Baseline Approach with Linear Regression

The baseline linear regression approach to mapping from cmu/us/art/faf mel-cepstral coefficients to msak0 EMA values was first attempted as follows:

1. Synthesize the msak0 utterances with the cmu/us/art/faf unit selection voice.
2. Extract the mel-cepstral coefficients from these utterances using SPTK with the previously mentioned parameters (and a 10ms frame advance rate).
3. Assume the identity mapping between cmu/us/art/faf mel-cepstral coefficients and msak0 mel-cepstral coefficients, and consider the results of the previous step to be msak0 mel-cepstral coefficients.
4. Apply the linear regression mapping from msak0 mel-cepstral coefficients to EMA values to these coefficients.

This approach was tried and the resulting predicted msak0 EMA values were compared with the measured msak0 EMA values using DTW with the Itakura rule [17] and a sum-squared-error metric. The mean of the sum-squared-errors (weighted by predicted utterance lengths) was taken, and the square root of that was taken to provide an error measure that was similar to RMS error. The result on the test set (msak0 test utterances) using mel-generalized cepstral coefficients was an RMS error of 10.248. The result on the test set using (non-generalized) mel-cepstral coefficients was an RMS error of 9.310.

The baseline linear regression approach to mapping from msak0 EMA values to cmu_us_art_faf mel-cepstral coefficients was first attempted as follows:

1. Take the EMA values from the msak0 test set.
2. Apply the linear regression mapping from msak0 EMA values to msak0 mel-cepstral coefficients.
3. Assume the identity mapping between msak0 mel-cepstral coefficients and cmu_us_art_faf mel-cepstral coefficients, and consider the results of the previous step to cmu_us_art_faf mel-cepstral coefficients.

The results of this approach were compared with the mel-generalized cepstral coefficients from the cmu_us_art_faf synthesized utterances using DTW with the Itakura rule [17] and a sum-squared-error metric. Again, the mean of the sum-squared-error (weighted by predicted utterance lengths) was taken, and the square root of that was taken to provide an error measure that was similar to RMS error. The result on the test set (msak0 test set utterances synthesized by cmu_us_art_faf) was an RMS error of 0.272. The result of trying this approach on (non-generalized) mel-cepstral coefficients was an RMS error of 1.599 and an average MCD of 9.431.

Baseline Approach with CART

The baseline CART approach to mapping from cmu_us_art_faf mel-cepstral coefficients to msak0 EMA values is the same as the baseline linear regression approach, except that the step which maps msak0 mel-cepstral to msak0 EMA values uses CART. Similarly, the baseline CART approach to mapping from msak0 EMA values to cmu_us_art_faf mel-cepstral values is the same as the baseline linear regression approach, except that the step which maps msak0 EMA values to msak0 mel-cepstral coefficients uses CART.

Cross-speaker Baseline Results

Interestingly, linear regression performed better than CART in all of the cross-speaker baseline trials, even though CART performed better than linear regression in almost all of the single-speaker baseline cases. Cross-speaker mapping results are summarized in Table 3.3 and Table 3.4.
Table 3.3: Cross-speaker Baseline Mapping Results

<table>
<thead>
<tr>
<th></th>
<th>FAF MGCEP</th>
<th>FAF MGCEP w/0th</th>
<th>FAF MCEP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FAF to MSAK0 EMA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Regression</td>
<td>2.527</td>
<td>2.507</td>
<td>2.299</td>
</tr>
<tr>
<td>CART</td>
<td>2.643</td>
<td>2.572</td>
<td>2.487</td>
</tr>
<tr>
<td><strong>MSAK0 EMA to FAF</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Regression</td>
<td>1.565 ± 0.577</td>
<td>1.858 ± 0.737</td>
<td>9.431 ± 2.732</td>
</tr>
<tr>
<td>CART</td>
<td>1.596 ± 0.585</td>
<td>1.890 ± 0.767</td>
<td>9.481 ± 2.777</td>
</tr>
<tr>
<td><strong>Round-Trip</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Regression</td>
<td>1.638 ± 0.595</td>
<td>1.767 ± 0.616</td>
<td>9.377 ± 2.436</td>
</tr>
<tr>
<td>CART</td>
<td>1.663 ± 0.672</td>
<td>1.823 ± 0.787</td>
<td>10.026 ± 2.462</td>
</tr>
</tbody>
</table>

Table 3.4: Cross-Speaker Baseline EMA RMSE Breakdowns

<table>
<thead>
<tr>
<th>Articulator</th>
<th>Lin. Regress.</th>
<th>CART</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MGCEP</td>
<td>w/0th</td>
</tr>
<tr>
<td>lower incisor x</td>
<td>0.735</td>
<td>0.728</td>
</tr>
<tr>
<td>upper lip x</td>
<td>1.059</td>
<td>1.054</td>
</tr>
<tr>
<td>lower lip x</td>
<td>1.745</td>
<td>1.684</td>
</tr>
<tr>
<td>tongue tip x</td>
<td>3.491</td>
<td>3.507</td>
</tr>
<tr>
<td>tongue dorsum x</td>
<td>3.242</td>
<td>3.253</td>
</tr>
<tr>
<td>velum x</td>
<td>1.917</td>
<td>1.813</td>
</tr>
<tr>
<td>lower incisor y</td>
<td>1.623</td>
<td>1.610</td>
</tr>
<tr>
<td>upper lip y</td>
<td>1.628</td>
<td>1.616</td>
</tr>
<tr>
<td>tongue tip y</td>
<td>4.171</td>
<td>4.170</td>
</tr>
<tr>
<td>tongue body y</td>
<td>3.540</td>
<td>3.518</td>
</tr>
<tr>
<td>tongue dorsum y</td>
<td>3.277</td>
<td>3.285</td>
</tr>
<tr>
<td>velum y</td>
<td>2.109</td>
<td>2.054</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>2.527</td>
<td>2.507</td>
</tr>
</tbody>
</table>
Figures 3.1, 3.2, and 3.3 contain graphs which show what plots of EMA values look like. The positions are (x,y) coordinates in the mid-sagittal plane, and the scale is thousandths of centimeters. The speaker is facing left. Figure 3.1 shows the actual msak0 test set EMA coordinates. Figure 3.2 shows the predicted msak0 EMA coordinates using the baseline linear regression mapping from faf mel-cepstral values. Figure 3.3 shows the predicted msak0 EMA coordinates using the baseline CART mapping from faf mel-cepstral values.

Looking at the graphs, a few trends can be observed. The actual msak0 test set EMA coordinates are more spread out than the EMA coordinates predicted from the cmu_us_art_faf mel-cepstral coordinates. In the actual test set data, the ranges of different articulators may overlap. In the baseline linear regression data, the ranges of the different articulators have no overlap (one could say they were articulated). In the baseline CART data, the ranges of some of the articulators overlap, but it is to a lesser extent than for the actual data.

3.4.3 Improving the Mappings between FAF MCEP and MSAK0 MCEP

The baseline results for the cross-speaker mapping were not very good when compared with the single-speaker baselines. One potential area for improvement is to reexamine the assumption that cmu_us_art_faf mel-cepstral coefficients can be treated equivalently to msak0 mel-cepstral coefficients. In the baseline cross-speaker mapping, the identity mapping was used to go between cmu_us_art_faf mel-cepstral coefficients and msak0 mel-cepstral coefficients. This is just one of a number of possibilities. Another approach would be to map between them based on z-scores. For example, given a source mel-cepstral coefficient, $m_s$, a target mel-cepstral coefficient, $m_t$, can be derived as follows:

$$m_t = \frac{(m_s - \overline{m_s})}{\sigma_{m_s}} \sigma_{m_t} + \overline{m_t}$$

where $\overline{m_s}$ is the mean of that particular coefficient over a training set for the source voice, $\sigma_{m_s}$ is the standard deviation of that coefficient over a training set for the source voice, $\overline{m_t}$ is the mean of that particular coefficient over a training set for the target voice, and $\sigma_{m_t}$ is the standard deviation of that particular coefficient over a training set for the target voice.

Using this type of mapping, another procedure for mapping from cmu_us_art_faf mel-cepstral coefficients to msak0 EMA values is:

1. Use the training set for msak0 and the training set for the cmu_us_art_faf synthesized utterances using msak0 training set text to collect mean and standard deviation statistics for the (1st through 24th) mel-cepstral coefficients.

2. Use the z-score map based on these statistics to map the cmu_us_art_faf mel-cepstral coefficients (from the synthesized msak0 test text) to msak0 mel-cepstral coefficients.
Figure 3.1: msak0 Test Set EMA
Figure 3.2: msak0 EMA predicted from FAF MGCEP (baseline linear regression)
Figure 3.3: msak0 EMA predicted from FAF MGCEP (baseline CART)
3. Use the previously learned msak0 maps to map the results of the previous step to msak0 EMA values.

4. Because the number of these predicted msak0 EMA values probably differs from the number of the EMA values for the corresponding msak0 utterance, use DTW with the Itakura rule \[17\] to align the results for evaluation.

Similarly, another procedure for mapping from msak0 EMA values to cmu us art faf mel-cepstral coefficients is:

1. Use the previously-learned msak0 maps to map the msak0 test set EMA values to msak0 mel-cepstral coefficients.

2. Use the training set for msak0 and the training set for the cmu us art faf synthesized utterances using msak0 training set text to collect mean and standard deviation statistics for the (1st through 24th) mel-cepstral coefficients.

3. Use the z-score map based on these statistics to map the predicted msak0 mel-cepstral coefficients to cmu us art faf mel-cepstral coefficients.

4. Because the number of these predicted cmu us art faf mel-cepstral coefficients may differ from the number of the coefficients for the corresponding cmu us art faf synthesized utterance, use DTW with the Itakura rule \[17\] to align the results for evaluation.

A “round-trip” map from cmu us art faf mel-cepstral coefficients to cmu us art faf mel-cepstral coefficients can be obtained by composing these two maps. In this case, the length of the predicted coefficients will match the length of the actual coefficients, and DTW will not be necessary during evaluation.

**Cross-speaker Z-score Mapping results**

The above mappings were tried, and the results were better than the baseline cross-speaker mappings. They also compared favorably to the single-speaker naive baselines, but were not quite as good as the best single-speaker maps using the same features. Interestingly, linear regression performed better than CART in all of the cross-speaker z-score mapping trials. The results appear in Table 3.5 and Table 3.6.

**3.4.4 Directly Mapping Between FAF MFCCs and MSAK0 EMA after DTW**

Another possible approach to mapping from cmu us art faf mel-cepstral coefficients to msak0 EMA values goes as follows:

- Use DTW to create utterances in cmu us art faf that match the lengths of the corresponding msak0 utterances.
Table 3.5: Cross-speaker Z-score Mapping Results

<table>
<thead>
<tr>
<th></th>
<th>FAF to MSAK0 EMA</th>
<th>MSAK0 EMA to FAF</th>
<th>Round-Trip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear Regression</td>
<td>CART</td>
<td>Linear Regression</td>
</tr>
<tr>
<td>Mean RMSE</td>
<td>2.162</td>
<td>2.134</td>
<td>2.232</td>
</tr>
<tr>
<td></td>
<td>2.232</td>
<td>2.233</td>
<td>2.207</td>
</tr>
<tr>
<td></td>
<td>1.289 ± 0.468</td>
<td>1.609 ± 0.693</td>
<td>7.632 ± 2.289</td>
</tr>
<tr>
<td></td>
<td>1.395 ± 0.433</td>
<td>1.723 ± 0.702</td>
<td>7.872 ± 2.395</td>
</tr>
<tr>
<td></td>
<td>1.245 ± 0.476</td>
<td>1.406 ± 0.530</td>
<td>7.267 ± 2.132</td>
</tr>
<tr>
<td></td>
<td>1.749 ± 0.644</td>
<td>1.980 ± 0.730</td>
<td>9.916 ± 2.432</td>
</tr>
</tbody>
</table>

Table 3.6: Cross-speaker EMA RMSE Z-score Result Breakdowns

<table>
<thead>
<tr>
<th>Articulator</th>
<th>Lin. Regress. MGCEP</th>
<th>w/0th</th>
<th>MCEP</th>
<th>CART MGCEP</th>
<th>w/0th</th>
<th>MCEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>lower incisor x</td>
<td>0.656</td>
<td>0.652</td>
<td>0.651</td>
<td>0.701</td>
<td>0.700</td>
<td>0.687</td>
</tr>
<tr>
<td>upper lip x</td>
<td>0.974</td>
<td>0.979</td>
<td>0.975</td>
<td>0.997</td>
<td>0.993</td>
<td>0.990</td>
</tr>
<tr>
<td>lower lip x</td>
<td>1.581</td>
<td>1.560</td>
<td>1.556</td>
<td>1.595</td>
<td>1.612</td>
<td>1.617</td>
</tr>
<tr>
<td>tongue tip x</td>
<td>3.130</td>
<td>3.118</td>
<td>3.158</td>
<td>3.265</td>
<td>3.195</td>
<td>3.167</td>
</tr>
<tr>
<td>tongue body x</td>
<td>2.716</td>
<td>2.722</td>
<td>2.763</td>
<td>2.855</td>
<td>2.822</td>
<td>2.825</td>
</tr>
<tr>
<td>tongue dorsum x</td>
<td>2.495</td>
<td>2.492</td>
<td>2.459</td>
<td>2.505</td>
<td>2.508</td>
<td>2.484</td>
</tr>
<tr>
<td>velum x</td>
<td>1.492</td>
<td>1.435</td>
<td>1.432</td>
<td>1.390</td>
<td>1.376</td>
<td>1.318</td>
</tr>
<tr>
<td>lower incisor y</td>
<td>1.355</td>
<td>1.342</td>
<td>1.305</td>
<td>1.457</td>
<td>1.429</td>
<td>1.387</td>
</tr>
<tr>
<td>upper lip y</td>
<td>1.546</td>
<td>1.548</td>
<td>1.581</td>
<td>1.568</td>
<td>1.590</td>
<td>1.611</td>
</tr>
<tr>
<td>lower lip y</td>
<td>3.004</td>
<td>2.961</td>
<td>2.941</td>
<td>3.140</td>
<td>3.230</td>
<td>3.051</td>
</tr>
<tr>
<td>tongue tip y</td>
<td>3.724</td>
<td>3.712</td>
<td>3.628</td>
<td>4.056</td>
<td>4.086</td>
<td>3.936</td>
</tr>
<tr>
<td>tongue body y</td>
<td>2.985</td>
<td>2.988</td>
<td>2.884</td>
<td>3.164</td>
<td>3.150</td>
<td>3.103</td>
</tr>
<tr>
<td>tongue dorsum y</td>
<td>2.892</td>
<td>2.900</td>
<td>2.884</td>
<td>2.954</td>
<td>2.977</td>
<td>3.085</td>
</tr>
<tr>
<td>velum y</td>
<td>1.711</td>
<td>1.672</td>
<td>1.663</td>
<td>1.600</td>
<td>1.599</td>
<td>1.637</td>
</tr>
<tr>
<td>Average</td>
<td>2.162</td>
<td>2.149</td>
<td>2.134</td>
<td>2.232</td>
<td>2.233</td>
<td>2.207</td>
</tr>
</tbody>
</table>
Table 3.7: Cross-speaker DTW Direct Mapping Results

<table>
<thead>
<tr>
<th></th>
<th>FAF MGCEP</th>
<th>FAF MGCEP w/0th</th>
<th>FAF MCEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAF to MSAK0 EMA</td>
<td>Mean RMSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Regression</td>
<td>2.310</td>
<td>2.308</td>
<td>2.261</td>
</tr>
<tr>
<td>CART</td>
<td>2.335</td>
<td>2.324</td>
<td>2.232</td>
</tr>
<tr>
<td>MSAK0 EMA to FAF</td>
<td>MCD mean/std</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Regression</td>
<td>1.221 ± 0.555</td>
<td>1.549 ± 0.763</td>
<td>7.902 ± 3.045</td>
</tr>
<tr>
<td>CART</td>
<td>1.260 ± 0.570</td>
<td>1.562 ± 0.813</td>
<td>7.892 ± 3.158</td>
</tr>
<tr>
<td>Round-Trip</td>
<td>MCD mean/std</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Regression</td>
<td>1.172 ± 0.496</td>
<td>1.445 ± 0.586</td>
<td>7.400 ± 2.545</td>
</tr>
<tr>
<td>CART</td>
<td>1.216 ± 0.524</td>
<td>1.451 ± 0.779</td>
<td>7.408 ± 2.689</td>
</tr>
</tbody>
</table>

- Use linear regression to construct maps directly from the mel-cepstral coefficients in these new cmu-us-art-faf utterances to the msak0 EMA values.

Similarly, another possible approach to mapping from msak0 EMA values to cmu-us-art-faf mel-cepstral coefficients is:

- Use DTW to create utterances in cmu-us-art-faf that match the lengths of the corresponding msak0 utterances.

- Use linear regression to construct maps directly from msak0 EMA values to the mel-cepstral coefficients in these new cmu-us-art-faf utterances.

Again, a “round-trip” from cmu-us-art-faf mel-cepstral values to cmu-us-art-faf mel-cepstral values can be constructed by composing the two maps. Also, the above mappings can be constructed using CART instead of linear regression.

Cross-speaker DTW Direct Mapping Results

Except for the one case where cmu-us-art-faf mel-cepstral coefficients were used to predict msak0 EMA values, linear regression performed better than CART in all of the cross-speaker DTW direct mapping trials.

Results from all these strategies to use DTW and create direct mappings between cmu-us-art-faf mel-cepstral coefficients and msak0 EMA values are in Table 3.7 and Table 3.8.

3.4.5 Discussion of Results

When looking at all of these numbers a couple basic questions to ask are:

1. Are the current prediction values good?
Table 3.8: Cross-speaker EMA RMSE DTW Direct Result Breakdowns

<table>
<thead>
<tr>
<th>Articulator</th>
<th>Lin. Regress.</th>
<th>CART</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MGCEP w/0th MCEP</td>
<td>MGCEP w/0th MCEP</td>
</tr>
<tr>
<td>lower incisor x</td>
<td>0.682 0.680 0.662</td>
<td>0.695 0.700 0.672</td>
</tr>
<tr>
<td>upper lip x</td>
<td>0.977 0.976 0.976</td>
<td>0.983 0.985 0.976</td>
</tr>
<tr>
<td>lower lip x</td>
<td>1.605 1.602 1.574</td>
<td>1.625 1.622 1.580</td>
</tr>
<tr>
<td>tongue dorsum x</td>
<td>2.728 2.727 2.670</td>
<td>2.801 2.780 2.641</td>
</tr>
<tr>
<td>velum x</td>
<td>1.235 1.230 1.297</td>
<td>1.216 1.177 1.202</td>
</tr>
<tr>
<td>lower incisor y</td>
<td>1.385 1.380 1.311</td>
<td>1.382 1.368 1.302</td>
</tr>
<tr>
<td>upper lip y</td>
<td>1.453 1.453 1.474</td>
<td>1.479 1.480 1.482</td>
</tr>
<tr>
<td>lower lip y</td>
<td>3.045 3.036 2.963</td>
<td>3.048 3.020 2.934</td>
</tr>
<tr>
<td>tongue dorsum y</td>
<td>3.185 3.186 3.121</td>
<td>3.239 3.235 3.144</td>
</tr>
<tr>
<td>velum y</td>
<td>1.575 1.571 1.611</td>
<td>1.590 1.578 1.572</td>
</tr>
<tr>
<td>Average</td>
<td>2.310 2.308 2.261</td>
<td>2.335 2.324 2.232</td>
</tr>
</tbody>
</table>

2. Can they be made better?

The answers to these questions are a bit subjective, because they depend on the metrics that are used to determine “goodness,” but they are still meaningful questions to ask. The following are possible guidelines for producing such metrics:

1. How well do EMA values match physical constraints? The motions of articulators are limited by anatomy. There may be limits to how far they can move in a sampling interval, and the positions of some articulators may prevent other articulators from appearing in certain positions.

2. What metrics are used in the literature? If similar experiments have already been conducted, it will be necessary to use the same metrics to compare results.

3. Is there a quantity that is correlated with the results of the final task? The goal of investigating many of these mappings is to use them in other models.

4. What statistical tests are available for goodness of fit and significance?

Considering the second guideline, the results for the single-speaker mappings can be compared to the results from [35] and [34]. The former reported numerous
results for mappings from msak0 EMA values to mel-cepstral coefficients. The baseline result, which only used EMA values for prediction, was a MCD mean of 5.59 and a MCD standard deviation of 2.23. For this metric, the lower the values, the better. Our best baseline results for this experiment used CART and had a MCD mean of 5.610 and a MCD standard deviation of 2.485. In this case the means only differ slightly in the third significant figure. The standard deviation is 11.4% greater for our baseline. So it appears that our mel-cepstral feature predictions are about as good on average, but they vary more, so they are slightly worse. It does appear that they can be improved, however. Predictions in [35] were improved by adding $F_0$ and power features for prediction. This led to results of a MCD mean of 4.59 and a MCD standard deviation of 1.61. Such features could also be added to our baseline linear regression and CART models to improve them. Some experiments on a subset of the msak0 data gave results that were consistent with this theory. (Of course, another option would be to try the GMM mapping technique used in [35].) As for the EMA value predictions, CART gave our best baseline RMSE average per articulator, which was 1.945. The best baseline results for this quantity in [34] was 1.63 before low-pass filtering was applied and 1.49 after low-pass filtering was applied. Our result was 19.3% and 30.5% worse than these results respectively. So our results for EMA value predictions were not as good. Again, it appears that they can be improved using ideas from the paper. These include: using the 0th mel-cepstral coefficient for prediction as well, using low-pass filtering of the end results, and using the GMM mapping technique.

For round-trip experiments, and cross-speaker experiments, there do not appear to be other examples in the literature, so other guidelines must be used to evaluate the results.

### 3.5 Using EMA to Improve Speech Synthesis Models

Although a number of experiments have been conducted which examine the extent to which articulatory positions can be predicted from speech signals, it still remains to determine whether articulatory position features (from both actual data and cross-speaker mappings) can be used to improve models used for speech synthesis. Although our goal is to improve Voice Transformation, it is still useful to investigate possible uses of articulatory position data in speech synthesis due to the overlap in goals and models between speech synthesis and voice transformation.

To investigate this area it is first necessary to consider what types of models are used for speech synthesis. The preliminary experiments will focus on the models that are used by a typical unit selection voice constructed with the Festival Speech Synthesis System [5]. Such models include duration models, $F_0$ models, power models, prominence models, unit selection, and joins. The preliminary experiments focused on duration, $F_0$, and power models.
3.5.1 Duration Models

The goal of a typical duration model in the Festival Speech Synthesis System is to provide times for the lengths of the segments (usually phones) in a synthesized utterance. There are numerous ways to model duration, varying in degrees of sophistication and accuracy. Approaches mentioned in the Festvox [4] documentation for creating synthetic voices with Festival include: always predicting the same duration, predicting phone durations based on averages collected from a training set, using a set of rules to modify these averages such as the Klatt model in [3], and training a CART.

Is it plausible to expect articulatory position data to aid in the prediction of segment durations? It is common to consider phones in terms of their phonetic features, including place of articulation, manner of articulation, nasalization and voicing. It seems reasonable to expect place of articulation to correspond with some articulatory position features and nasalization to correspond with velum position features, although other phonetic features, such as voicing, may not directly correspond with articulatory positions. So it appears that articulatory position data may help distinguish between some classes of phones, and could potentially be used to place boundaries between phones, which in turn could be used to determine lengths of phones. This approach, however, appears to be different from the previously mentioned duration models.

Can articulatory position data be used with the type of duration models used in the Festival Speech Synthesis System? There appears to be an inherent difficulty with using these models with articulatory position data. The articulatory position data we have is sampled periodically over time, without knowledge of which segment they belong to. This does not create a problem for the Festival-style models during training, because the segment durations are known, and it is possible to determine which articulatory positions belong to which segments. There is some question of what statistics concerning the variable-length list of articulatory positions per segment are important, but it would be possible to try a number of figures, such as averages, minima, and maxima and derive an empirical answer. The true difficulty arises, however, during testing. For the test set, these models would attempt to predict the durations of the segments and these values would not be known \textit{a priori}. These models would not have knowledge of which articulatory positions belong to which segments, and could not use them for predictions.

This does not necessarily mean that articulatory position data cannot be used for duration modeling. It just suggests that there may be no straightforward way to incorporate it into the duration models currently used in Festival. It may be possible to create a model which starts with an initial guess for the durations, and then modifies it iteratively, based on a probabilistic model. Such an approach would be quite different from the previously-mentioned models.
3.5.2 $F_0$ Models

The goal of a typical $F_0$ model in the Festival Speech Synthesis System is to predict values for the fundamental frequency ($F_0$) for the start, middle, and end of each syllable. The approaches mentioned in the Festvox documentation include: rule-based, linear regression, CART, and Tilt modeling.

Is it plausible to expect articulatory position data to aid in the prediction of $F_0$ values? Although the fundamental frequency of speech is most closely associated with the larynx, which is not an articulator, there may still be some correspondence with the positions of the articulators that are measured. Trying to predict $F_0$ values with articulatory positions appears to be a worthwhile area to investigate.

Can the current $F_0$ models be extended with articulatory position data? The baseline CART model [4] predicts $F_0$ targets for the start, middle and end of each syllable, based on the following syllabic features:

1. Number of phones in the syllable
2. Onset type
3. Position in word
4. A window of five syllable break indexes, centered on the current syllable
5. A window of five stresses, centered on the current syllable
6. Syllables to start and end of current phrase
7. Stressed syllables to start and end of current phrase
8. Sub-phrases.

The pda program distributed with Festival was used to extract $F_0$ values from the msak0 acoustic files. Then Festival utterances were created for the msak0 data. After the previously mentioned problematic utterances were removed, the traintest script from Festival was used to construct data sets for training and testing the baseline CART model for $F_0$ prediction. Trees were built in a stepwise manner using a stop value of 15.

Then, another set of utterances was created which included articulatory position values at every 10ms. In addition to the baseline syllabic features, the following features were also extracted:

1. All 14 articulator coordinates for the first frame in the syllable
2. All 14 articulator coordinates for the last frame in the syllable.

The results are summarized in Table 3.9.
### Table 3.9: msak0 $F_0$ Prediction Results

<table>
<thead>
<tr>
<th>Scenario</th>
<th>RMSE</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start (Baseline)</td>
<td>11.5216</td>
<td>0.5199</td>
</tr>
<tr>
<td>Middle (Baseline)</td>
<td>17.0915</td>
<td>0.5622</td>
</tr>
<tr>
<td>End (Baseline)</td>
<td>14.0939</td>
<td>0.8731</td>
</tr>
<tr>
<td>Start (w/EMA)</td>
<td>11.5216</td>
<td>0.5199</td>
</tr>
<tr>
<td>Middle (w/EMA)</td>
<td>16.2363</td>
<td>0.6181</td>
</tr>
<tr>
<td>End (w/EMA)</td>
<td>13.8934</td>
<td>0.8781</td>
</tr>
<tr>
<td>Start (+ mid)</td>
<td>11.8573</td>
<td>0.4926</td>
</tr>
<tr>
<td>Middle (+ mid)</td>
<td>16.2363</td>
<td>0.6181</td>
</tr>
<tr>
<td>End (+ mid)</td>
<td>13.8052</td>
<td>0.8791</td>
</tr>
<tr>
<td>Start (+ dist)</td>
<td>11.8573</td>
<td>0.4926</td>
</tr>
<tr>
<td>Middle (+ dist)</td>
<td>16.2555</td>
<td>0.6171</td>
</tr>
<tr>
<td>End (+ dist)</td>
<td>13.6540</td>
<td>0.8822</td>
</tr>
<tr>
<td>Start (EMA+avgs)</td>
<td>13.4140</td>
<td>0.2025</td>
</tr>
<tr>
<td>Middle (EMA+avgs)</td>
<td>20.5901</td>
<td>0.2282</td>
</tr>
<tr>
<td>End (EMA+avgs)</td>
<td>27.5300</td>
<td>0.3571</td>
</tr>
<tr>
<td>Start (EMA only)</td>
<td>13.4140</td>
<td>0.2025</td>
</tr>
<tr>
<td>Middle (EMA only)</td>
<td>19.5805</td>
<td>0.3257</td>
</tr>
<tr>
<td>End (EMA only)</td>
<td>27.5300</td>
<td>0.3571</td>
</tr>
<tr>
<td>Start (LR EMA+avgs)</td>
<td>12.1266</td>
<td>0.4243</td>
</tr>
<tr>
<td>Middle (LR EMA+avgs)</td>
<td>18.9360</td>
<td>0.4046</td>
</tr>
<tr>
<td>End (LR EMA+avgs)</td>
<td>25.7304</td>
<td>0.4546</td>
</tr>
<tr>
<td>Start (LR EMA only)</td>
<td>12.1249</td>
<td>0.4241</td>
</tr>
<tr>
<td>Middle (LR EMA only)</td>
<td>19.4663</td>
<td>0.3359</td>
</tr>
<tr>
<td>End (LR EMA only)</td>
<td>25.9795</td>
<td>0.4379</td>
</tr>
<tr>
<td>Start (+EMA sme+dist+smdiffs)</td>
<td>11.8573</td>
<td>0.4926</td>
</tr>
<tr>
<td>Middle (+EMA sme+dist+smdiffs)</td>
<td>17.1855</td>
<td>0.5562</td>
</tr>
<tr>
<td>End (+EMA sme+dist+smdiffs)</td>
<td>12.4744</td>
<td>0.9025</td>
</tr>
</tbody>
</table>
### Cross-Speaker $F_0$ Results

One question that was raised was whether cross-speaker articulatory position predictions would be helpful for predicting $F_0$. Experiments using cross-speaker mappings from fsew0 to msak0 and from cmu to faf were performed. The procedure was as follows:

1. Get utterances (direct from fsew0 or synthesized from faf)
2. Extract MFCCs with SPTK toolkit
3. Z-score map them to the msak0 MFCC ranges
4. Apply a previously learned msak0 MFCC to msak0 EMA map (linear regression)
5. Incorporate these predicted EMA values into Festival utterances
6. Generate features for syllable start, middle, and end, distances, differences, previous and next segment
7. Apply a previously learned msak0 EMA to msak0 $F_0$ map (linear regression)

The results of these experiments in terms of correlations are presented in Table 3.10. Correlations for $F_0$ values at the starts and ends of the syllables suggest there is some predictive power in these cross-speaker articulatory position predictions. However, the $F_0$ values in the middles of the syllables are very strongly uncorrelated.

### 3.5.3 Power Model

Another possible type of prosodic model is one which predicts the power of each segment (phone) in an utterance. In the following experiments, the power was derived by computing the energy from the signal with the sig2fv program from the Edinburgh Speech Tools [5] and then averaging over the length of the segment.

Is it plausible to expect articulatory position data to aid in the prediction of power values? Although the articulators are not the base cause of the strength of

<table>
<thead>
<tr>
<th>Syl Positions</th>
<th>msak0</th>
<th>fsew0</th>
<th>fsew0 cross</th>
<th>faf cross</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>0.659</td>
<td>0.626</td>
<td>0.479</td>
<td>0.479</td>
</tr>
<tr>
<td>Middle</td>
<td>0.481</td>
<td>0.257</td>
<td>0.066</td>
<td>0.022</td>
</tr>
<tr>
<td>End</td>
<td>0.425</td>
<td>0.463</td>
<td>0.283</td>
<td>0.389</td>
</tr>
</tbody>
</table>

Table 3.10: Cross-Speaker $F_0$ Prediction Results
Table 3.11: msak0 Power Prediction Results

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline CART</td>
<td>0.8689</td>
<td>0.5165</td>
</tr>
<tr>
<td>+ EMA Begin/End</td>
<td>0.8689</td>
<td>0.5165</td>
</tr>
<tr>
<td>+ EMA Mid/Dist.</td>
<td>0.8863</td>
<td>0.4978</td>
</tr>
</tbody>
</table>

the signal, it does not appear unreasonable to expect there to be some correlation between power and articulator movement. Emphasis could be an underlying cause that might simultaneously affect both articulator motions and power.

The results of some experiment that added articulatory position information to power models are in Table 3.11. It can be seen that for these trials, adding the articulatory position data did not help. Although loud, emphatic speech may contain a combination of greater power and larger jaw movements (leading to a wider range of lower incisor measurements), the MOCHA data did not vary enough in these regards to show a relationship.

### 3.5.4 Phonetic Feature Mapping

After finding it very difficult to use articulatory position data to predict various prosodic quantities, such as duration, $F_0$ and power, we searched for other useful speech parameters that would be more closely related to articulatory positions. One promising type was acoustic-phonetic features. The work mentioned below on mappings from articulatory positions to phonetic features is from a paper that was presented at the Interspeech2005 conference [37].

There have been some lines of work concerned with what have traditionally been called “acoustic-phonetic” features [25], but are occasionally referred to as “articulatory” features [23] [14] [38]. These features are categorial and describe phones when taken together. Some examples include voicing and placement of articulation. To minimize confusion, we will refer to such features as “phonetic” features in this proposal. Recent work on phonetic features has included an attempt to go beyond the “beads-on-a-string” approach to modeling speech [24] to models based on parallel streams of phonetic features. Such an approach has been demonstrated to improve speech recognition [23].

As many of the traditional phonetic features are related to notions of placement in the vocal tract, it seems natural to consider the connection between them and actual positions of articulators as measured by an EMA.

There are number of ways to represent phonetic features. One strategy is to use a set of multi-valued features, such as the manner, place, voicing, rounding, front-back, and static features described in [14]. A potential complication of this approach is that such multi-valued phonetic features are typically conceived of in a hierarchical manner. For instance, some features such as high and low are typically considered only for vowels, while other features such as labial and velar are typically considered only for consonants. In [14], all of these values are
possible for the place feature. The model used in [14] approaches this problem by conditioning the place value on the manner value, which can be vowel, silence, or one of a number of consonant types. Without some sort of hierarchy, though, place values associated with vowels may be confusable with place values for consonants. This may degrade performance.

Another strategy is to use a set of binary features that are either present or absent as in [23]. In this approach, a hierarchy of features is not necessary, but one potential complication is that many more features are needed to describe the phone set, and the cross-product of the values can be quite large. Based on how the features are used, however, this may not be a problem. This was the approach that we used.

Stepwise CART was used to construct models for predicting 18 binary phonetic features from the articulatory positions. This was fewer than the full set of 76 binary features used in [23] but sufficient for the purpose of demonstrating a relationship between the phonetic and articulatory position features.

Stepwise CART was chosen as a model because it can ignore predictor features when it does not find a high correlation with the predictee. This was considered important because it is believed that the positions of some articulators may be irrelevant to the values of some phonetic features. The stop-size for the trees was determined by cross-validation. For each speaker, 8/10 of the utterances were used for training, with an additional 1/10 used as a held-out set for the stepwise processing. The remaining 1/10 were used for testing. A few utterances were not used due to corrupt data. During training, as suggested by [23], only the center frames of the phones were used in order to minimize the effects of co-articulation. The centers of the phones were derived by automatically labeling the boundaries with SphinxTrain [7]. The center of each phone was labeled with the phone’s canonical phonetic features.

Other work [23] has used MFCCs to predict binary phonetic features. This work used different corpora that weren’t “phonetically balanced” like the MOCHA data and only provided overall accuracies for the phonetic feature recognizers, so the results cannot be compared. However, this work does raise the question that MFCCs may have predictive value for phonetic features. As this appears to be the case, we decided to also conduct experiments that would predict phonetic features from MFCCs and from a combination of articulatory positions and MFCCs. Because MFCCs are readily derived from the speech signal, using articulatory positions to predict phonetic features would only be useful in cases where the performance was improved or the speech signal was not available. The trials in the following experiments used the 0th through 24th MFCCs.

The results of the trials for the msk0 utterances from the MOCHA database are listed in Table 3.12. The listed results are f-scores that were derived by combining precision and recall. An alpha value of 0.5 was used to equally weight them.

For the 18 features that were tried, 5 were better predicted from MFCCs (unvoiced, vowel, high vowel, mid vowel, low vowel), 6 were better predicted from articulatory positions (lateral, labial, palatal, velar, back vowel, diphthong), and 6 were better predicted from a combination of the two (stop, nasal, fricative,
Table 3.12: *msak0 Binary Phonetic Feature Prediction*

<table>
<thead>
<tr>
<th>Feature</th>
<th>MFCC</th>
<th>EMA</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>unvoiced</td>
<td>0.683</td>
<td>0.203</td>
<td>0.683</td>
</tr>
<tr>
<td>stop</td>
<td>0.386</td>
<td>0.254</td>
<td>0.573</td>
</tr>
<tr>
<td>vowel</td>
<td>0.511</td>
<td>0.407</td>
<td>0.511</td>
</tr>
<tr>
<td>lateral</td>
<td>0.028</td>
<td>0.136</td>
<td>0.136</td>
</tr>
<tr>
<td>nasal</td>
<td>0.280</td>
<td>0.234</td>
<td>0.287</td>
</tr>
<tr>
<td>fricative</td>
<td>0.447</td>
<td>0.507</td>
<td>0.515</td>
</tr>
<tr>
<td>labial</td>
<td>0.175</td>
<td>0.457</td>
<td>0.457</td>
</tr>
<tr>
<td>palatal</td>
<td>0.037</td>
<td>0.368</td>
<td>0.037</td>
</tr>
<tr>
<td>velar</td>
<td>0.088</td>
<td>0.550</td>
<td>0.408</td>
</tr>
<tr>
<td>glottal</td>
<td>undef.</td>
<td>undef.</td>
<td>undef.</td>
</tr>
<tr>
<td>high vow.</td>
<td>0.270</td>
<td>0.132</td>
<td>0.132</td>
</tr>
<tr>
<td>mid vow.</td>
<td>0.205</td>
<td>0.197</td>
<td>0.205</td>
</tr>
<tr>
<td>low vow.</td>
<td>0.333</td>
<td>0.201</td>
<td>0.239</td>
</tr>
<tr>
<td>front vow.</td>
<td>0.198</td>
<td>0.184</td>
<td>0.406</td>
</tr>
<tr>
<td>back vow.</td>
<td>0.062</td>
<td>0.141</td>
<td>0.062</td>
</tr>
<tr>
<td>diphthong</td>
<td>0.072</td>
<td>0.182</td>
<td>0.072</td>
</tr>
<tr>
<td>round</td>
<td>0.154</td>
<td>0.139</td>
<td>0.256</td>
</tr>
<tr>
<td>alv. fric.</td>
<td>0.586</td>
<td>0.338</td>
<td>0.601</td>
</tr>
</tbody>
</table>

The experimental results demonstrated that the prediction of some phonetic features was indeed improved by using articulatory positions as predictors. Most of the features that were better predicted by articulatory positions were related to placement, which was expected. The MFCCs were much better at predicting whether a phone was unvoiced. This is not surprising because voicing is controlled by the larynx, which was not treated as an articulator in these experiments.

**Cross-Speaker Phonetic Feature Prediction**

Phonetic feature prediction is one possible candidate for measuring the usefulness of cross-speaker articulatory position prediction because articulatory positions have been demonstrated to be useful for predicting some phonetic features in the single-speaker case.

We investigated this possibility by conducting experiments using the fsew0 and FAF data to predict msak0 articulatory positions, which were then used to predict phonetic features.
Table 3.13: fsew0 Binary Phonetic Feature Prediction

<table>
<thead>
<tr>
<th>Feature</th>
<th>MFCC</th>
<th>EMA</th>
<th>Both</th>
<th>pEMA</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>unvoiced</td>
<td>0.645</td>
<td>0.356</td>
<td>0.598</td>
<td>0.318</td>
<td>0.681</td>
</tr>
<tr>
<td>stop</td>
<td>0.580</td>
<td>0.198</td>
<td>0.569</td>
<td>0.183</td>
<td>0.550</td>
</tr>
<tr>
<td>vowel</td>
<td>0.603</td>
<td>0.519</td>
<td>0.653</td>
<td>0.428</td>
<td>0.603</td>
</tr>
<tr>
<td>lateral</td>
<td>0.060</td>
<td>0.067</td>
<td>0.060</td>
<td>undef.</td>
<td>0.040</td>
</tr>
<tr>
<td>nasal</td>
<td>0.088</td>
<td>0.209</td>
<td>0.481</td>
<td>0.099</td>
<td>0.400</td>
</tr>
<tr>
<td>fricative</td>
<td>0.562</td>
<td>0.466</td>
<td>0.539</td>
<td>0.217</td>
<td>0.496</td>
</tr>
<tr>
<td>labial</td>
<td>0.052</td>
<td>0.436</td>
<td>0.429</td>
<td>0.097</td>
<td>0.053</td>
</tr>
<tr>
<td>palatal</td>
<td>0.429</td>
<td>0.145</td>
<td>0.595</td>
<td>0.047</td>
<td>0.086</td>
</tr>
<tr>
<td>velar</td>
<td>0.136</td>
<td>0.328</td>
<td>0.460</td>
<td>0.016</td>
<td>0.042</td>
</tr>
<tr>
<td>glottal</td>
<td>0.067</td>
<td>undef.</td>
<td>undef.</td>
<td>undef.</td>
<td>undef.</td>
</tr>
<tr>
<td>high vow.</td>
<td>0.383</td>
<td>0.254</td>
<td>0.339</td>
<td>0.102</td>
<td>0.383</td>
</tr>
<tr>
<td>mid vow.</td>
<td>0.273</td>
<td>0.194</td>
<td>0.273</td>
<td>0.197</td>
<td>0.273</td>
</tr>
<tr>
<td>low vow.</td>
<td>0.298</td>
<td>0.377</td>
<td>0.298</td>
<td>0.262</td>
<td>0.411</td>
</tr>
<tr>
<td>front vow.</td>
<td>0.379</td>
<td>0.446</td>
<td>0.451</td>
<td>0.130</td>
<td>0.310</td>
</tr>
<tr>
<td>back vow.</td>
<td>0.206</td>
<td>0.082</td>
<td>0.206</td>
<td>0.130</td>
<td>0.206</td>
</tr>
<tr>
<td>diphthong</td>
<td>0.047</td>
<td>0.163</td>
<td>0.047</td>
<td>0.081</td>
<td>0.045</td>
</tr>
<tr>
<td>round</td>
<td>0.086</td>
<td>0.052</td>
<td>0.086</td>
<td>0.058</td>
<td>0.027</td>
</tr>
<tr>
<td>alv. fric.</td>
<td>0.705</td>
<td>0.514</td>
<td>0.680</td>
<td>0.269</td>
<td>0.705</td>
</tr>
</tbody>
</table>

**fsew0 Phonetic Feature Prediction**

For the first cross-speaker phonetic feature prediction experiments, msak0 articulatory positions were predicted from the fsew0 MFCCs using the z-score mapping technique that was previously described. These articulatory positions were then used to learn decision trees that predicted phonetic features. The results are compared to prediction of fsew0 phonetic features based on actual fsew0 articulatory positions in Table 3.13. The results listed in the EMA column were for predictions from actual fsew0 articulatory positions from the 7 EMA (x,y)-coordinate pairs. The results listed in the pEMA column were predicted from articulatory position predictions for the msak0 speaker based on the fsew0 MFCCs using the z-score mapping cross-speaker approach. Again, the reported results are f-scores based on precision and recall, using an alpha value of 0.5.

For the cases using actual fsew0 articulatory positions, four phonetic features were best predicted by articulatory position data alone (lateral, labial, low vowel, diphthong). This was similar to the msak0 trials in Table 3.12 where lateral, labial, and diphthong were also best predicted by articulatory position data alone. Although palatal and velar were best predicted by articulatory positions alone for msak0, they were best predicted by the combination of articulatory positions and MFCCs for fsew0. Of the phonetic features best predicted by articulatory position alone for msak0, only back vowel was best predicted by MFCCs alone for fsew0. However, actual articulatory data predicted low vowel
better than MFCCs for fsaw0, which was not the case for msak0.

Considering the cases that used cross-speaker articulatory position predictions, labial, diphthong and round were the only cases where only using cross-speaker predicted articulatory positions was not improved by adding actual fsaw0 MFCCs. The combination of cross-speaker predicted articulatory positions and actual MFCCs gave the best results overall for unvoiced and low vowel. In the cases of high vowel, mid vowel, back vowel, and alveolar fricative, this combination tied the best performance, but that was because the MFCCs were responsible for that performance, and the cross-speaker articulatory positions were allowed to be ignored in the CART framework.

Table 3.14: FAF Binary Phonetic Feature Prediction

<table>
<thead>
<tr>
<th>Feature</th>
<th>MFCC</th>
<th>pEMA</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>unvoiced</td>
<td>0.291</td>
<td>0.237</td>
<td>0.237</td>
</tr>
<tr>
<td>stop</td>
<td>0.179</td>
<td>0.180</td>
<td>0.180</td>
</tr>
<tr>
<td>vowel</td>
<td>0.431</td>
<td>0.421</td>
<td>0.431</td>
</tr>
<tr>
<td>lateral</td>
<td>0.051</td>
<td>0.022</td>
<td>0.022</td>
</tr>
<tr>
<td>nasal</td>
<td>0.124</td>
<td>0.082</td>
<td>0.082</td>
</tr>
<tr>
<td>fricative</td>
<td>0.186</td>
<td>0.116</td>
<td>0.116</td>
</tr>
<tr>
<td>labial</td>
<td>0.083</td>
<td>0.125</td>
<td>0.125</td>
</tr>
<tr>
<td>palatal</td>
<td>0.109</td>
<td>0.125</td>
<td>0.125</td>
</tr>
<tr>
<td>velar</td>
<td>0.113</td>
<td>0.051</td>
<td>0.113</td>
</tr>
<tr>
<td>glottal</td>
<td>0.133</td>
<td>0.111</td>
<td>0.111</td>
</tr>
<tr>
<td>high vow.</td>
<td>0.110</td>
<td>0.130</td>
<td>0.130</td>
</tr>
<tr>
<td>mid vow.</td>
<td>0.240</td>
<td>0.247</td>
<td>0.247</td>
</tr>
<tr>
<td>low vow.</td>
<td>0.168</td>
<td>0.044</td>
<td>0.044</td>
</tr>
<tr>
<td>front vow.</td>
<td>0.124</td>
<td>0.112</td>
<td>0.112</td>
</tr>
<tr>
<td>back vow.</td>
<td>0.161</td>
<td>0.099</td>
<td>0.099</td>
</tr>
<tr>
<td>diphthong</td>
<td>0.123</td>
<td>0.079</td>
<td>0.079</td>
</tr>
<tr>
<td>round</td>
<td>0.135</td>
<td>0.044</td>
<td>0.044</td>
</tr>
<tr>
<td>alv. fric.</td>
<td>0.096</td>
<td>0.045</td>
<td>0.045</td>
</tr>
</tbody>
</table>

For the next round of cross-speaker phonetic feature experiments, msak0 articulatory positions were predicted from the FAF MFCCs using the z-score mapping technique. Again, these were then used to learn decision trees that predicted phonetic features. The f-score results are listed in Table 3.14. These experiments differed from the fsaw0 experiments in that no actual articulatory position data was available for the FAF utterances. Thus the results listed in the pEMA and Both columns used cross-speaker articulatory position predictions. The cross-speaker articulatory features were better at predicting stop, labial, palatal, high vowel, and mid vowel, and the MFCCs were better at predicting the remaining features. For FAF, there weren’t any cases where the combination outperformed the individual feature sets.

Overall, it appears that articulatory position data can be used to improve
the prediction of phonetic features. For one speaker (msak0), the addition of articulatory position data improved the recognition of 12 out of 18 phonetic features. For another speaker (fsew0), its addition improved the recognition of 9 out of 18 phonetic features. There is a considerable degree of overlap between the phonetic features that were best predicted for both speakers by adding articulatory position data.

For one speaker (fsew0), adding cross-speaker articulatory positions gave the best results for 2 out of 18 phonetic features. For another speaker (FAF), adding cross-speaker articulatory features gave the best results for 5 out of 18 phonetic features. These results are a bit less consistent than the results from using actual articulatory position data but show some promise.

3.6 Producing Speech from Articulatory Positions

3.6.1 Synthesis using Articulatory Positions

As mentioned in the Related Work chapter, articulatory position data has been used by at least two groups to predict spectral features, which were in turn used in combination with source features to synthesize speech [35] [29].

We performed experiments similar to the ones in [35] to investigate this area and also to serve as a basis for later experiments involving Voice Transformation. It is hoped that this type of synthesis will provide the ability to modify synthesized speech through modifications of the articulatory parameters, but this is not one of the goals of this proposal.

In our experiments, mappings were learned from the 14 articulatory position coordinates to the 1st through 24th MFCCs from the SPTK toolkit. We used the Gaussian Mixture Mapping technique described in [35] to learn mappings for both the msak0 and fsew0 speakers from the MOCHA database. In each case, 9/10 of the utterances were used as a training set. The articulatory positions and MFCCs were extracted at 10ms intervals, using 25ms windows for the MFCCs. The values for each frame were treated as components of 38-dimensional vectors, and the resulting vectors were used to train Gaussian Mixture Models with 64 Gaussians. Mapping from articulatory positions to MFCCs is performed by fixing the articulatory position components and using the EM algorithm [9] to provide a maximum likelihood estimate for the remaining MFCC components based on the learned Gaussian Mixture Model.

The preliminary experiments did not use additional features such as $F_0$, power, and so-called "dynamic features", which were shown to improve the mapping quality in terms of Mel-Cepstral Distortion.

Once the 1st through 24th MFCCs were predicted from an utterance, they were combined with actual 0th MFCCs and $F_0$ estimates to synthesize speech using the synthesis program in the Voice Transformation tools that are being added to the Festvox distribution [4]. This program uses pulse or noise excitation
with the Mel Log Spectral Approximation (MLSA) filter from [16] to synthesize speech.

It was found that intelligible, though somewhat degraded, speech could be produced using this method.

### 3.6.2 Voice Transformation using Articulatory Positions

Preliminary experiments were conducted using articulatory positions for Voice Transformation. The basic strategy was to modify the MFCC-based Voice Transformation methods used in the tools being added to the Festvox distribution [4] to use articulatory positions instead of MFCCs. The resulting articulatory position estimates for the target speaker were then used to synthesize speech as mentioned above by using them to predict the 1st through 24th MFCCs and then combining them with the $F_0$ estimates and the actual 0th MFCC.

The resulting speech sounded roughly similar to the original speech, but the individual words were mostly unintelligible.

### 3.7 Evaluation

Because we feel the evaluation of Voice Transformation still has numerous open questions and want to explore them, we conducted a listening experiment involving Voice Transformation. One issue with Voice Transformation is that familiarity with the speakers may influence the listeners’ perceptions of identity. We created the following experiment to examine this.

Two groups of people, called Group A and Group B, were selected for a Voice Transformation listening experiment based on the following criteria:

- Each group had 1 pair of male speakers and 1 pair of female speakers.
- When selecting speakers, priority was given to speakers with similar voices based on our subjective opinions.
- The listeners in each group knew the speakers in their group and did not know the speakers in the other group.

For Group A, the female speakers were clb and slt, and the male speakers were ehn and ref. For Group B, the female speakers were hb and jm, and the male speakers were mo and rf. Each speaker was recorded reading the first 30 sentences of the CMU ARCTIC corpus [20]. Then Voice Transformation models were trained in both directions for each of the speaker pairs (1 male pair and 1 female pair for both groups for a total of 4 pairs). Voice Transformation was performed by scaling pitch estimates, using a Gaussian Mixture Model mapping to transform mel-cepstral coefficients, and using a MLSA filter [16] for synthesis as described in [36].

For each speaker pair, a pair comparison evaluation with 10 trials was constructed. The utterances in each pair had different text to avoid confusion from
the unmodified portions of source speaker prosody, such as power, that were carried over to the transformed speech. Some trials consisted of recordings from different speakers, some consisted of transformed speech in different directions between the speakers, and some consisted of a recording and transformed speech. The original recordings were analyzed and resynthesized using the same MLSA filter technique [16] employed by the Voice Transformation process, in order to minimize differences perceived from artifacts due to the vocoding process used during transformation. Listeners were asked to rate the similarity of the speakers in each trial on a scale from 1 to 5, where 1 meant the speakers were very similar and 5 meant the speakers were very different. How the listeners were to judge speaker similarity and difference was left to them. In total, 10 listeners (5 from each group) listened to 40 utterance pairs (10 utterance pairs for each of 4 speaker pairs). With this setup we were able to collect data to investigate whether knowing the speakers made a difference in the judgment of speaker recognizability for Voice Transformation.

One thing we wanted to know immediately was whether the Voice Transformation was “successful.” One measure of this was whether the transformed speech was consistently judged as being more similar to the target speaker than the source speaker. This, indeed, was the case when considering the average similarity scores for each speaker pair across all listeners. These averages are shown in Figure 3.4, where “s1” stands for the first speaker in each pair, “s2” stands for the second speaker in each pair, “s1→s2” stands for transformed speech with the first speaker as the source and the second speaker as the target, and “s2→s1” stands for transformed speech with the second speaker as the source and the first speaker as the target. The scores comparing the target speakers with the transformed speech (s2,s1→s2 and s1,s2→s1) were lower, and thus more similar, than the scores comparing the source speakers with the transformed speech (s1,s2→s1 and s2,s2→s1).

Looking at the bars in Figure 3.4, a few more trends become apparent. Moving from the leftmost group of bars to the rightmost group, the bars for each speaker pair tend to get higher, showing greater differences in the compared speech. It appears that as the speakers are themselves judged further apart, the transformed speech is also judged as being further from the speakers. The Group A male speakers stand out as having the only exceptions to this general rule. Interestingly, there is a strong asymmetry with the Group A speakers. The bar comparing the transformation s2→s1 with its target speaker, s1, is much shorter than the bar comparing the transformation s1→s2 with its target speaker, s2. This suggests that the transformation from speaker s2 to s1 was much more successful than the transformation from speaker s1 to s2.

The next question was whether knowing the speakers made a difference. A breakdown of the results according to whether the listeners knew the speakers is given in Figure 3.5. Not only did the same general trend appear, where the transformed speech was judged as being more similar to the target speech than the source speech, but the scores for each type of compared speech were very close regardless of whether the listeners knew the speakers.
Figure 3.4: Similarities by Speaker Pair

Figure 3.5: Similarities by Knowledge of Speaker
3.7.1 Transformation Triangle Diagrams

As we looked at numerous graphs similar to the ones in Figure 3.4 and Figure 3.5, we realized that we wanted a better way to summarize multiple bars in the graphs and show how their values were related to each other. This led us to create Transformation Triangle Diagrams (TTDs) for each speaker pair. Some examples of these are in Figure 3.6, Figure 3.7, Figure 3.8, and Figure 3.9. TTDs can be interpreted as follows:

- The numbers in the diagrams are calculated by subtracting 1 from the similarity scores to compute 0-based similarity “distances” where 0 is most similar and 4 is most different.

- The distance between speech from the two speakers in a pair is represented by a horizontal line, with the names of the speakers listed at either end.

- Each diagram is composed of two directed triangles. The upper triangle represents comparisons made using the left speaker in the TTD as the source for Voice Transformation and the right speaker as the target. The lower triangle represents comparisons made using the right speaker as the source for Voice Transformation and the left speaker as the target. The arrows serve as reminders for the directions of the transformations.

- The vertices that are off the horizontal baseline represent transformed speech, and the remaining triangle edges represent the distances from the speakers’ speech to the transformed speech. For example, in the first TTD in Figure 3.6, the distance between speaker a1 and speech transformed from a1 to a2 is 1.9, the distance between speech transformed from a1 to a2 and speaker a2 is 0.7, the distance between speaker a2 and speech transformed from a2 to a1 is 2.1, and the distance between speech transformed from a2 to a1 and speaker a1 is 0.5.

- It should be noted that TTDs make no attempt to compare transformed speech using one speaker as the source with transformed speech using the other speaker as the source.

A few examples of TTDs are given in Figure 3.6, Figure 3.7, Figure 3.8, and Figure 3.9. Figure 3.6 represents a pair of speakers called a1 and a2, where both transformations were mostly successful in that the transformed speech was considerably closer to the targets than the sources in both cases.

Figure 3.7 represents a pair of speakers called b1 and b2, where both transformations were fairly unsuccessful in that the transformed speech was closer to the source than the target. As transformation becomes more successful, the
TTDs tend to skew so the upper triangle is crushed to the right and the lower triangle is crushed to the left.

However, distance from a vertex representing transformed speech to the horizontal baseline can make a difference as well. In Figure 3.8 representing speakers c1 and c2 and in Figure 3.9 representing speakers d1 and d2, the vertices representing the transformed speech would project to the same location on the horizontal baselines, but the transformations between c1 and c2 were more successful than the ones between d1 and d2 because the transformed speech is closer to the targets. One additional point is that the length of the horizontal baselines vary according to the similarity of the speakers. The more similar the speakers are, the narrower the baseline is.

In the ideal case, both transformations would coincide with their targets, and the TTD would collapse to a horizontal line with arrowheads pointing outward.
at each end. In a case where the transformation was completely unsuccessful and the transformed speech sounded like the source voice, the TTD would again collapse to a horizontal line, but there would be inward pointing arrows as well.

It is important to note that the distances in these diagrams may not actually be distances in a Euclidean sense, and it may not be possible to construct triangles for some combinations of scores if the lengths of the edges do not satisfy the triangle inequality. One pathological case would be when the horizontal bar is longer than the sum of the other two sides of a triangle. That would mean that the distance between the source and target speakers is actually greater than the combined distances of the transformed speech to both the source and target speakers. The other pathological case would be when the distance from the transformed speech to one of the speakers was greater than the sum of the distance from the transformed speech to the other speaker plus the distance between the two speakers themselves. In such a case, it would also be impossible to construct a triangle. However, it should be noted that for all the examples we tried based on our data, we were able to construct triangles.

TTDs are not the first attempt to try to represent distances between speech in Voice Transformation. Others have used Multi-Dimensional Scaling (MDS) techniques to accomplish this [1]. In MDS, distances are calculated among multiple quantities in a multi-dimensional space, and the results are projected onto a plane for comparison. Although MDS is an interesting and useful technique for analyzing data, we find that TTDs are a compact, simpler-to-understand way of depicting the specific relationships we are trying to compare in Voice Transformation.

3.7.2 Evaluating Voice Transformation with TTDs

The TTDs for results from our listening experiment broken down by speaker pair are in Figure 3.10. These results correspond to the four speaker pairs from the graph in Figure 3.4. Looking at these TTDs, a number of things become readily apparent. First of all, the transformations were mostly successful in the sense that the triangles are skewed so the transformed speech is closer to the target speech than the source speech in each case. Another point is that the speakers in the first pair were considered much more similar than the others based on the widths of the diagrams. One interesting thing that appears in the third pair is that the transformation from ref to ehn is much more successful than the transformation from ehn to ref, as shown by the asymmetry in the diagram. This is another visual depiction of the same asymmetry mentioned earlier in the section on Data Analysis.

In our listening experiment, we found that whether the listeners knew the speakers did not appear to significantly affect how they judged speaker similarity. This knowledge will guide us in designing further experiments of this nature because we will not be concerned with finding listeners who either know or don’t know the speakers. We have also created a new type of diagram called a Transformation Triangle Diagram (TDD) that was useful in representing certain relationships in a compact, understandable manner. Future work
Figure 3.10: Transformation Triangle Diagrams by Speaker Pair
will involve investigating further methods of visualizing Voice Transformation results. While this paper investigates the area of speaker recognizability, there are other areas of Voice Transformation evaluation, such as intelligibility and naturalness, where different forms of analysis may be necessary.
Chapter 4

Proposed Experiments

This chapter describes the experiments we propose to conduct in investigating our thesis in Section 4.1 and lists our expected contributions in Section 4.2.

4.1 Experiments

Our thesis is that better underlying models of speech will improve Voice Transformation quality in terms of intelligibility, naturalness and identity. We propose using articulatory position data to improve these underlying models, and we also propose investigation of other model features, such as residuals, that may be used to improve transformed speech. It is difficult to judge improvement in the quality of speech. For this reason, a good part of our work will be concerned with evaluation.

Our preliminary experiments investigated the use of articulatory positions for improving speech models at prosodic and phonetic levels. At the prosodic level, connections between duration, \( F_0 \), and power were explored. The evidence suggested that in the type of recordings that were used, the connection between prosodic factors and articulatory positions was weak. At the phonetic level, connections between articulatory positions and acoustic-phonetic features were explored. These appeared to be stronger because certain phonetic features were predicted better through the use of articulatory position data.

These preliminary experiments used the articulatory position data directly without considering their physical nature very carefully. No constraints were placed on the positions and motions of the articulators, and no attempts were made to model the remainder of the vocal tract.

We propose investigating the following areas: articulatory positions in speech models, parameterization for Voice Transformation, and evaluation of Voice Transformation quality.
4.1.1 Articulatory Positions in Speech Models

We propose the construction of speech models, using articulatory positions, that take into account some of the physical constraints of articulators. Because our final goal is to improve Voice Transformation, it will be necessary for these models to predict quantities that are useful for Voice Transformation. Such quantities include, but are not limited to, MFCCs. A number of strategies for modeling constraints will be pursued.

Combinations of Positions

The preliminary experiments only used the positions of the articulators, which consisted of 7 (x,y) coordinate pairs in the mid-sagittal plane. Some important features may appear through the combination of features. For example, using a combination of the three tongue points could give information on the slope of the tongue, whether it curves up or down, and what is its highest point. Also, the position of the jaw, as represented by the location of the lower incisor may influence the effect of the lower lip. We propose adding the above combinations of articulators, along with others, to our MFCC and phonetic feature prediction experiments.

Modeling Vocal Tract Constriction

Constrictions in the vocal tract are related to properties of the speech signal. Adding some notion of the position of the palate could improve our models. We propose trying to predict the location of the palate based on articulatory position data and the electropalatograph data, which we have not previously used. The prediction of the palate location would then be used to predict constrictions, which would be used as additional features in experiments to predict MFCCs and phonetic features.

Using Static Position Constraints

Some popular models of articulators, including the GMM-based mapping approach from MFCCs, allow the articulators to be anywhere. The possible positions of articulators are obviously limited. We propose the use of models that assign zero probability to positions that are clearly outside of the realistic range of motion. Such models again will be used to predict MFCCs and phonetic features.

Using Dynamic Constraints

There are also limits to the speed and acceleration of articulators. This, in turn, should make the positions more predictable given history of the motions. Some form of linear dynamic system, or possibly even a linear predictive filter, may improve the predictions of positions. We propose the use of such models to predict articulatory positions.
4.1.2 Parameterization for Voice Transformation

While attempting to improve predictions of various quantities used in Voice Transformation, it will be necessary to step back and examine whether overall Voice Transformation quality is being improved. This will first require introducing the newly predicted quantities into the Voice Transformation process. However, there are also a number of areas which are not as directly related to articulatory positions that merit further investigation. In order to determine whether our attempts to improve Voice Transformation using articulatory position data are truly worthwhile, these areas should be examined as well.

One such area is the use of the residual in Voice Transformation systems that use an underlying LPC model for analysis and synthesis of speech. In speech synthesis based on residual-excited LPC, the perceived identity of the synthesized speech can be changed by using different excitations. Using the residual from a different speaker for excitation can make the synthesized speech sound like it came from that speaker. This approach is based on the theoretical result that for a system that truly is LPC-based, the residual equals the excitation. However, we know that speech production is not really based on LPC, although LPC appears to approximate it fairly well for some tasks. Attempting to predict residuals for Voice Transformation and then using them for excitations is a possible approach that may improve identity perception, although it is not theoretically correct.

Another residual-based approach is to consider the spectral domain and think of the residual as the deviation in the model’s predicted spectrum from the actual spectrum. This is theoretically cleaner than using the residual as the excitation. A number of researchers have taken this approach and experimented with residual predictions to improve the perceived identity of the transformed speech [18] [40] [33]. However, the residual from an LPC model is actually the error, or the difference between the model’s representation and the truth. In this case, modeling the residual means accepting the flaws in the original speech model and creating a second level in the model in an attempt to compensate for them. The model behind LPC is already considered to be a flawed representation of speech as it is an all-pole model, and proper representation of nasals theoretically requires zeros as well. In practice, this doesn’t cause much trouble for speech recognition systems, but as mentioned earlier, the error from the model appears to carry identity information for speech synthesis (and Voice Transformation systems based on such synthesis), and the error in part comes from the misrepresentation of nasals.

There are a couple directions that can be taken from this point. One is to continue to predict residuals for Voice Transformation and to consider the connection between the articulators and residuals. For example, because the nasality of speech is controlled by motions of the velum, the articulatory position data from MOCHA may be related in part to the residual. Furthermore, the other articulators may also be related to the LPC residual. The articulatory positions can be used to try to predict residuals.

Another direction is to use other initial speech models that do not have the
same flaws as LPC. Then the natural questions that arise include:

1. What are the flaws in these other speech models?
2. Do these models have any quantities that appear to correspond with speaker identity, in the way that the residual appears to for LPC?
3. Are there other benefits in terms of intelligibility, naturalness, or identity for these other models?
4. Do these other models have notions similar to residual excitation, and is there anything to be gained by exploring the analogous quantities?
5. If these other models do lead to better Voice Transformation, will any improvements from using articulatory position data still hold up?

**Prediction of Residuals Using Articulators**

We propose to use articulatory positions to aid in the prediction of residuals. We will use methods similar to those used in [18], [40] and [33], but have additional articulatory position features.

**Incorporation of Quantities Predicted from Articulators into Voice Transformation Systems**

After we have used articulators to predict quantities such as MFCCs, phonetic features, and residuals, we will want to incorporate these features into Voice Transformation systems. We propose doing this and comparing the results to non-enhanced Voice Transformation systems.

**Improvement of Articulatory-Based Voice Transformation**

In the preliminary experiments, we also performed Voice Transformation based on mapping the source speaker’s articulators to the target speaker’s articulators, followed by mapping from the predicted articulators to MFCCs and performing MLSA filter-based synthesis. The results of this were not as intelligible as we would have liked. We propose attempting to improve this type of Voice Transformation in a number of ways. Dynamic features were not used in the original articulatory position to MFCC map. We will add these and compare the quality of the transformed speech to the original approach. Another possibility would be to use MFCCs in addition to articulatory positions in the transformation mapping.

**Comparison of Voice Transformation Quality using Different Underlying Speech Models**

In the literature, a number of different underlying models have been used for the analysis and synthesis of speech in Voice Transformation. Some of the models are based on LPC, some are based on MFCCs and the MLSA filter, and some
are based on harmonic sinusoidal speech. These different techniques may lead to different quality in the transformed speech, and some may work better with transformation mappings than others. Without getting a better understanding of this, it will be hard to determine whether using articulatory positions is worthwhile. We propose performing Voice Transformation using all three of these analysis/synthesis methods and comparing the quality of the results with both objective and subjective metrics.

4.1.3 Evaluation of Voice Transformation Quality

Evaluation strategies will be an important part of the proposed experiments. Our goals of intelligibility, naturalness, and identity must be examined more closely to determine what they actually are. Numerous types of objective and subjective tests have been devised to measure the quality of Voice Transformation, but we really need to know whether these tests are measuring the correct things. A deeper understanding of the factors that contribute to intelligibility, naturalness, and identity will guide our evaluations and possibly how we construct our models.

Identity is a quality that is more typically associated with the goals of Voice Transformation than for speech recognition or even speech synthesis. But as mentioned earlier, there are different aspects to identity. Some tasks may require transformed speech to sound like it was produced by a specific, known individual. In other cases, for instance when reading a story, it may be desirable to have transformed speech that sounds like it came from the same speaker, but the style is changed to represent different characters in the story. These two aspects of identity were called specificity and distinctness earlier. These are not the only aspects of identity. Furthermore, it is useful to intelligibility and naturalness may also have aspects worth further consideration. There is also the question of which goals or aspects are most important. Is there a way to favor one over another for different tasks?

Evaluation

It is more difficult to list specific experiments for this area, because evaluation is an ongoing part of all the experiments, and new insights will arise as new data is collected. To begin with, our evaluation experiments will involve objective and subjective metrics that have been used in the past. These include metrics such as Mel-Cepstral Distortion, and listening tests, such as pair-comparison tests. We consider tests involving human perception to be the most important, although they are often more time-consuming than calculating quantities such as Mel-Cepstral Distortion. There are a couple directions to follow. One is to determine whether tests such as pair-comparison tests actually measure what we want to measure. For example, can they measure factors of identity such as specificity and distinctness? Another direction is to determine how well the objective metrics correspond to human perception. Are MFCCs enough? Should we also include pitch or other features in our metrics based on the quantity we
are trying to measure? Our work will be guided by various studies that have already been performed in the psycholinguistic literature.

4.2 Expected Contributions

The expected contributions of this work are:

- A demonstration that using articulatory positions aids Voice Transformation.
- Better quality Voice Transformation.
- A better understanding of how to evaluate Voice Transformation.
Chapter 5

Timeline

The estimated timeline for the completion of the proposed experiments and other work related to this thesis is as follows.

- January 2006: Combinations of Positions
- February 2006: Using Static Position Constraints
- March - May 2006: Using Dynamic Constraints
- February - May 2006: Modeling Vocal Tract Constriction
- June - July 2006: Prediction of Residuals Using Articulators
- August - October 2006: Improvement of Articulatory-Based Voice Transformation
- November - December 2006: Evaluation
Bibliography


