INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6” x 9” black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700  800/521-0600
Generating Gestural Scores
From Acoustic Data
Using Temporal Decomposition

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By

Michael J. Collins, M.S.

The Ohio State University

1998

Dissertation Committee:

Dr. Stanley C. Ahalt, Co-Adviser
Dr. Ashok K. Krishnamurthy, Co-Adviser
Dr. Mary E. Beckman

Approved by

Stanley C. Ahalt
Co-Adviser

Ashok K. Krishnamurthy
Co-Adviser

Department of Electrical Engineering
© Copyright by
Michael J. Collins
1998
ABSTRACT

Articulatory phonology [1] provides a description of speech in terms of the articulatory actions necessary to produce speech, instead of in terms of the acoustic consequences of those actions. These articulatory actions can be described abstractly using “gestures”. and the relative phasing and duration of these gestures in speech is shown on a “gestural score”. This articulatory model of speech offers potential improvement in speech coding, synthesis, and recognition. However, there presently exists a shortage of methods for generating the abstract gestural score from observed data such as an acoustic waveform.

Jung [2] and Collins [3] have explored the feasibility of generating gestural scores from articulatory data, and have produced an automated method to generate gestural scores from articulatory data corresponding to CV-C (consonant-vowel-consonant) tokens. McGowan [4, 5] has investigated recovering gestural scores from acoustics using an analysis-by-synthesis approach with synthetic speech, but achieved only moderate success on an extremely constrained problem.

The goal of the current research is to examine the feasibility of generating acoustic-based gestural scores, using an approach similar to that of Jung and Collins in their work with articulatory-based gestural scores. In particular, the present work evaluates the suitability of temporal decomposition [6, 7] for determining the overlapping
temporal locations of the gestures, using \textit{a priori} knowledge to perform the gesture tier placement.

To evaluate this mapping, several experiments are performed, including: comparing the acoustic-based gestural scores with gestural scores generated from simultaneously recorded articulatory data, applying the acoustic-based gestural scores to speech phenomena predicted by articulatory phonology, and resynthesizing speech from temporal decomposition target functions that have been modified by inserting, removing, and hiding target functions.
For my wife and family, who have been there through it all.
ACKNOWLEDGMENTS

I must first give my special thanks to Dr. Stanley Ahalt and Dr. Ashok Krishnamurthy, my advisers, for their helpful guidance, patience, and motivational encouragement. I would like to thank Dr. Mary Beckman for her interest and expertise, and for going beyond the "call of duty" at just about every occasion. In addition, I would like to thank my wife, family and friends for their support and advice, especially the IPS Lab (Bill, Tim, Mark, Nick, Ashu, Ed, Jayanth, et. al.), which has been like family to me.
VITA


April 1991 ............................................................. B.S. Electrical Engineering,
Cum Laude
The University of Pittsburgh
Pittsburgh, Pennsylvania. U.S.A.

March 1995 ........................................................... M.S. Electrical Engineering.
The Ohio State University
Columbus, Ohio. U.S.A.

Department of Electrical Engineering
The Ohio State University

Department of Electrical Engineering
The Ohio State University

PUBLICATIONS

"Variability in the Production of Quantal Vowels Revisited." Journal of the Acous­
murthy, T. Jung, et al.).

"Generating Gestural Scores From Articulatory Data Using Temporal Decompo­
sition", IEEE Transactions on Speech and Audio Processing, accepted. (with A. Kr­
ishnamurthy and S. Ahalt).

FIELDS OF STUDY

Major Field: Electrical Engineering
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. Background</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Gestural Scores</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Temporal Decomposition</td>
<td>16</td>
</tr>
<tr>
<td>2.2.1 Atal's temporal decomposition algorithm</td>
<td>16</td>
</tr>
<tr>
<td>2.2.2 Modifications by van Dijk-Kappers and Marcus</td>
<td>22</td>
</tr>
<tr>
<td>3. Generating Gestural Scores From Articulatory Data</td>
<td>26</td>
</tr>
<tr>
<td>3.1 Generating Articulatory Gestural Scores</td>
<td>26</td>
</tr>
<tr>
<td>3.1.1 Data</td>
<td>26</td>
</tr>
<tr>
<td>3.1.2 Data Warping and Normalization</td>
<td>28</td>
</tr>
<tr>
<td>3.1.3 Weighted Temporal Decomposition</td>
<td>33</td>
</tr>
<tr>
<td>3.1.4 Mapping Target Functions to Gestures</td>
<td>35</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>3.1.5</td>
<td>Recognition Tests</td>
</tr>
<tr>
<td>3.2</td>
<td>Automatic Generation of Articulatory Gestural Scores</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Temporal Decomposition</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Mapping Target Functions to Gestures</td>
</tr>
<tr>
<td>4.</td>
<td>Generating Gestural Scores From Acoustic Data</td>
</tr>
<tr>
<td>4.1</td>
<td>Acoustic Parameter Sets</td>
</tr>
<tr>
<td>4.2</td>
<td>Seeded Temporal Decomposition</td>
</tr>
<tr>
<td>4.3</td>
<td>Automatic Seeding/Segmenting</td>
</tr>
<tr>
<td>4.4</td>
<td>Mapping Target Functions to Gestures</td>
</tr>
<tr>
<td>4.5</td>
<td>System Diagram</td>
</tr>
<tr>
<td>5.</td>
<td>Evaluating the Gestural Scores Generated From Acoustic Data</td>
</tr>
<tr>
<td>5.1</td>
<td>Data</td>
</tr>
<tr>
<td>5.1.1</td>
<td>X-ray Microbeam Data</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Electropalatograph Data</td>
</tr>
<tr>
<td>5.2</td>
<td>Gestural Score Similarity Metric</td>
</tr>
<tr>
<td>5.3</td>
<td>Comparison With Articulatory Gestural Scores</td>
</tr>
<tr>
<td>5.4</td>
<td>Recognition Tests</td>
</tr>
<tr>
<td>5.5</td>
<td>Application to Coarticulation Phenomena</td>
</tr>
<tr>
<td>5.6</td>
<td>Resynthesis</td>
</tr>
<tr>
<td>6.</td>
<td>Conclusions</td>
</tr>
<tr>
<td>6.1</td>
<td>Summary of Current Work</td>
</tr>
<tr>
<td>6.2</td>
<td>Future Work</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Placement for each pellet.</td>
<td>27</td>
</tr>
<tr>
<td>3.2 The critical channels for the stop consonants and five vowels.</td>
<td>34</td>
</tr>
<tr>
<td>4.1 Similarity scores between gestural scores for each parameter set and the articulatory gestural score.</td>
<td>53</td>
</tr>
<tr>
<td>4.2 Phoneme mappings for the tier information.</td>
<td>67</td>
</tr>
<tr>
<td>5.1 Complete list of tokens used to compose the electropalatograph data set.</td>
<td>74</td>
</tr>
<tr>
<td>5.2 Similarity scores between gestural scores for each parameter set and the articulatory gestural score.</td>
<td>77</td>
</tr>
<tr>
<td>5.3 Percent of temporal decomposition target functions corresponding to each hand labeled phoneme for each of the parameter sets of interest.</td>
<td>78</td>
</tr>
<tr>
<td>5.4 Breakdown of missed phonemes by position within the CVC.</td>
<td>78</td>
</tr>
<tr>
<td>5.5 Breakdown of doubled phonemes by position within the CVC token.</td>
<td>79</td>
</tr>
<tr>
<td>5.6 Human viewer recognition accuracy for the acoustic gestural scores.</td>
<td>83</td>
</tr>
<tr>
<td>5.7 Percentage of the deleted phonemes detected by temporal decomposition at each speaking tempo.</td>
<td>86</td>
</tr>
<tr>
<td>5.8 Lengths of the gestures corresponding to deleted phonemes for each of the speaking rates.</td>
<td>90</td>
</tr>
<tr>
<td>5.9 Percentage overlap of the gestures corresponding to deleted phonemes for each of the speaking rates.</td>
<td>91</td>
</tr>
</tbody>
</table>
5.10 Listener identification accuracy for resynthesized tokens of each of three types of modifications.
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>The Haskins Laboratory forward system using gestural scores for articulatory dynamics (adapted from Browman and Goldstein, 1990).</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Vocal tract variables (adapted from Browman and Goldstein. 1989).</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Hypothetical gestural score fragments for “tan pants” (a) spoken deliberately showing no assimilation, and (b) in conversational speech showing assimilation.</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>Hypothetical gestural score fragments for “perfect memory” (a) words read individually showing no deletion, and (b) as a phrase in conversational speech showing deletion.</td>
</tr>
<tr>
<td>Figure 2.5</td>
<td>Hypothetical gestural scores for (a) “princess” showing no epenthesis and (b) “prince” showing epenthesis.</td>
</tr>
<tr>
<td>Figure 2.6</td>
<td>The variables used in adjusting the analysis window.</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Jung’s method for generating gestural scores from raw articulatory data.</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Original and warped Cartesian spaces.</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Original and transformed data spaces for /i/ and /a/.</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Original and transformed articulatory traces for the token “a kook o[f]” /a kuk a/, showing how the warping transforms the consonant /k/s to become “humps” on the velar pellet traces. The marked segments indicate the approximate locations of the /k/s.</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Lip gestures derived from the target function.</td>
</tr>
</tbody>
</table>
3.6 Collins' method for generating gestural scores from raw articulatory data.

3.7 Histograms for X (constriction location) and Y (constriction degree) channels of the LIP pellet. The dark distribution is for reconstruction weights corresponding to labial gestures, while the light distribution is for reconstruction weights not corresponding to a labial gesture.

3.8 Error plot for the LIP X channel reconstruction weights decision boundary.

4.1 Current system for generating gestural scores from acoustic data.

4.2 Averaged magnitude spectrum across all pellets and all tokens of articulatory data.

4.3 The utterance “a beep o[f]” (/a bip a/) for four of the examined parameter sets low-pass filtered to 50 Hz. Each plot shows the multi-dimensional data as the upper traces and the resulting temporal decomposition target functions at the bottom.

4.4 The utterance “a beep o[f]” (/a bip a/) for the remaining four examined parameter sets low-pass filtered to 50 Hz. Each plot shows the multi-dimensional data as the upper traces and the resulting temporal decomposition target functions at the bottom.

4.5 The utterance “a beep o[f]” (/a bip a/) for four of the examined parameter sets low-pass filtered to 10 Hz. Each plot shows the multi-dimensional data as the upper traces and the resulting temporal decomposition target functions at the bottom.

4.6 The utterance “a beep o[f]” (/a bip a/) for the remaining four examined parameter sets low-pass filtered to 10 Hz. Each plot shows the multi-dimensional data as the upper traces and the resulting temporal decomposition target functions at the bottom.

4.7 Similarity plots for acoustic gestural scores using cepstral coefficients low-pass filtered with five cutoff frequencies and articulatory gestural scores offset in time by various amounts.
4.8 Similarity plots for acoustic gestural scores using log-area parameters low-pass filtered with five cutoff frequencies and articulatory gestural scores offset in time by various amounts. ................................................................. 55

4.9 Similarity plots for acoustic gestural scores using log-area ratios low-pass filtered with five cutoff frequencies and articulatory gestural scores offset in time by various amounts. ................................................................. 56

4.10 Three results of temporal decomposition on the same articulatory data, differing only in initial analysis window size. The marked segments show the initial location of the analysis windows. The data is from the utterance “a beep of[l]” (/a bip a/). ................................................................. 58

4.11 The three possible cases for determining the next starting window location for seeded TD. The vertical lines show the provided “hard” boundaries. The shaded box shows the starting location of the next analysis window. ................................................................. 60

4.12 Comparison of the standard TD used by Jung and Collins to seeded TD. The data shown here is articulatory data for the utterance “a beep of[l]” (/a bip a/), and the vertical lines indicate the seeded boundaries. 62

4.13 Plot of each of the three segmenting metrics and their sum for the utterance “a beep of[l]” (/a bip a/). ................................................................. 64

4.14 Mapping an acoustic target function to a gesture. ................................................................. 65

4.15 Gestural score for the utterance “beak” (/bik/) generated by the acoustic-to-gestural score system. A plot of the original acoustic waveform, the log-area parameters and the seeded temporal decomposition target functions is shown for reference. ................................................................. 70

5.1 Examples of missed target functions for both articulatory data and log-area parameters. The tokens are “beeg” /big/ and “teeg” /tig/. Asterisks mark the seeded interval in which no target function was found. 80

5.2 Examples where one target function is found for each underlying event. Both articulatory data and log-area parameters. The tokens are “derg” /dɔɡ/ and “beeg” /big/. ................................................................. 81
5.3 Examples of doubled target functions for both articulatory data and log-area parameters. The tokens are “curd” /kərd/ and “gerg” /ɡəɡ/. Asterisks mark the seeded interval in which two target functions were found.

5.4 Hypothetical gestures demonstrating hiding. The top plot shows gestures becoming increasingly overlapped as the speaking rate increases. The bottom plot shows the resulting acoustics.

5.5 Temporal decomposition results for “a Galatians o[f]” (/ə ɡælətʃænz ə/) at the normal and fastest speaking rates.

5.6 Temporal decomposition results for “a gorilla o[f]” (/ə ɡərɪlə ə/) at the normal and fastest speaking rates.

5.7 Target functions for “a Galatians o[f]” (/ə ɡælətʃænz ə/). Notice that the target function corresponding to the deleted schwa (marked with an asterisk) is still present even at the fastest speaking rate. The acoustic waveform is shown for reference.

5.8 Temporal decomposition data for the original tokens “cape” (/kɛp/) and “scape” (/skep/). The target function corresponding to the /s/ in “scape” (marked with an asterisk) is moved to the beginning of “cape”, to form a synthesized “scape” (see Figure 5.9).

5.9 Temporal decomposition data and spectrogram for the synthesized token “scape”, an example of resynthesizing a token after inserting a target function.

5.10 Temporal decomposition data for the original tokens “grill” (/ɡrɪl/) and “rill” (/rɪl/). The target function corresponding to the /ɡ/ in “grill” (marked with an asterisk) is removed to form a synthesized “rill” (see Figure 5.11).

5.11 Temporal decomposition data and spectrogram for the synthesized token “rill”, an example of resynthesizing a token after removing a target function.
CHAPTER 1

INTRODUCTION

Mankind has long dreamed of the possibility of using speech to communicate with machines. With the advent of computers midway through the twentieth century, popular belief was that this dream would soon be a reality. However, human speech production and perception is complex, and the lack of complete knowledge about these systems has made man-machine speech communication more difficult to achieve than originally believed. Some of this difficulty may lie in the use of an inappropriate internal representation.

A fundamental principle of language organization is that the larger structures of speech are composed of smaller parts in coordination. For example, the sentence “your memory is perfect” is composed of the smaller parts “your”, “memory”, “is” and “perfect”. The smaller parts can be recombined in other ways to form different larger structures. Parts from “your memory is perfect” could be recombined with other words to form a different phrase, such as “my perfect memory”. Words are similarly composed of recurring subparts. For example, “my” combines the beginning sound of “memory” with the ending of other words such as “tie” and “guy”.

In actual speech, however, the subparts of words do not recombine neatly—they cannot be treated simply as “beads of a string”. The reality of speech production is
that it is not simply a concatenation of discrete, static, invariant elements, but is much more complex, with each sound changing and being changed by its neighbors. This influence of one phoneme on its neighbors is called coarticulation, and it manifests itself in many ways. The most common coarticulation phenomena are referred to as assimilation, deletion and epenthesis. A brief traditional explanation of each follows.

Assimilation occurs when one phoneme becomes more like a nearby phoneme. For example, in the phrase “tan pants” (/tæn#pænts/), the final consonant in “tan” assimilates the place of articulation of the first consonant in “pants”. This results in a phrase which sounds more like “tam pants” ([tæm pænts]). Another example is the English consonant “k”, which is produced differently depending on the place of articulation of the following vowel. A “front k” is used before “front” sounds, such as in “key” [kɪ]; whereas a “back k” is used before “back” sounds, such as in “coo” [ku].

Deletion occurs when a phoneme appears to be removed in certain contexts. For example, when spoken quickly, “support” /səpɔːrt/ sounds more like “sport” [sport]. Another example is “perfect memory” (/pəˈfɛkt mɛməri/), where the “t” sounds like it is dropped in casual speech ([pəˈfɛk mɛməri])

Finally, epenthesis is when a phoneme appears to be inserted, often in a cluster of unrelated sounds. The inserted phoneme typically has characteristics of the sounds surrounding it, forming a “bridge” between the unrelated sounds. Examples of epenthesis are the words “warmth” /wɔːmt/ and “prince” /prɪns/, which sound like “warmpth” [wɔːmpθ] and “prints” [prɪnts] in conversational speech.

The effects of coarticulation are problematic for traditional phonetic representations. The difficulty lies in that, unlike the simplified descriptions presented above might indicate, coarticulation cannot be quantified into a discrete set of “states”. 
There exists a continuum of coarticulation effects. Drawing on the previous examples, /k/ isn’t produced in only a “front” and “back” variety—its place of articulation varies along that entire region of the palate. Deleted and inserted phonemes are present (or not present) in various degrees, instead of just “present” or “not present”. Referring again to the previous examples, “support” doesn’t actually become “sport”, but rather something in between [8, 9].

To address these effects, traditional phonology uses phonetic rules to describe the changes that occur in each context. However, these rules fail to adequately capture the continuous nature of the phenomenon. Additionally, these segment-based rules fail to provide insight as to why these effects occur. They are just lists detailing where each coarticulation effect has been observed. There is nothing which relates the coarticulation effects, nor do the rules provide a way to predict how and where the effects occur.

Several attempts have been made to improve the modern speech recognition systems which are based on these representations. These improvements include: developing neural network systems that can learn temporal patterns (recurrent neural networks) [10], adding transition nodes to hidden Markov models to improve their temporal properties (segment models) [11], and using “trigram” models to explicitly encode all of the coarticulation effects [12]. However, these efforts are attempts to overcome or compensate for the effects of coarticulation, when a better solution might be develop a new representation which can more accurately model it.

One such attempt at a representation which captured coarticulation was presented by Browman and Goldstein [13, 14, 15, 1], who suggested a description of the speech signal using dynamically specified units of articulatory actions called gestures. The
relative phasing and duration of the gestures is called a *gestural score*. The gestural score model holds the promise of providing a more unified and parsimonious description of the speech signal, naturally linking the linguistic and physical domains of the signal.

The gestural score is attractive because it describes speech using timing in an explicit and natural manner. Effects such as coarticulation fall out as natural consequences of the speech production system. Assimilation is modeled by two overlapping gestures—the overlap results in an acoustic consequence which shares features of both sounds. Deletion occurs when two gestures completely overlap a third gesture that is between them, resulting in little or no acoustic consequence for the hidden gesture. Epenthesis can occur when consecutive multi-channel gestures overlap. The resulting combination of overlapping gestures produces the inserted sound. Unlike the way in which the phonetic rules dictate that phonemes are being inserted and deleted, there is no inserting or deleting of gestures during coarticulation. In all cases, the coarticulation effects can be described by simple modifications of the relative positions, magnitudes and phasing of the involved gestures. For illustration, the earlier example of deletion for the /t/ in “perfect memory” (/pəfɛkt mɛməri/) is shown on a gestural score by the gestures for the preceding /k/ and following /m/ encroaching on the time spanned by the gestures for the /t/, effectively covering them.

However, one of the drawbacks of this model is the lack of methods to infer the abstract gestural score from physical signals that can be easily measured. Ideally, one would like to infer the gestural score from the speech signal waveform; however, work in this area has been limited, with only a few preliminary steps taken towards this end. As a first step toward a goal of using the speech signal waveform, Jung
et. al. [2] developed a method of obtaining the gestural score from articulatory data, and showed that those gestural scores could be used to recognize consonant-vowel-consonant (CVC) tokens. However, Jung’s method for generating gestural scores required a priori knowledge. This knowledge consisted of the gestural score tier on which to place each gesture once that gesture’s temporal location was determined. Collins [3] extended and improved Jung’s method, resulting in an automated method for generating gestural scores from articulatory data without the use of a priori knowledge.

McGowan [4, 5] considered a method for recovering gestural scores from the acoustic signal using a genetic algorithm [16] in an analysis-by-synthesis [17, 18, 19] framework. Working completely with synthesized data, McGowan attempted to recreate four of the vocal tract variables for four different tokens. With enough constraints, McGowan was able to extract gestural scores with moderate success, but at high computational cost. Unfortunately, this method cannot be readily extended to analyzing natural speech data, because it requires a forward model similar to that which produced the speech being analyzed. The data of McGowan’s work was synthetic, so the same speech synthesizer that was used to produce the original speech was also available as the forward model. This is not the case for real speech.

The current work explores generating acoustic-based gestural scores (hereafter referred to as acoustic gestural scores for brevity) by applying techniques similar to that of Jung and Collins. To this end, a system is developed to generate gestural scores from acoustics, using a modified temporal decomposition algorithm for the gesture timing information and a priori knowledge for the gesture tier information.
To ascertain the effectiveness of this system, in particular the effectiveness of temporal decomposition to determine the gesture timing, several experiments are conducted, including: comparing the acoustic gestural scores with gestural scores generated from simultaneously recorded articulatory data, applying the acoustic gestural scores to speech phenomena predicted by articulatory phonology, and resynthesizing speech from temporal decomposition target functions that have been modified by inserting, removing, and hiding target functions.

This dissertation is organized as follows. Background to the problem, including a description of gestural scores and a derivation of temporal decomposition, is presented in Chapter 2. The previous work of Jung and Collins on generating gestural scores from articulatory data is presented in Chapter 3. Chapter 4 details the current system for generating gestural scores from acoustics. Chapter 5 describes the recognition, coarticulation and resynthesis experiments conducted to evaluate the system. Finally, Chapter 6 summarizes the work and proposes future research.
CHAPTER 2

BACKGROUND

The goal of this research is to investigate the generation of gestural scores from acoustic data. The following sections provide background context for the task.

2.1 Gestural Scores

Believing that the inherently multidimensional nature of articulation could be utilized to concisely explain many phonological phenomena, Browman and Goldstein [14, 15] have developed a representation for speech called the *gestural score*. The gestural score is a multi-channel representation of the relative magnitudes, locations and durations of *gestures*. A gesture is an abstract entity which represents a coordinated articulatory movement, often with the intent of achieving a constriction in the vocal tract.

Gestural scores are part of a computational model for speech being developed at Haskins Laboratory by Browman, Goldstein, Saltzman and Rubin [20, 13]. A block diagram of their system is displayed in Figure 2.1. This system is a complete forward model which creates synthesized acoustic output from an intended utterance, using gestural scores and articulatory trajectories as intermediate representations.
The vocal tract model in this system is an articulatory synthesizer dubbed “ASY” [21], which is based on Mermelstein’s articulatory model [22]. ASY is driven by articulatory trajectories, which are provided by a task dynamic model.

The task dynamic model is that proposed by Saltzman [23, 24]. Briefly, task dynamics attempts to model the human body’s ability to perform tasks in a “goal oriented” manner. For example, when reaching for a glass of water, the hand approaches the glass in a quasi-straight line, from any direction, despite the complex motion required by the articulators (the joints and muscles) involved. In addition to this trajectory shaping, research has shown that these goal oriented motions also demonstrate immediate compensation—the articulators reorganize themselves when faced with an external perturbation without requiring direct control by the brain [25, 26]. For example, the hand can still accomplish its “task” even if the elbow is restricted.
by a sling, or if the hand is briefly pushed on its way to the glass. Kelso et al. [27] showed that the speech articulators also demonstrate this type of behavior.

To model the trajectory shaping and immediate compensation task dynamic behavior of the human body, Saltzman uses a system of linear, uncoupled, second order equations. This system transforms discrete activation levels in a task space into articulator trajectories in the articulator space. That is, it converts task dynamics to articulator kinematics. From Figure 2.1, it is clear that gestures are the dynamic inputs which result in speech articulator kinematics; gestures are one step more abstract than speech articulations.

The task space used by gestural scores is a constriction space defined using vocal tract variables. Figure 2.2 shows all of the vocal tract variables proposed by the research group at Haskins Laboratory; this set is an initial working set hypothesized to be the smallest needed to uniquely represent all of the phonemes in American English. Each tract variable is associated with a set of articulators, and represents quantities such as constriction degree or constriction location. It is important to remember, however, that each tract variable represents a quantity in the task space, not the positions of any of the individual articulators used to accomplish this goal. For example, the opening formed by the lips is reflected in the lip aperture (LA) tract variable. It represents the culmination of the positions of the upper lip, lower lip, and the jaw. However, the specific positions of any of these articulators is unimportant (from a task dynamics viewpoint) so long as the correct lip aperture is produced.

Note that many of the vocal tract variables are grouped into horizontal-vertical pairs, where both variables share the same set of articulators. For instance, tongue
body gestures are specified in terms of the tract variables of tongue body constriction location (TBCL) and constriction degree (TBCD).

The gestural score is a multi-channel representation of the activation levels, locations, and durations of gestures defined in the vocal tract variable space. Viewed as a phonological representation for speech, the gestural score can easily represent coarticulation phenomena such as assimilation, reduction and epenthesis. Figures 2.3 through 2.5 demonstrate how gestural scores can account for the various coarticulation effects.

The first figure of this series shows a part of the hypothetical gestural score for the phrase “tan pants”. Figure 2.3 (a) shows the phrase spoken carefully, while Figure 2.3 (b) shows the phrase used in conversational speech. When used in conversational speech, the consonant /n/ in “tan” apparently assimilates to the place of articulation of the following /p/ consonant (/tæn pænts/ → [tæm pænts]). On the gestural score, this is shown by the increased overlap between the velum gesture for the /n/ and the LA gesture for the /p/.

The next figure of this series shows a part of the hypothetical gestural score for the phrase “perfect memory”. Figure 2.4 (a) shows the two words read individually from a list of words, while Figure 2.4 (b) shows the two words used as a phrase in conversational speech. When produced in conversational speech, the final /t/ consonant of “perfect” appears to be deleted (/pər'fekt# məməri/ → [pə'fek məməri]). On the gestural score, this is shown by the /t/ gesture on the TTCD tier being completely hidden by the /k/ and /m/ gestures on the TBCD and LA tiers. Because of the topology of the vocal tract, the early closure of the lips for the /m/ gesture prevents any acoustic consequence for the /t/ gesture from being produced. The
<table>
<thead>
<tr>
<th>Tract Variables</th>
<th>Articulators Involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lip Protrusion (LP)</td>
<td></td>
</tr>
<tr>
<td>Lip Aperture (LA)</td>
<td>upper and lower lips, jaw</td>
</tr>
<tr>
<td>Tongue Tip Constriction Location (TTCL)</td>
<td></td>
</tr>
<tr>
<td>Tongue Tip Constriction Degree (TTCD)</td>
<td>tongue tip, body, jaw</td>
</tr>
<tr>
<td>Tongue Body Constriction Location (TBCL)</td>
<td></td>
</tr>
<tr>
<td>Tongue Body Constriction Degree (TBCD)</td>
<td>tongue body, jaw</td>
</tr>
<tr>
<td>Velic Aperture (VEL)</td>
<td>velum</td>
</tr>
<tr>
<td>Glottal Aperture (GLO)</td>
<td>glottis</td>
</tr>
</tbody>
</table>

Figure 2.2: Vocal tract variables (adapted from Browman and Goldstein, 1989).
Figure 2.3: Hypothetical gestural score fragments for “tan pants” (a) spoken deliberately showing no assimilation, and (b) in conversational speech showing assimilation.
closure of the lips is the dominant constriction, overshadowing the closure formed by the tongue tip.

The last figure of this series shows a part of the hypothetical gestural score for the words “prince” and “princess”. Figure 2.4 (a) shows the word “princess”, while Figure 2.4 (b) shows the word “prince”. When produced in conversational speech, epenthesis occurs in “prince”, with a /t/ appearing between the /n/ and the /s/ (/prins/ → [prints]). On the gestural score, this is shown by the phasing of the /n/ gestures. Because the VEL and GLO gestures for the /n/ end before the TTCD gesture for the /s/ begins, there is a brief time in which the vocal tract configuration is that of a /t/. This same change does not happen in “princess” however, due to the differing syllabic structure and consequently differing gesture timing.

In all three cases, it is important to note that the coarticulation effect is explained by simple modifications of the relative positions, magnitudes and phasing of the involved gestures. Gestures aren’t actually inserted or deleted as phonemes are in the traditional phonological view of coarticulation. Nor are there the numerous context-sensitive rules describing each case. All of the effects are handled by adjusting the relative gesture timings.

Unfortunately, Browman and Goldstein only provided the forward model for gestural scores. Therefore, Jung [2] investigated the usefulness of Browman and Goldstein’s gestural score as an articulatory-phonetic representation and developed a method to generate gestural scores from articulatory data (see Section 3.1).
Figure 2.4: Hypothetical gestural score fragments for "perfect memory" (a) words read individually showing no deletion, and (b) as a phrase in conversational speech showing deletion.
Figure 2.5: Hypothetical gestural scores for (a) “princess” showing no epenthesis and (b) “prince” showing epenthesis.
2.2 Temporal Decomposition

The works of Jung and Collins, as well as the present work, use temporal decomposition to determine the temporal placement of gestures. Temporal decomposition, which was originally proposed by Atal [6], models a set of parameters by a group of overlapping target functions and target vectors. The target functions are a temporal basis set for the parameters, and the target vectors are the weights necessary to reconstruct the original parameters from the target functions. Atal developed the temporal decomposition algorithm as a method for low bit-rate encoding of speech, but later work has focused on using temporal decomposition as a tool for speech segmentation [7].

2.2.1 Atal's temporal decomposition algorithm

The set of speech parameters chosen by Atal [6] to describe the speech signal are log-area parameters low-pass filtered to 50 Hz. The log-area parameters are derived from linear prediction analysis [28, 29] on the acoustic waveform. Briefly, the log-area parameters are found by taking the logarithm of the tube areas derived from the reflection coefficients of Levinson’s recursion algorithm [30]. These parameters are highly correlated and vary slowly in time, which are two criteria for successful application of temporal decomposition [31].

Using the notation of Atal, let \( y_i(n) \) be the \( i \)th log-area parameter at the \( n \)th sampling instant, \( \phi_k(n) \) be the \( k \)th target function, and \( a_{ik} \) be the contribution of the \( k \)th target function to the \( i \)th area parameter. The number \( n \) is a discrete index of time and varies between 1 and \( N \). Atal assumes that one can approximate the original
speech parameter, $y_i(n)$, by a linear combination of target vectors and functions

$$
\hat{y}_i(n) = \sum_{k=1}^{m} a_{ik} \phi_k(n), \quad 1 \leq n \leq N, \quad 1 \leq i \leq p
$$

(2.1)

where $p$ is the total number of log-area parameters, and $m$ is the number of target functions (roughly proportional to the number of speech events), and $N$ is the duration of the speech utterance in samples.

In matrix notation, Equation 2.1 is expressed as

$$
Y = \mathcal{A}\Phi
$$

(2.2)

where $Y$ is a $p \times N$ matrix whose $(i, n)$ element (ith row and nth column, or the value for the $i$th log-area parameter at time $n$) is $y_i(n)$. $\mathcal{A}$ is a $p \times m$ matrix whose $(i, k)$ element is $a_{ik}$ (weight for $k$th function’s contribution to $i$th parameter). and $\Phi$ is an $m \times N$ matrix whose $(k, n)$ element is $\phi_k(n)$. For our purposes, $Y$ is the matrix of known log-area parameter values, while both $\mathcal{A}$ and $\Phi$ are unknown and must be determined during the analysis.

The temporal decomposition algorithm derives the target functions $\phi_k(n)$ sequentially, making them chronologically ordered. That is, the function $\phi_{z+1}(n)$ occurs later than the function $\phi_z(n)$ and so on. Each target function roughly corresponds to a speech event in the utterance.

In general, the rank of $Y$ is greater than or equal to $m$; therefore, taking the pseudo-inverse of Equation 2.2 yields

$$
\Phi = (\mathcal{A}^T \mathcal{A})^{-1} \mathcal{A}^T Y.
$$

(2.3)

This implies that a target function $\phi_k(n)$ can be represented as

$$
\phi_k(n) = \sum_{i=1}^{p} w_{ki} y_i(n), \quad 1 \leq k \leq m, \quad 1 \leq n \leq N
$$

(2.4)
That is, that the target functions are a linear combination of the original speech parameters. Here, the $w_{ki}$ values are unknown, as $A$ is unknown in Equation 2.3. Note that for efficient coding, Atal assumed that the target functions were compact in time, implying that the majority of the $w_{ki}$ would be zero ($W$ would be sparse). This constraint reduces the large number of unknown $w_{ki}$ to a manageable number.

In Atal's analysis, the input dimensionality is reduced by first performing a singular value decomposition (SVD) on the log-area parameters. Only the first few singular values are retained and the small singular values are zeroed. That is,

$$Y' = UDV',$$

where $U$ is an $N \times p$ orthogonal matrix, $V$ is a $p \times p$ orthogonal matrix, and $D$ is a diagonal matrix of singular values (the square roots of the eigenvalues of $Y'Y$). Typically, the largest 3 to 5 singular values are enough to contain more than 95% of the variance of the original signals. Therefore, the number of input vectors can be reduced if the log-area parameters are approximated with only the first 3 to 5 singular vectors. The target functions $\phi_k(n)$ can then be represented as

$$\phi_k(n) = \sum_{i=1}^{m} b_{ki} u_i(n).$$

where the $b_{ki}$ are a set of amplitude coefficients and $u_i(n)$ is the element in the $n$th row and the $i$th column of the matrix $U$.

In order to determine the target functions, Atal defines a measure of spread $\theta(n)$ from the sample $n = l$ as

$$\theta(l) = \left( \frac{\sum_n \alpha(n) \phi^2(n)}{\sum_n \phi^2(n)} \right)^{1/2}$$

18
where $\alpha(n)$ is the weighting factor and the sum over the index $n$ extends over the entire speech segment. Atal proposed a quadratic weighting factor for $\alpha(n)$:

$$\alpha(n) = (n - l)^2. \quad (2.8)$$

The target functions should be as compact as possible so that the majority of the $w_{ki}$ of Equation 2.4 are zero, implying that the measure of distance $\theta(n)$ should be minimized. To obtain the optimal target function, replace $\phi(n)$ of Equation 2.7 by the expression of Equation 2.6. We then find the derivatives of $\theta(n)$ with respect to the unknown amplitude coefficient $b_{ki}$. Using $\ln(\theta(n))$ gives the same results as differentiating $\theta(n)$ but is computationally more efficient. The differentiation yields

$$\sum_n (n - l)^2 \frac{\partial}{\partial b_r} \phi^2(n) = \lambda \sum_n \frac{\partial}{\partial b_r} \phi^2(n), \quad 1 \leq r \leq m. \quad (2.9)$$

where

$$\lambda = \frac{\sum_n (n - l)^2 \phi^2(n)}{\sum_n \phi^2(n)} = \theta_{\text{min}}^2. \quad (2.10)$$

From Equation 2.6.

$$\phi^2(n) = \sum_{i=1}^m \sum_{j=1}^m b_{ij} u_i(n) u_j(n). \quad (2.11)$$

where the subscript $k$ has been dropped. Then,

$$\frac{\partial}{\partial b_r} \phi^2(n) = 2 \sum_{i=1}^m b_i u_i(n) u_r(n), \quad 1 \leq r \leq m. \quad (2.12)$$

Combining Equation 2.9 and 2.12 yields

$$\sum_{i=1}^m b_i \sum_n (n - l)^2 u_i(n) u_r(n) = \lambda \sum_{i=1}^m b_i \sum_n u_i(n) u_r(n) = \lambda b_r. \quad (2.13)$$

which can also be expressed in matrix notation as

$$Rb = \lambda b, \quad (2.14)$$
where the element in the \( i \)th row and \( r \)th column of the matrix \( R \) is given by

\[
R_{ir} = \sum_n (n - l)^2 u_i(n) u_r(n).
\]  

(2.15)

This is an eigenvalue problem, and Equation 2.14 has exactly \( m \) solutions. If all the \( \lambda 's \) are different, the solution will be the vector corresponding to the smallest \( \lambda \). In case they are not, the minimum value of \( \lambda \) determines the optimum \( b \), although the optimum \( b \) is not unique. Replacing the \( b \) of Equation 2.6 with the optimum \( b \) from the above procedure, one obtains the target function in the interval of speech segment.

The amplitude coefficients \( a_{ik} \) of Equation 2.1 are determined by minimizing the mean-squared error \( E \) defined by

\[
E = \sum_n [y_i(n) - \sum_{k=1}^{M} a_{ik} \phi_k(n)]^2
\]  

(2.16)

where \( M \) represents the total number of speech events. Taking the partial derivatives of \( E \) with respect to the coefficients \( a_{ik} \) and setting them equal to zero, one obtains a set of simultaneous linear equations

\[
\sum_{k=1}^{M} a_{ik} \sum_n \phi_k(n) \phi_r(n) = \sum_n y_i(n) \phi_r(n), \quad 1 \leq r \leq M, \quad 1 \leq i \leq p,
\]  

(2.17)

which can be solved for the unknown coefficients \( a_{ik} \). Atal used an iterative refinement process to provide better estimates of \( \phi_k(n) \) and \( a_{ik} \).

To analyze an entire utterance, Atal finds a target function every 10 ms. using a fixed length window. This approach produces numerous redundant target functions, so Atal employed a reduction algorithm to eliminate them. Among similar target functions, only the target function with the center of mass closest the center of its analysis window is kept.
Atal’s procedure temporally decomposes the set of parameter functions derived from the speech signal into a set of temporal target functions and their reconstruction weights (target vectors). It is important to note that the target functions overlap in time. Later investigations into temporal decomposition examine it in the context of being a tool for performing speech segmentation. The temporal overlap of the target functions overcomes a major limitation of traditional segmentation. The temporally overlapping “events” allow effects such as those resulting from coarticulation to be modeled. The motivating idea is that the target vectors may be associated with ideal articulatory positions, and the target functions describe the temporal evolution of these targets [31, 32].

Atal [6] performed temporal decomposition on several sentences spoken by male and female speakers. As mentioned earlier, for input Atal used eight log-area parameters derived from the acoustic waveform, and used the target functions and target vectors to reconstruct the speech signal. The reconstructed signal agreed well with the original speech signal. However, in Atal’s work, temporal decomposition was developed for efficient speech coding. Atal achieved a reduction in the bit-rate for transmitting the speech segment, but did not consider the possible underlying phonetic meaning of the decomposition.

Van Dijk-Kappers and Marcus [7] modified temporal decomposition to improve the determination of the number and location of the target functions. They also showed that there is a correspondence between the locations of the target functions and the locations of hand-labeled phonemes. These improvements are described in the following section.
2.2.2 Modifications by van Dijk-Kappers and Marcus

Van Dijk-Kappers and Marcus [7] differed from Atal in their approach to temporal decomposition. Atal simply used temporal decomposition as a method of speech compression, while van Dijk-Kappers and Marcus used temporal decomposition to perform segmentation. They investigated the possible phonetic meaning of the target vectors and target functions. In addition, they provided several improvements to the algorithm.

The first modification concerns the weight factor for measuring the distance of $\theta(n)$ from the sample $n = l$ (refer to Equation 2.8). The choice of this weighting factor can influence the shape of the target functions. Since Atal's weighting factor is quadratic and the distance metric of Equation 2.7 is minimized, the procedure tends to force target functions to be centrally located [7] and "hump" shaped. However, if the target functions truly reflect underlying articulatory gestures, they could be asymmetric with respect to the center of the analysis window, as the analysis window may not be centered on the underlying gesture. Consequently, van Dijk-Kappers and Marcus [7] proposed an alternative weighting factor of an adaptable rectangular window:

$$\alpha(n) = \begin{cases} 
1 & \text{for } n_1 \leq n \leq n_2 \\
0 & \text{elsewhere} 
\end{cases} \quad (2.18)$$

where $n_1$ and $n_2$ are inside the bounds of $n$ in Equation 2.7, making $\alpha(n)$ act as a small windowing function.

In this case, $\theta(n)$ is a measure of how well the target function fits in the adaptable rectangular window. If $n_1$ and $n_2$ extend over the whole analysis window, then $\theta(n) = 1$, and the window is the exact size of the target function. Therefore, Equation 2.7 should be maximized in this case in order to obtain the optimal target function.
(minimizing $\theta(n)$ yields the degenerate case $n_1 = n_2$). Since both the location and duration of the target function are unknown for each analysis window, the values of $(n_1, n_2)$ are also unknown. Van Dijk-Kappers and Marcus employed an iterative procedure to determine the optimal target function.

The procedure proposed by van Dijk-Kappers and Marcus starts with a small rectangular window placed in the center of the analysis window. A first approximation to the target function is determined accordingly. The location and extent of the next $(n_1, n_2)$ is between the frames where the current target function has the threshold value $h_w$. This procedure is repeated until the location and extent of $(n_1, n_2)$ becomes stable. At the end of the iterative procedure an optimal target function for this analysis window has been derived, normalized to a peak value of one.

A second modification involves the length of the analysis window. A fixed-length analysis window was used in Atal’s work. Marcus and van Lieshout [33] found Atal’s algorithm could only find target functions roughly equal in size to the analysis window. Changing the window size resulted in a different set of target functions. Since Van Dijk-Kappers and Marcus hoped to have the temporal decomposition target functions have phonetic meaning (e.g., the location and the extent of an underlying segment), it was obvious that they would have to modify temporal decomposition to overcome this limitation.

One of the clearest problems with the use of fixed-length windows arises when the window size is too large. When this occurs, edge effects (leakage) from neighboring speech events become a problem. Van Dijk-Kappers and Marcus avoid this problem, as well as the that of an undersized window, by adapting the window size to the length of the target function. In order to adjust the location and size of the window,
the maximum of the target function \( r_{max} \) within the window is determined. Next the locations of \( n_{left} \) and \( n_{right} \), the closest indices to the left and right respectively of \( r_{max} \) where the current target function has values less than a threshold value \( h_w \), are determined. If there is no such \( n_{left} \) within the window, then the target function is incomplete on the left bound and the first frame of the window, \( n_0 \), will be assigned to \( n_{left} \). The same procedure applies to the last frame of the window and \( n_{right} \). (reference Figure 2.6)

![Diagram](image.png)

**Figure 2.6:** The variables used in adjusting the analysis window.

On the other hand, if there are an \( n_{left} \) and \( n_{right} \) which satisfy the rule, the edge effects are examined by calculating the measures

\[
\text{left edge effect} = \sum_{n_{left}}^{n_0} \phi^2(n)
\]

and

\[
\text{right edge effect} = \sum_{n_{right}}^{n_n} \phi^2(n).
\]

where \( n_n \) is the last frame of the window. If either edge effect exceeds a threshold value, \( S \), the corresponding boundary of the window is shortened. A more detailed
description of this technique can be found in [7]. Related work [34] has shown that good choices for the values of \( h_w \) and \( S \) are 0.24 and 0.06, respectively.

These iterative procedures continue until an optimal target function is found within the analysis window. The analysis window is then shifted temporally to the next speech event of interest. Instead of using Atal’s procedure of performing an analysis every 10 ms, van Dijk-Kappers and Marcus place the next analysis window where we expect to find the next target function. Van Dijk-Kappers and Marcus suggested that the best choice is the \( n_{\text{right}} \) of the previous target function. This removes the need for the reduction step of Atal.

With the modifications provided by van Dijk-Kappers and Marcus, Atal’s temporal decomposition technique has become a powerful tool in analyzing speech. Studies show that the temporal decomposition process used by van Dijk-Kappers and Marcus has a fair degree of correspondence between the number and locations of the target functions and the number and locations of the phonemes in the speech segment that the target functions were derived from [33, 31]. This desirable quality makes temporal decomposition a candidate for finding the locations and durations of the underlying gestures of a speech segment.
CHAPTER 3

GENERATING GESTURAL SCORES FROM ARTICULATORY DATA

The goal of this research is to investigate the generation of gestural scores from acoustic data. The following sections detail previous work in generating articulatory gestural scores, upon which the current work draws heavily.

3.1 Generating Articulatory Gestural Scores

In his work, Jung investigated the usefulness of Browman and Goldstein's gestural score as an articulatory-phonetic representation. Jung [2] developed a method to convert raw articulatory data to gestural scores (using a priori knowledge) similar to those proposed by Browman and Goldstein [15]. The consonant-vowel-consonant (CVC) tokens that each gestural score represented were recognizable with a high degree of accuracy by a trained artificial neural network. Jung's work involved several steps, as shown in Figure 3.1. An explanation of each block in the figure is presented in the following sections. For a complete discussion, please see [2].

3.1.1 Data

The articulatory data used by Jung was collected at the X-ray microbeam facility [35] at the University of Wisconsin. The facility allows the tracking of up to
Figure 3.1: Jung’s method for generating gestural scores from raw articulatory data.

ten gold pellets for an extended period of time by tracking the pellets using X-ray microbeams instead of using full raster X-ray scans. The locations of the pellets are shown in Table 3.1. Electroglottograph and acoustic data were synchronously recorded.

<table>
<thead>
<tr>
<th>Pellets</th>
<th>Placement</th>
<th>Size (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference 1</td>
<td>at gum line between the two maxillary incisors</td>
<td>3.0</td>
</tr>
<tr>
<td>Reference 2</td>
<td>bridge of nose, 58 mm above reference pellet 1</td>
<td>3.0</td>
</tr>
<tr>
<td>jaw pellet</td>
<td>at gum line between the mandibular incisors</td>
<td>3.0</td>
</tr>
<tr>
<td>upper lip</td>
<td>center of vermillion border</td>
<td>3.0</td>
</tr>
<tr>
<td>lower lip</td>
<td>center of vermillion border</td>
<td>3.0</td>
</tr>
<tr>
<td>tongue blade</td>
<td>10 mm behind the tongue tip</td>
<td>2.5</td>
</tr>
<tr>
<td>tongue body anterior</td>
<td>35 mm behind the tongue tip</td>
<td>2.5</td>
</tr>
<tr>
<td>tongue body posterior</td>
<td>48 mm behind the tongue tip</td>
<td>2.5</td>
</tr>
<tr>
<td>tongue dorsum</td>
<td>70 mm behind the tongue tip</td>
<td>2.5</td>
</tr>
</tbody>
</table>

Table 3.1: Placement for each pellet.

The subject was a male who is a native speaker of a Midwestern dialect of American English. The subject was seated so that all of the data would be collected as if viewing the subject from his right side. A full-head X-ray scan was taken of the
subject. In addition, a “finger trace” of the subject’s palate was generated by tracing the palate with a gold pellet glued on the finger of a glove. Both of these pieces of data, along with the inventory of extreme back positions of the tongue dorsum pellet, are used to reconstruct the shape of the vocal tract wall (upper palate and velum).

The database consists of 420 utterances produced by the subject. The target tokens are CVCs embedded in the sentence *Say a CVC(1) of a CVC(2) again*. The consonants are drawn from the set of English stop consonants /p,t,k,b,d,g/ and the vowel is drawn from the set /i,u,æ,a,ə/. In every utterance, the nuclear accent was placed on the first CVC. Each CVC token appeared at least twice in both stressed and unstressed contexts.

### 3.1.2 Data Warping and Normalization

The raw articulatory data as collected from the University of Wisconsin X-ray microbeam facility records the locations of the pellets using a Cartesian coordinate system. Unfortunately, this coordinate system is not optimal for our methods in determining the gestural score, as we are primarily attempting to locate constrictions. The task of generating the gestural score might be simplified if the articulatory data were “warped” to a space resembling the vocal tract variables used by Browman and Goldstein [14, 15]. Jung used a warping transformation where the x dimension represents constriction location and the y dimension represents constriction degree.

First, a 6th degree polynomial was fit to an estimation of the vocal tract wall and evaluated at 500 equally spaced x coordinates [34]. For each time instant and each tongue pellet, all of the Euclidean distances were calculated between the pellet’s x-y
coordinates and all the points along the vocal tract wall. The closest point on the vocal tract yielded the warped x-y coordinate:

- The x coordinate is the distance along the vocal tract.
- The y coordinate is the distance from the vocal tract wall.

Figure 3.2 shows the original and warped spaces. The lip pellets were warped into the lip protrusion and lip aperture vocal tract variables:

- The x coordinate is the x coordinate of the lower lip.
- The y coordinate is the distance between the upper and lower lip pellets.

The effect of these transformations is to warp the raw articulatory data into a space where constriction articulations are far easier to detect. For example, consult Figure 3.3. The top two figures both show the distributions of the four tongue pellets during “target points” for all of instances of the phoneme /i/. If the vocal tract trace were removed from the tract, however, it would be unclear as to which phoneme was being displayed. In the warped space, it is obvious that the vowel is the high, front vowel /i/. Similarly, the bottom two figures show the distributions of the four tongue pellets for /a/. For consonants, the effect of the transformation into the warped space is to make the articulations (constrictions) appear “hump” shaped on the relevant articulatory traces, allowing temporal decomposition to detect them much easier. Figure 3.4 shows the warped and unwarped articulatory traces for the token “a kook o[f]” (/ə kuk a/). Notice how the consonant closures for the two /k/s are more pronounced on the TB3 traces. The marked segments indicate the approximate locations of the /k/s in each utterance. One additional step was taken to condition the data.
Figure 3.2: Original and warped Cartesian spaces.
Figure 3.3: Original and transformed data spaces for /i/ and /a/. 
Figure 3.4: Original and transformed articulatory traces for the token “a kook o[f]” /ə kük a/. showing how the warping transforms the consonant /k/s to become “humps” on the velar pellet traces. The marked segments indicate the approximate locations of the /k/s.
Each pellet was normalized so that each pellet covered the range [0, 1]. That is,

\[ x_{\text{new}} = \frac{x_{\text{old}} - \min(x_{\text{old}})}{\max(x_{\text{old}}) - \min(x_{\text{old}})} \]  

(3.1)

and

\[ y_{\text{new}} = \frac{y_{\text{old}} - \min(y_{\text{old}})}{\max(y_{\text{old}}) - \min(y_{\text{old}})} \]  

(3.2)

where the \text{min} and \text{max} were taken over the entire database of a particular trace.

In this way, temporal decomposition will treat each channel as equally important, for every pellet moves from 0 to 1 as it covers its full range of movement.

Finally, to preserve the direction of the movement of the tongue, the warped trajectories of pellets in the y-dimension were reversed with respect to the vocal tract wall as

\[ y\text{-value} = 1 - \text{distance from the vocal tract wall} \]  

(3.3)

such that the higher the value is, the higher the tongue position is (and the less the amount of opening). In this way, gestures appear as humps, not valleys.

### 3.1.3 Weighted Temporal Decomposition

Because temporal decomposition approximates the trajectories in a mean-squared-error fashion across all of the channels, too much emphasis can be placed on minimizing the overall error of all channels at the expense of an accurate model for the trajectories of the “critical” channel(s). Typically, only a subset of the articulatory channels (e.g., lip channels for bilabial constrictions) is relevant for analyzing the articulatory movements associated with a given gesture. As a goal of Jung’s work was to design accurate gestural scores for testing, he found it inefficient for temporal decomposition to use all of the information to model any given gesture, since only a
few channels are "critical" in articulating a particular sound. Therefore Jung made a modification to the temporal decomposition algorithm called weighted temporal decomposition (WTD). Under WTD, only the presumed critical channel(s) among the multi-dimensional articulatory data are actually used. In effect, the inputs become

\[ y_t(n) = \begin{cases} x_t(n) & \text{if channel } i \text{ is critical} \\ 0 & \text{otherwise} \end{cases} \]  

(3.4)

for \( n_1 \leq n \leq n_2 \), where \((n_1, n_2)\) is the duration within which the sound is articulated. This modification has the additional benefit of concentrating temporal decomposition on the timing of events only on the critical channels for any gesture, instead of the timing of all concurrent events present for that instant.

The critical channel changes according to the expected gesture, and as such, requires \textit{a priori} knowledge of the inputs. The critical channels used by Jung are shown in Table 3.2. Please see [34] for a complete discussion of weighted temporal decomposition.

<table>
<thead>
<tr>
<th>Pellet</th>
<th>Consonants</th>
<th>Vowels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p &amp; b</td>
<td>t &amp; d</td>
</tr>
<tr>
<td>Lip Protrusion</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Lip Aperture</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Tongue Tip Location</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Tongue Tip Degree</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Tongue Blade Location</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Tongue Blade Degree</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Tongue Body Location</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Tongue Body Degree</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Tongue Dorsum Location</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Tongue Dorsum Degree</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 3.2: The critical channels for the stop consonants and five vowels.
Using the *a priori* knowledge of the critical channels, weighted temporal decomposition was applied to the CVC database to obtain the corresponding target functions and target vectors. Next, those target functions and vectors were converted to a gestural scores.

### 3.1.4 Mapping Target Functions to Gestures

Temporal decomposition provides target functions and reconstruction weights (target vectors) as output. The target functions provide the location and duration of the underlying temporal events, and the reconstruction weights provide the values to recreate approximations of the original multi-channel data from a weighted sum of the target functions. The task is to determine the amplitude, duration, and channel of the gestures to be placed on the gestural score using the information derived by temporal decomposition.

The CVC database contains the stop consonants /p, b, t, d, k/ and /g/. Stop consonants normally involve the closure of the vocal tract at the place of articulation. Since these consonants involve the formation and release of a constriction, they should show a constriction maximum on the critical y-dimension channels during the activation intervals. Temporal decomposition should find these “humps”: the resulting target functions provide their location and duration. Jung set the starting frame of the gesture to the first frame where the value of the target function was non-zero, and set the ending frame to the one where the target function reached its maximum. In some sense, the gesture can be viewed as a step activation to the task dynamic system, with the resulting articulatory trajectories being the step response. The activation level of the gesture was set to the value of the reconstruction weight.
corresponding to the critical channel. Once determined, the gestures were placed on the channels corresponding to the critical channels. That is, they were placed using \textit{a priori} knowledge. Figure 3.5 shows an example of generating a gesture from a target function for a bilabial stop consonant.

Vowels are slightly less straightforward. They are described by Browman and Goldstein in terms of their place of articulation and tongue height (which corresponds to constriction degree). On the gestural score, vowels are mainly modeled by gestures on the tongue body (TB) tier, with lip protrusion and aperture (LP and LA) also being used for vowels with lip rounding. Thus, it is desirable that the magnitudes of gestures on the TB channels reflect the quantities of constriction location and degree. For example, for the high front vowel /i/, the location of the tongue body constriction should be around the anterior portion of vocal tract, suggesting that the magnitude of the gesture on TBCL is rather small. To find the appropriate constriction degree and constriction location values, Jung conducted a principal components analysis on the positional values of the tongue pellets in the new warped space for all of the vowels of a particular speaker. The analysis enabled him to determine the contribution each
pellet makes to each vowel. He made use of those vowel quantities to derive the
gestural scores for the vowel, setting the values on the appropriate channels to values
determined by his analysis as representative of the "front"ness and "high"ness of the
various vowels.

Voicing is also represented on the gestural score. The simultaneously recorded
electroglottograph data are used to determine the voicing of consonants in various
utterances. The EGG signal was first filtered to remove the low frequencies (under
100 Hz) caused by the gross movements of the glottis in the electroglottograph data.
Next, the energy of the electroglottograph data was computed, and the voicing ac­
tivity was determined by comparing the energy with a threshold value. The sound
segments are voiced if their energy exceeds a threshold value, otherwise they are
marked as unvoiced.

3.1.5 Recognition Tests

To test if the generated gestural scores were a viable representation for the CVC
tokens. Jung performed two sets of recognition tests. In the first test, three Elman
Recurrent Networks [10] were trained (one for each phoneme) to recognize the various
phonemes that would appear in their respective positions (initial consonant, vowel,
final consonant). For example, one network was trained to recognize the vowel, its
five outputs being one each for /i, u, æ, a/ and /ɛ/. Each node was trained so that
the sum of its outputs remained constant. Recognition accuracy was extremely high
(98% on average), showing that the gestural scores were a suitable representation for
recognition of the CVC tokens. To verify that they were also phonetically significant.
human recognition tests were also conducted. Two graduate students from the Linguistics department at The Ohio State University performed the tests. Each had the channels on the gestural score explained, then attempted to classify all of the CVC tokens from their gestural scores. Again, recognition accuracy was extremely high, implying that the gestural score possesses phonetic significance.

Convinced of the suitability of the gestural score as a phonetic representation, the next task was the refinement and automation of Jung's process for generating gestural scores from articulatory data.

3.2 Automatic Generation of Articulatory Gestural Scores

Collins investigated the automated generation of gestural scores from articulatory data for a class of CVC tokens. The work was a continuation of that initiated by Jung, with emphasis on the generation of articulatory gestural scores without using a priori knowledge. The block diagram of the approach is shown in Figure 3.6 (compare with Figure 3.1). Much of the systems of Jung and Collins are the same, so only the major differences will be detailed.

![Figure 3.6: Collins' method for generating gestural scores from raw articulatory data.](image)

38
3.2.1 Temporal Decomposition

The temporal decomposition algorithm used by Collins to perform the segmentation of the tokens was that proposed by van Dijk-Kappers and Marcus (see Section 2.2.2). This is a departure from Jung’s work, as Jung used the weighted temporal decomposition algorithm of his own design. However, weighted temporal decomposition requires *a priori* knowledge in the form of the weighting, and this information was not available in the automatic system of Collins’ work. Therefore, modified temporal decomposition as proposed by van Dijk-Kappers and Marcus was used in its original form. Note that without the weighting, the temporal decomposition target functions likely correspond to the gesture ensemble timing, instead of the timing of the individual gestures for the critical articulator(s). That is, the target functions corresponded to the timing of the effective constriction [14], not solely the timing of the critical articulator’s gestures.

3.2.2 Mapping Target Functions to Gestures

As with Jung, Collins converted the temporal decomposition target functions to gestures phased relative to each other on the gestural score. The same method for generating the gestures from the target functions was used by Jung and Collins (see Section 3.1.4), however, the mechanism for placing the gestures on the gestural score was different. The goal of Jung’s work was to produce accurate gestural scores for testing. As such, *a priori* knowledge was used to place the gestures on the gestural score. The work of Collins was to automate the generation of the gestural scores from the articulatory data; consequently, the *a priori* information was not available. An
alternate statistical-based method was developed to place the gestures on the gestural score.

Collins postulated that since the target functions are normalized to a maximum height of unity, the reconstruction weights for each channel are roughly proportional to the maximum for the channel during the time frames for which the target function is active. In the CVC data used by Jung and Collins, this claim is certainly true for the channels on which there are stop gestures: the articulatory movements for the stops appear as "humps" which temporal decomposition should capture. Assuming that gestures manifest themselves on the warped data as "humps" which are normalized to unity, reconstruction weights corresponding to gestures should also be close to unity. That is, since

\[ \text{trace max} \approx \text{reconstruction weight} \times \text{target function max} \quad (3.5) \]

and since the target function maximum is fixed at one, one can conclude that, roughly speaking,

\[ \text{trace max} \approx \text{reconstruction weight} \quad (3.6) \]

Since Collins assumed that gestures manifest themselves in the warped data as humps close to unity. Equation 3.6 implies that their corresponding reconstruction weights should also be close to unity, while the reconstruction weights of target functions not mapping to gestures would be somewhat smaller.

Histograms of the LIP reconstruction weights for target functions corresponding to bilabial consonant gestures and those which do not correspond to bilabial consonant gestures are shown in Figure 3.7. In these figures, the abscissa represents the
Figure 3.7: Histograms for $X$ (constriction location) and $Y$ (constriction degree) channels of the LIP pellet. The dark distribution is for reconstruction weights corresponding to labial gestures, while the light distribution is for reconstruction weights not corresponding to a labial gesture.

value of the reconstruction weight, and the ordinate represents the number of occurrences. The dark histograms are the distributions of the reconstruction weights for target functions corresponding to bilabial gestures, while the light histograms belong to the reconstruction weights of target functions which do not correspond to bilabial consonant gestures. Plots of the tongue tip (TT) reconstruction weights for dental consonant gestures, of the tongue dorsum (TD) reconstruction weights for velar consonant gestures and vowel gestures are all qualitatively identical to that shown. These distributions provided supporting evidence for the claim that the reconstruction weight of target functions corresponding to gestures should be larger than those which do not correspond to a gesture, as the dark distributions consistently had a higher mean than the light distributions. In addition, with the exception of the three velar $X$ distributions, all the pellets show a reasonable degree of separation. The
three velar X pellets apparently did not correlate well with the presence of consonant gestures (though the velar Y pellets do), and consequently were not used in later procedures. This is because English velar consonants are influenced by the following vowel, and vary widely in place of articulation.

Decision boundaries were chosen from the experimentally determined classification error curve to minimize total error. A typical error plot is shown in Figure 3.8. Using the obtained decision boundary, the reconstruction weights were classified and their corresponding gestures placed on the gestural score. For each target function, the gestures were placed such that the total error was minimized. Accuracy tests found 92% of the gestures to be correctly placed.
CHAPTER 4

GENERATING GESTURAL SCORES FROM ACOUSTIC DATA

The goal of this research is to investigate the generation of gestural scores from acoustic data. To this end, a system has been designed and implemented; the block diagram of the system is shown in Figure 4.1. Compare this block diagram with Figures 3.1 and 3.6, the block diagrams of the articulatory systems of Jung and Collins. The following sections provide a description of the various system blocks and

![Diagram](image)

Figure 4.1: Current system for generating gestural scores from acoustic data.
the procedure used to generate gestural scores from acoustic data. An evaluation of this system can be found in Chapter 5.

4.1 Acoustic Parameter Sets

Comparing the system that generates gestural scores from acoustics (Figure 4.1) with the articulatory systems of Jung and Collins (Figures 3.1 and 3.6), it is obvious that the present system draws heavily on its predecessors. In particular, temporal decomposition is still used to determine the overlapping temporal locations of the gestures. Temporal decomposition, as explained in Section 2.2, requires an input signal which is multi-channel and highly correlated. In the research of Jung and Collins, X-ray microbeam articulatory data were used as the input, which satisfy both of these criteria. Unfortunately, the raw acoustic data of interest in the present research does not meet either criteria. Therefore, our first task in generating acoustic gestural scores is to transform the acoustic signal into an appropriate representation.

In the preliminary stages of this research, various acoustic parameter sets were examined and evaluated for their suitability. The suitability was based on three measures.

First, the representation has to capture the temporal locations of the underlying gestures as accurately as possible. As the “true” gesture locations are unknown, the present work uses both articulatory target functions and hand-labeled phoneme locations to evaluate the acoustic target function locations. Target functions extracted by temporal decomposition from the multi-channel acoustic representation have to reasonably align with one of these “gold standards”, which are used as surrogates for the “true, underlying” gestures. Otherwise, there are grounds to suspect that the
determined target functions are capturing some phenomena other than the underlying gestures. This is undesirable, as the generated gestural scores should be phonetically relevant.

The other two constraints placed on the input representation are that the representation has to be multi-dimensional, and that the data should be highly correlated. These criteria are necessary if temporal decomposition is to be used to determine the underlying temporal locations of the gestures, as temporal decomposition performs best with highly correlated multi-channel data.

The task of examining parameter sets for their suitability to temporal decomposition was considered by van Dijk-Kappers [31], who compared several acoustic parameter sets to determine the correspondence between the hand-labeled phoneme locations and the location of the target functions determined by temporal decomposition. All of the parameter sets considered were based on linear predictive coefficients (LPC) with the exception of filter bank outputs. Van Dijk-Kappers found that the log-area parameters of Atal were the most desirable, having a one-to-one correspondence between phonemes and target functions 73% of the time. Additionally, 98% of the phonemes were covered by one or two target functions.

The current problem is similar to that considered previously by van Dijk-Kappers, so many of the same parameter sets are evaluated for their suitability to the present task. Cepstral parameter sets [36] are also included, as they been successfully applied to various speech problems [37, 38]. The parameter sets considered are:

- Autocorrelation coefficients
- Area coefficients
- Cepstral coefficients
- Log-area parameters
- Log-area ratios
- Line spectral frequencies
- Reflection coefficients

Four of these parameter sets—area coefficients, log-area parameters, log-area ratios, and reflection coefficients—are closely related, so a brief word of clarification may be in order. Reflection coefficients are obtained from the LPC analysis using Levinson’s recursion algorithm [30]. The reflection coefficients result from the boundaries in a tube model of the vocal tract, which is an equivalent representation of the linear prediction model. Using the reflection coefficients, one can obtain the areas of the tubes (area coefficients), the logarithm of the areas of the tubes (log-area parameters), and the logarithm of the ratios of the areas of the tubes (log-area ratios). Note that while strictly speaking the tube models are derived from the acoustic speech waveform, they can not be considered to reflect the “actual” vocal tract which generated the speech.

Each parameter set is extracted from the acoustic signal using 16 ms frames with an 8 ms overlap. The frames are chosen to provide the same sampling rate as the articulatory data. Due to the physical limitations of the speech production system, speech events occur slowly in time. The articulatory data confirms this. Figure 4.2 shows the averaged magnitude spectrum across all of the channels of all of the articulatory tokens. The bandwidth is clearly less than 10 Hz, as the magnitude is down 30 dB from peak at that point.
Figure 4.2: Averaged magnitude spectrum across all pellets and all tokens of articulatory data.
Unfortunately, the bandwidth of all of the acoustically-derived parameter sets is practically the full 60 Hz. The presence of the time-varying glottal waveform, as well as model-mismatch, introduce analysis artifacts. Therefore, to improve the correlation between the acoustic parameter sets and the articulatory data, the multi-channel data is low-pass filtered with a 30 tap FIR filter. The filtering also improves the performance of temporal decomposition, which performs better with slow-changing data. Filters with cutoff frequencies of 50 Hz, 40 Hz, 30 Hz, 20 Hz, and 10 Hz are all examined. After filtering, seeded temporal decomposition is applied to each of the parameter sets. Figures 4.3 through 4.6 show each parameter set, generated for the example utterance “a beep of” (/a bip ə/), filtered first at 50 Hz and then filtered at 10 Hz. The articulatory data are also shown. Additionally, the corresponding temporal decomposition target functions for each parameter set are shown for comparison purposes.

The resulting target functions are compared against those of the corresponding articulatory data using the metric detailed later in Section 5.2. Briefly, the metric, termed similarity, is a measure of how similar the gesture locations and durations from one gestural score are to the gesture locations and durations in another gestural score. The similarity scores for each parameter set are shown in Table 4.1. Three representations performed well for both cutoff frequencies: cepstral coefficients, log-area parameters, and log-area ratios. Figures 4.7, 4.8, and 4.9 show similarity plots of the three representations for various delays and low-pass filter cutoff frequencies. The same general trends appear in all three plots; increased filtering increases similarity, and the maximum similarity occurs when the articulatory gestural scores are delayed
Figure 4.3: The utterance “a beep of” (/ə bip ə/) for four of the examined parameter sets low-pass filtered to 50 Hz. Each plot shows the multi-dimensional data as the upper traces and the resulting temporal decomposition target functions at the bottom.
Figure 4.4: The utterance "a beep o[f]" (/a bip a/) for the remaining four examined parameter sets low-pass filtered to 50 Hz. Each plot shows the multi-dimensional data as the upper traces and the resulting temporal decomposition target functions at the bottom.
Figure 4.5: The utterance “a beep o[l]” ([a bip a]/) for four of the examined parameter sets low-pass filtered to 10 Hz. Each plot shows the multi-dimensional data as the upper traces and the resulting temporal decomposition target functions at the bottom.
Figure 4.6: The utterance “a beep of” (/a bip a/) for the remaining four examined parameter sets low-pass filtered to 10 Hz. Each plot shows the multi-dimensional data as the upper traces and the resulting temporal decomposition target functions at the bottom.
Table 4.1: Similarity scores between gestural scores for each parameter set and the articulatory gestural score.

<table>
<thead>
<tr>
<th>Parameter Set</th>
<th>Similarity 50 Hz</th>
<th>Similarity 10 Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autocorrelation coefficients</td>
<td>0.678</td>
<td>0.726</td>
</tr>
<tr>
<td>Area coefficients</td>
<td>0.670</td>
<td>0.725</td>
</tr>
<tr>
<td>Cepstral coefficients</td>
<td>0.725</td>
<td>0.731</td>
</tr>
<tr>
<td>Log area parameters</td>
<td>0.702</td>
<td>0.736</td>
</tr>
<tr>
<td>Log area ratios</td>
<td>0.715</td>
<td>0.736</td>
</tr>
<tr>
<td>Line spectral frequencies</td>
<td>0.694</td>
<td>0.716</td>
</tr>
<tr>
<td>Reflection coefficients</td>
<td>0.676</td>
<td>0.729</td>
</tr>
</tbody>
</table>

15-30 ms. However, only the log-area parameters show improvement with each successively lower low-pass filter cutoff frequency. In particular, the 10 Hz filtered cepstral coefficients exhibit some peculiar behavior, showing a peak performance with the acoustic gestures shifted 16 ms prior to the articulatory gesture, a dramatic drop in performance with shifts from -16 ms to 108 ms, then a sharp rise again in performance with extreme shifts beyond 108 ms.

From this examination, log-area parameters are considered the best for our purposes, and are therefore used as the primary input representation throughout the rest of this research.

### 4.2 Seeded Temporal Decomposition

Temporal decomposition has been an important segmentation tool in previous work on generating gestural scores [2, 3]. It was chosen over other tools because it allows "soft" boundaries between overlapping segments, whereas traditional segmentation techniques yield "hard" boundaries between strictly abutting segments.
Figure 4.7: Similarity plots for acoustic gestural scores using cepstral coefficients low-pass filtered with five cutoff frequencies and articulatory gestural scores offset in time by various amounts.
Figure 4.8: Similarity plots for acoustic gestural scores using log-area parameters low-pass filtered with five cutoff frequencies and articulatory gestural scores offset in time by various amounts.
Figure 4.9: Similarity plots for acoustic gestural scores using log-area ratios low-pass filtered with five cutoff frequencies and articulatory gestural scores offset in time by various amounts.
However, temporal decomposition is not without disadvantages. As Marcus and van Lieshout [33] documented, Atal's algorithm could only find target functions roughly equal in size to the analysis window. In response, Van Dijk-Kappers and Marcus [7] modified temporal decomposition to allow the analysis window to change in size. However, temporal decomposition is still influenced by the initial window size and window overlap chosen for the analysis. Compare Figures 4.10 (a), (b) and (c), the output of three temporal decomposition analysis differing only in the initial window size for each target function. The marked segments reflect the locations of the initial analysis window for each target function. As is evident from the figures, temporal decomposition is strongly predisposed to find target functions that are in roughly the same size and location as the specified windows. The presence of this inherent bias indicates that standard temporal decomposition cannot be utilized to accurately locate speech events with locations and durations markedly different from those of the analysis window. As speech segments vary in length, it is therefore difficult to pick one temporal window initial analysis window size which will accurately find all segments.

To circumvent the difficulties posed by the bias of standard temporal decomposition, we have introduced a mechanism for dynamically specifying the location and duration of the starting analysis window. This new version of temporal decomposition, dubbed "seeded temporal decomposition," takes advantage of this inherent predisposition of the algorithm by using windows which approximately correspond to the location of the underlying speech events. Using such windows encourages the temporal decomposition algorithm to locate these speech events. That is, we can
Figure 4.10: Three results of temporal decomposition on the same articulatory data, differing only in initial analysis window size. The marked segments show the initial location of the analysis windows. The data is from the utterance “a beep o[ʃ]” (/a bip a/).
“seed” the temporal decomposition algorithm with “hard” segment boundaries, and then allow the algorithm to “soften” the boundaries into overlapping segments.

The seeded temporal decomposition implementation differs from standard temporal decomposition only at the point at which a target function has been found and a new analysis window needs to be established. Standard temporal decomposition places a fixed length analysis window (typically 150-250 ms) after the current window, with a small overlap (typically 50 ms). As mentioned above, this predisposes the standard temporal decomposition algorithm to find a target function in approximately this location and of this duration. However, the resulting target function may not always be the best match for the underlying temporal event. Therefore, seeded temporal decomposition uses the supplied hard segment boundaries to try to position the next analysis window in the location most likely to yield a correctly placed target function.

If the current target function approximately spans the original location of the current analysis window, seeded temporal decomposition places the next analysis window with its right edge at the next specified hard boundary. This is the most desirable case, where one target function is found which roughly corresponds in location and duration with each specified underlying temporal event. This case is shown in Figure 4.11(a).

If the current target function is sufficiently smaller, the right of edge of the next analysis is the same as the original right edge of the current analysis window. This case implies that the specified underlying temporal event is composed of more than one acoustic consequence, and therefore, more than one target function (such as in a closure and release for a stop consonant). This case is shown in Figure 4.11(b).
If the current target function is large enough that it spans multiple hard boundaries, the right edge of the next analysis window is set to the first hard boundary not already covered. This case results in a target function which covers multiple underlying temporal events, possibly indicating a dominant gesture sufficiently covered by its neighbor so as to have little independent acoustic consequence and therefore be unrecoverable as a separate entity by our method. This case is shown in Figure 4.11(c).

Figure 4.11: The three possible cases for determining the next starting window location for seeded TD. The vertical lines show the provided “hard” boundaries. The shaded box shows the starting location of the next analysis window.

In all cases, the left edge of the next analysis window is placed overlapping the current window by a small amount (typically 50 ms).

In this way, the seeded temporal decomposition is able to dynamically place each analysis window in such a way as to provide the highest correspondence between the specified underlying temporal event locations and the resulting target functions.
Figure 4.12 shows a comparison between the articulatory target functions derived by standard and seeded temporal decomposition. The vertical lines indicate the seeded boundaries, which are shown for illustration. Notice that the target functions derived by seeded temporal decomposition vary more widely in length than those of standard temporal decomposition, indicating that seeded temporal decomposition is more capable of finding underlying temporal events which vary in size.

With these modifications, seeded temporal decomposition should prove to be a more accurate tool for determining the underlying temporal events in speech.

4.3 Automatic Seeding/Segmenting

For the seeded TD algorithm to dynamically adjust the initial window location for each target function, segment boundaries must be provided.

To automatically segment the utterances, a combination of three metrics are used: signal energy difference, zero crossing rate difference, and cepstral distance. Each of the three metrics indicates a change in the speech waveform: signal energy difference shows changes between sonorants and obstruents (and to a lesser degree voiced and unvoiced sounds), zero crossing rate difference shows changes between voiced and unvoiced as well as between fricatives and non-fricatives, and the cepstral distance indicates changes in the vocal tract shape. Each of these metrics is calculated in a straightforward manner.

The signal energy is calculated by taking the sum across all the points of the FFT for an analysis frame. The frames are 32 ms long with a 16 ms overlap. The signal energy difference is simply the vector formed by the difference of successive frames.
Figure 4.12: Comparison of the standard TD used by Jung and Collins to seeded TD. The data shown here is articulatory data for the utterance “a beep o[f]” (/ə bip a/), and the vertical lines indicate the seeded boundaries.
The zero crossing rate is calculated by first centering the speech waveform by subtracting off the mean, then counting the number of transitions from either positive to negative or negative to positive. This count is performed for each of the 32 ms long frames (again with 16 ms overlap). The zero crossing rate difference is the vector formed by the difference of successive frames.

The cepstral coefficients are calculated from 16th-order LPC values found for 32 ms frames with 16 ms overlap. The cepstral difference is the vector formed by the difference of successive frames of cepstral coefficients.

Each of the three metrics are calculated separately and normalized to have a peak magnitude of unity. Afterwards, they are added together in a weighted fashion to form one metric. Currently, the three metrics are equally weighted. Figure 4.13 shows an example of the three metrics and the combined metric. The vertical lines on the sum metric represent hand labeled boundaries.

The combined metric is a measure of similarity from sample to sample: spikes in the metric indicate borders between dissimilar sections of the speech utterance. The locations of these spikes for each token are therefore used as the boundaries for the following seeded temporal decomposition step.

4.4 Mapping Target Functions to Gestures

The temporal location of the gestures are derived from the temporal decomposition target functions in a manner similar to that of Jung, differing only in the duration of the derived gesture. In Jung's work, the starting frame of the gesture bundle was the first frame where the value of the target function is non-zero, and the ending frame was where the target function reached its maximum. In the present work, the
Figure 4.13: Plot of each of the three segmenting metrics and their sum for the utterance “a beep [f]” (/ə bip ə/).
starting frame is the same, but the ending is the last frame that the target function is non-zero. That is, the gesture bundle has exactly the same location and duration as the target function. Figure 4.14 shows an example of converting the timing of an acoustic target function to a gesture. This choice is motivated by the examples of

![Target function and Gesture](image)

**Figure 4.14**: Mapping an acoustic target function to a gesture.

Browman and Goldstein (see [15], Figures 19.13 through 19.15).

The glottal tier gestures are derived from the Entropics Waves program get_f0, which provides a "probability of voicing" [39] along with its primary purpose of tracking the fundamental frequency (F0). The sound segments are marked as voiced (active) if their probability of voicing exceeds a threshold value, otherwise they are marked as unvoiced.

At this point, there are no velic gestures, as there are no nasals in the corpus.

In addition to determining the temporal location of each gesture, the tier placement and gesture magnitude must also be determined to place the gesture on the gestural score. In Jung's work [2], this task was accomplished using a priori knowledge of the tier location and gesture magnitude. Subsequently, a method for statistically determining the necessary information was developed by Collins [3] for articulatory gestural scores.
As the goal of the current work is similar to that of Jung, that is, to determine the feasibility of generating gestural scores from a particular input representation and then evaluate their usefulness, we will adopt a similar approach to that of Jung. In Jung's work, the tier location and gesture magnitude for each oral gesture were specified explicitly. In the present research, a level of abstraction is added—the phoneme associated with each dominant oral gesture is specified, and the tier location and gesture magnitude appropriate to that phoneme is obtained from a lookup table. The values used in the phone-to-gesture procedure are shown in Table 4.2.

The present work employs two simplifications of the gestural scores to aid in implementation. These same simplifications exist in both Jung's and Collins' previous work with articulatory gestural scores, and the influence of these articulatory systems cause the simplifications to exist in the present acoustic system. The limitations introduced by the simplifications appear to be minor, particularly for the corpus being studied, but future work may endeavor to remove the limitations in the interest of completeness.

The first simplification is that all of the tiers on the generated gestural scores reflect when the appropriate articulators are "active". Gestures on the oral tiers indicate an intended constriction, gestures on the velic tier indicate that the nasopharyngeal port is opening to provide nasality, and gestures on the glottal tier indicate voicing. The current system does not provide a mechanism for explicit oral opening gestures, nor explicit glottal spreading gestures.

The second simplification involves the independence of the gesture timing among the various tiers. In a "true" gestural score, every tier should have its own independent timing. However, in the present system, all the gestures on the oral tiers are
<table>
<thead>
<tr>
<th>Phoneme</th>
<th>LA</th>
<th>LP</th>
<th>TTCD</th>
<th>TTCL</th>
<th>TBCD</th>
<th>TBCL</th>
<th>VEL</th>
<th>GLO</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>/b/</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/m/</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/f/</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/v/</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>/w/</td>
<td>0.8</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/t/</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>/d/</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/s/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>/z/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/n/</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>/l/</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/r/</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/f/</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>/h/</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/s/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/z/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.9</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>/n/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/l/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.9</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/r/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/l/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/e/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/a/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/æ/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/u/</td>
<td>0.8</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/u/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/o/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/ø/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/a/</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 4.2: Phoneme mappings for the tier information.
synchronized (forming a "gesture bundle"). As temporal decomposition finds only one target function for each underlying temporal event. This implies that the temporal decomposition target functions likely correspond to the gesture ensemble timing, instead of the timing of individual gestures. That is, the target functions correspond to the timing of the effective constriction [14]. While the present system cannot capture the subtlety of oral gestures slightly phased with one another, it can capture primary effects of interest such as overlap of gestures on the oral tract, or phasing between the oral, glottal and velic tiers. The present system captures both of these primary effects, as the temporal decomposition target functions overlap with their neighbors, and the oral and glottal tier have independent timing (the velic tier is unused in the present corpus). Therefore, it has been surmised that the simplifications are not overly limiting for the constrained corpus used in this study.

4.5 System Diagram

The block diagram of the acoustic-to-gestural-score system is shown in Figure 4.1. The inputs to the system are the acoustic data and the a priori list of phonemes. The acoustic data is the input that the system will operate on: the list of phonemes is knowledge required by the system at this time. Future work should endeavor to remove this a priori knowledge.

The acoustic data drives two of the blocks: the feature extraction block and the automatic segmenting block. The feature extraction block takes the input acoustic signal and extracts the filtered log-area parameters described in Section 4.1. The automatic segmenting block uses the input acoustic signal to generate hard segmenting boundaries (see Section 4.3). The output of both of these blocks, the filtered log-area
parameters and the hard segment boundaries, is used to drive the next block: seeded temporal decomposition.

As discussed in Section 4.2, seeded temporal decomposition needs hard segment boundaries in addition to the input signal data. In this system, the input data are provided from the feature extraction block as low-pass filtered log-area parameters, and the hard segmenting information is derived either from hand-labelling or from the automatic segmenting block. The seeded temporal decomposition takes these two pieces of information and generates the gesture timing information, in the form of the temporal decomposition target functions.

In addition to the gesture timing information, gesture tier information is needed to generate a gestural score. The tier information is assumed to conform to a canonical specification of gestures for each phoneme, and so can be retrieved from a simple lookup table, as described in Section 4.4.

Together, the tier and timing information are used to generate a gestural score. This occurs in the onset and durational analysis block, which operates as described in Section 4.4.

An example of a complete gestural score generated by this system is shown in Figure 4.15 along with a plot of the original acoustic waveform, the log-area parameters, and the seeded temporal decomposition target functions.

To evaluate the acoustic-to-gestural score system, gestural scores were generated on this system and compared with gestural scores derived from simultaneously recorded articulatory data. These experiments are detailed in the following chapter, along with additional experiments aimed at ascertaining the utility of the acoustic gestural scores.
Figure 4.15: Gestural score for the utterance "beak" (/bik/) generated by the acoustic-to-gestural score system. A plot of the original acoustic waveform, the log-area parameters and the seeded temporal decomposition target functions is shown for reference.
CHAPTER 5

EVALUATING THE GESTURAL SCORES GENERATED FROM ACOUSTIC DATA

The goal of this research is to investigate the generation of gestural scores from acoustic data. As discussed in the previous chapter, a system has been designed and implemented for performing such a task. This system provides a first step towards using acoustic data instead of articulatory data to generate gestural scores: to gauge the worth of this system we must ascertain the usefulness of those acoustic gestural scores.

In prior works, gestural score evaluation consisted of human and artificial neural network recognition tests. These tests were well suited to determine if the gestural score contained sufficient information for recognition. The present work conducts recognition tests while adding several new experiments. This chapter discusses the data used and experiments conducted to evaluate the generated acoustic gestural scores.
5.1 Data

Two sets of data. X-ray microbeam data and electropalatograph (EPG) data, are used in evaluating the utility of the acoustic gestural scores. Both data sets will be discussed in the following sections.

5.1.1 X-ray Microbeam Data

The first is the data set used by Jung and Collins in their work with articulatory gestural scores. This consists of simultaneously recorded articulatory and acoustic data for 55 consonant-vowel-consonant (CVC) tokens in the frame sentence Say a CVC\(_1\) of a CVC\(_2\) again. The consonants for the tokens are drawn from the set of English stop consonants /p t k b d g/ and the vowel is drawn from the set /i u a e a ə/ The nuclear accent was placed on the first CVC in every utterance. This data set is completely detailed in Section 3.1.1; however, there is one departure from the method of Jung and Collins which will be considered here.

In the previous work using this corpus, the tokens were extracted and cropped to 583 ms in length. For the present work, a longer token size of 833 ms is used. This guarantees that the entire CVC is always present, embedded in a portion of the surrounding schwas. Temporal decomposition typically finds five target functions (instead of the three target functions of previous work), but only the target functions pertaining to the CVC are retained. This ensures that the target functions corresponding to the CVC are well-formed, which results in a more accurate gestural score.
Consisting of both articulator and simultaneously recorded acoustic data, this data is used in the comparison and recognition experiments of Section 5.3 and Section 5.4.

5.1.2 Electropalatograph Data

The second corpus is a new data set recorded for the present research. It consists of simultaneously recorded electropalatograph and acoustic data.

The electropalatograph data consists of tongue contact locations on a custom mouthpiece fitted to the subject’s palate. By dividing the palate into place-of-articulation regions and tabulating the percent of area contacted within these regions, articulatory data for the constriction locations and degrees can be inferred. However, this data was not yet analyzed and was thus unavailable to test for the present research, so only the acoustic waveform is actually used.

The subjects are a male and a female native speaker of American English. The subjects are given a short time in which to become accustomed to the acrylic palate, after which they record the database in one sitting.

The database consists of 27 sentences, each spoken at five different speaking rates, resulting in 135 total utterances for each speaker. It is designed using the same frame sentence as the X-ray microbeam data: Say a \textit{TOKEN}(1) of a \textit{TOKEN}(2) again. However, the tokens are not limited to only \textit{CVC} tokens in the present corpus. Instead, the tokens form several classes, each composed of tokens which differ (largely) only in the starting consonant cluster. The consonant clusters consist of 1 to 3 consonants, with consonants appearing singly and within clusters, where possible. The intent is to provide similar tokens between which target functions can be easily
interchanged, inserted, or deleted for the resynthesis experiment (Section 5.6). For example, consider a class of tokens “raid”, “trade”, “sade”, “staid” and “strayed”. The target functions from each can manipulated to form other tokens in the class: e.g., removing the “r” from “strayed” yields “staid”, inserting an “s” at the beginning of “trade” gives “strayed”, etc. Some additional tokens are chosen to provide examples of the coarticulation phenomenon deletion. The complete list of tokens is shown in Table 5.1. The deleted phonemes are the vowels in the first unstressed syllable of each of the multi-syllabic tokens.

<table>
<thead>
<tr>
<th>accolade</th>
<th>gorilla</th>
<th>shade</th>
</tr>
</thead>
<tbody>
<tr>
<td>cape</td>
<td>grade</td>
<td>shred</td>
</tr>
<tr>
<td>clay</td>
<td>grill</td>
<td>slayed</td>
</tr>
<tr>
<td>courageous</td>
<td>laid</td>
<td>staid</td>
</tr>
<tr>
<td>crepe</td>
<td>raid</td>
<td>strayed</td>
</tr>
<tr>
<td>disclaimer</td>
<td>rill</td>
<td>tape</td>
</tr>
<tr>
<td>Galatians</td>
<td>sade</td>
<td>terrainium</td>
</tr>
<tr>
<td>gate</td>
<td>scape</td>
<td>Toledo</td>
</tr>
<tr>
<td>glade</td>
<td>scrape</td>
<td>trade</td>
</tr>
</tbody>
</table>

Table 5.1: Complete list of tokens used to compose the electropalatograph data set.

The five speaking rates are categorized as slowest, slow, normal, fast and fastest. The subject performs the rates as normal, slow, slowest, then repeats the utterance as normal, fast, fastest. This provides a consistent method for generating the five speaking rates.

This data is used in the coarticulation and resynthesis experiments of Section 5.5 and Section 5.6.
5.2 Gestural Score Similarity Metric

The current work provides a first step towards using acoustic data instead of articulatory data, so an obvious additional evaluation tactic would be to compare the acoustic gestural scores to the previously generated articulatory gestural scores.

A direct evaluation method would be a comparison of the acoustic gestural scores with their articulatory counterparts. At the present, no method exists for such a comparison. For this research, a priori knowledge is used for the gesture tier information: therefore, we only are interested in comparing the number of target functions and their relative locations. A method to compare gestural scores based on this criteria was implemented, and is described as follows.

Consider first the case of comparing the location and duration of a single gesture $g_1$ with the location and duration of a second gesture $g_2$. The percentage coverage of $g_1$ by $g_2$ would indicate how much of the time that $g_1$ was active that $g_2$ was also active. That is

$$\text{coverage}(g_1, g_2) = \frac{\text{overlap}(g_1, g_2)}{\text{length}(g_1)}$$

This equation for coverage is not commutative; that is, \(\text{coverage}(g_1, g_2) \neq \text{coverage}(g_2, g_1)\). Therefore, a more viable measure of the similarity of location and duration for two gestures would be the average overlap between the two gestures as a percentage of each gesture’s length

$$\text{similarity}(g_1, g_2) = \frac{1}{2} \left[ \text{coverage}(g_1, g_2) + \text{coverage}(g_2, g_1) \right]$$

or

$$\text{similarity}(g_1, g_2) = \frac{1}{2} \left[ \frac{\text{overlap}(g_1, g_2)}{\text{length}(g_1)} + \frac{\text{overlap}(g_2, g_1)}{\text{length}(g_2)} \right]$$
This measure ranges from 0 (when the gestures have no overlap) to 1 (when the gestures span identical times). To find the similarity of two gestural scores $G_1$ and $G_2$, take the average of the best pairwise similarities between all the gestures $g_1(i)$ in $G_1$ with all of the gestures $g_2(k)$ in $G_2$. That is,

$$similarity(G_1, G_2) = \frac{1}{n} \sum_{i=1}^{n} \max_k (similarity(g_1(i), g_2(k))) \quad (5.4)$$

where $n$ and $m$ are the number of gestures in $G_1$ and $G_2$ respectively and $k$ ranges from 1 to $m$. Again, this number will range from 0 to 1, providing a quantitative scale with which to compare the timings of two gestural scores. Tests using this scale to evaluate the acoustic gestural scores are detailed in the following section.

5.3 Comparison With Articulatory Gestural Scores

The current system is the most recent step in the achieving the long term goal of an automatic acoustic-to-gestural score system. Previous work has concentrated on producing accurate gestural scores from articulatory data. Therefore, the most straightforward method to evaluate the acoustic gestural scores is to directly compare them to their corresponding "gold-standard" articulatory gestural scores. One approach to this comparison is the similarity score detailed in the previous section, with the results of that comparison shown previously in Table 4.1 and repeated in Table 5.2 for convenience. A second comparison approach is to compare the relationship between the locations of hand-labeled phonemes and the target functions resulting from seeded temporal decomposition. Using the simplistic assumption that there is a one-to-one correspondence between oral gesture bundles and phonemes, it would follow that we would desire to have a one-to-one correspondence between phonemes and
Table 5.2: Similarity scores between gestural scores for each parameter set and the articulatory gestural score.

<table>
<thead>
<tr>
<th>Parameter Set</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50 Hz</td>
</tr>
<tr>
<td>Autocorrelation coefficients</td>
<td>0.678</td>
</tr>
<tr>
<td>Area coefficients</td>
<td>0.670</td>
</tr>
<tr>
<td>Cepstral coefficients</td>
<td>0.725</td>
</tr>
<tr>
<td>Log area parameters</td>
<td>0.702</td>
</tr>
<tr>
<td>Log area ratios</td>
<td>0.715</td>
</tr>
<tr>
<td>Line spectral frequencies</td>
<td>0.694</td>
</tr>
<tr>
<td>Reflection coefficients</td>
<td>0.676</td>
</tr>
</tbody>
</table>

The temporal decomposition target functions. Note that the glottal tier, calculated separately, is independent of phoneme location.

To perform the comparison, each hand-labeled phoneme is examined for how many target functions span the corresponding time—zero, one or two (no phonemes were spanned by more than two target functions). A phoneme which has zero target functions corresponding to it constitutes a “missed” phoneme. Phonemes with two target functions spanning them often result from phonemes that have two different acoustic consequences, such as the closure and release for stop consonants.

The results of this examination for each of the considered parameters sets is shown in Table 5.3. Tables 5.4 and 5.5 show the breakdown of misses and doubles by position within the CVC token. As missing a target function completely is more severe than finding two target functions, log-area parameters again appear to perform the best of the acoustic parameter sets.

It is interesting to note that more than half of the errors occur in the initial position, especially for the two-to-one mappings. Often times, the transition between
Table 5.3: Percent of temporal decomposition target functions corresponding to each hand labeled phoneme for each of the parameter sets of interest.

<table>
<thead>
<tr>
<th>Parameter Set</th>
<th>Target functions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>zero</td>
</tr>
<tr>
<td>Articulatory parameters</td>
<td>1 of 165 (1%)</td>
</tr>
<tr>
<td>Cepstral coefficients</td>
<td>5 of 165 (3%)</td>
</tr>
<tr>
<td>Log area parameters</td>
<td>2 of 165 (1%)</td>
</tr>
<tr>
<td>Log area ratios</td>
<td>8 of 165 (5%)</td>
</tr>
</tbody>
</table>

Table 5.4: Breakdown of missed phonemes by position within the CV'C.

<table>
<thead>
<tr>
<th>Parameter Set</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>initial</td>
</tr>
<tr>
<td>Articulatory parameters</td>
<td>1 of 1 (100%)</td>
</tr>
<tr>
<td>Cepstral coefficients</td>
<td>4 of 5 (80%)</td>
</tr>
<tr>
<td>Log area parameters</td>
<td>0 of 2 (0%)</td>
</tr>
<tr>
<td>Log area ratios</td>
<td>4 of 8 (50%)</td>
</tr>
</tbody>
</table>

the first carrier schwa and the initial consonant is modeled with a separate target function. This target function may extend into the initial consonant, also modeling the closure (and pre-voicing for voiced consonants). A second target function results from the burst. This is particularly true for the log-area parameters, where almost all of the two-to-one instances (7 or 9) occur with voiced consonants, even though unvoiced consonants account for more than half of the initial consonants in the corpus.

Figures 5.1 through 5.3 provide an example of a temporal decomposition of the parameter sets, as well as illustrate the cases where a one-to-one mapping between phonemes (and by our assumption, gestures) and target functions is not attained. The
plots are for both articulatory data and log-area parameters, and the vertical lines indicate the seeded boundaries provided to temporal decomposition by the automatic segmenting. Asterisks show the intervals where a one-to-one mapping is not obtained.

These results indicate that the acoustic gestural scores are accurately capturing the locations of the phonemes, and by our assumption, the gestures. To further verify this accuracy, recognition tests were also performed. The recognition tests are the subject of the following section.

### 5.4 Recognition Tests

As another comparison of the acoustic and articulatory gestural scores, human recognition tests similar to that of Jung and Collins (see Section 3.1.5) were performed on the gestural scores generated from the log-area parameters. For these tests, the gestural scores with zero-to-one or two-to-one mappings between the underlying phonetic events and the target functions were removed. In this way, the accuracy of the gestural score generation process does not bias whether or not the score contains sufficient information for recognition.

Table 5.5: Breakdown of doubled phonemes by position within the CVC token.

<table>
<thead>
<tr>
<th>Parameter Set</th>
<th>Position</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>initial</td>
<td>vowel</td>
<td>final</td>
</tr>
<tr>
<td>Articulatory parameters</td>
<td>8 of 14 (57%)</td>
<td>5 of 14 (36%)</td>
<td>1 of 14 (7%)</td>
</tr>
<tr>
<td>Cepstral coefficients</td>
<td>5 of 9 (56%)</td>
<td>3 of 9 (33%)</td>
<td>1 of 9 (11%)</td>
</tr>
<tr>
<td>Log area parameters</td>
<td>9 of 17 (52%)</td>
<td>4 of 17 (24%)</td>
<td>4 of 17 (24%)</td>
</tr>
<tr>
<td>Log area ratios</td>
<td>12 of 27 (44%)</td>
<td>11 of 27 (41%)</td>
<td>4 of 27 (15%)</td>
</tr>
</tbody>
</table>
Figure 5.1: Examples of missed target functions for both articulatory data and log-area parameters. The tokens are “beeg” /big/ and “teeg” /tig/. Asterisks mark the seeded interval in which no target function was found.
Figure 5.2: Examples where one target function is found for each underlying event, both articulatory data and log-area parameters. The tokens are “derg” /dəɡ/ and “beeg” /bɪɡ/.
Figure 5.3: Examples of doubled target functions for both articulatory data and log-area parameters. The tokens are "curd" /kɔːd/ and "gerg" /gɔːɡ/. Asterisks mark the seeded interval in which two target functions were found.
Human recognition tests were conducted in the same way as those performed by Jung and Collins. Two phonetically trained students at The Ohio State University were given brief training in how to read the gestural score, and were shown a few example scores with their answers. After training, each student was asked to identify all the CVC tokens from their gestural scores. During testing, no feedback was given. However, both students were able to correctly identify all of the tokens, as shown in Table 5.6.

<table>
<thead>
<tr>
<th>Position</th>
<th>Subject 1</th>
<th>Subject 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>38 of 38 (100%)</td>
<td>38 of 38 (100%)</td>
</tr>
<tr>
<td>Vowel</td>
<td>38 of 38 (100%)</td>
<td>38 of 38 (100%)</td>
</tr>
<tr>
<td>Final</td>
<td>38 of 38 (100%)</td>
<td>38 of 38 (100%)</td>
</tr>
</tbody>
</table>

Table 5.6: Human viewer recognition accuracy for the acoustic gestural scores.

Again, the results show the acoustic gestural scores performing extremely well: competitive with articulatory gestural scores and retaining sufficient phonetic information to perform recognition.

In addition to the comparison and recognition tests, additional tests were performed to determine whether or not the acoustic gestural scores were able to capture more complex phonetic effects predicted by gestural phonology, such as coarticulation, and whether the overlapping segments can be interchanged, providing a novel approach to synthesis. These tests are the topics of the remaining sections of this chapter.
5.5 Application to Coarticulation Phenomena

As conversational speech is spoken at an increasingly faster rate, some phonemes shorten in length until they eventually have little or no acoustic consequence. The coarticulation phenomena of deletion. However, even though there might be no acoustic evidence of the phoneme, gestural phonology predicts that the gestures corresponding to a deleted phoneme are still present. Instead of the deleted phoneme’s gesture being deleted, it is covered by the gestures of the adjacent phonemes. The gestures of the adjacent phonemes are overlapping the deleted phoneme’s gesture, hiding it. The deleted phoneme’s gesture is not removed, as the acoustic signal might seem to indicate—it is merely hidden. Note that even when there appears to be no direct acoustic evidence of the removed phoneme, cues in the neighboring phonemes allow listeners to correctly identify the correct word in the overwhelming majority of the cases [8, 9].

For illustration, Figure 5.4 shows a hypothetical set of gestures, and the resulting acoustics. As the speaking rate increases, the gestures become more and more overlapped, resulting in a shorter and shorter acoustic consequence. A traditional “beads on a string” phonetic representation cannot accurately capture this effect. With such as phonetic representation, a phoneme is either present or not—differing “degrees” of deletion cannot be captured. We would like to demonstrate that our acoustic gestural scores can, at least in part, capture this effect.

For this experiment, the data of Section 5.1.2 are used. As mentioned previously, a part this data set is designed to demonstrate deletion. Six tokens have been chosen for this purpose: “Galatians”, “Toledo”, “courageous”, “accolade”, “gorilla”
Figure 5.4: Hypothetical gestures demonstrating hiding. The top plot shows gestures becoming increasingly overlapped as the speaking rate increases. The bottom plot shows the resulting acoustics.
and “terrainium”. By having the speakers perform each utterance at five different speaking tempos, we have examples of the gradual onset of deletion.

To determine if the generated acoustic gestural scores can capture this effect, we considered the deleted phoneme. For all of the tokens, the deleted phoneme is the vowel in the first unstressed syllable. We would like for the gesture bundle corresponding to this phoneme to still be detected by temporal decomposition, even at the fastest speaking rate. Additionally, the phoneme’s gesture should become increasingly overlapped by the gestures of the neighbor phonemes as the speaking rate increases.

The first condition can be demonstrated by simply tabulating whether or not temporal decomposition finds a target function for the deleted phoneme. This tabulation occurs for each of the speaking rates, across all six of the tokens. Table 5.7 shows the occurrence of the deleted phonemes at each of the speaking rates. The table shows

<table>
<thead>
<tr>
<th>Speaking Rate</th>
<th>Percentage Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slowest</td>
<td>6 of 6 (100%)</td>
</tr>
<tr>
<td>Slow</td>
<td>6 of 6 (100%)</td>
</tr>
<tr>
<td>Normal</td>
<td>6 of 6 (100%)</td>
</tr>
<tr>
<td>Fast</td>
<td>4 of 6 (66%)</td>
</tr>
<tr>
<td>Fastest</td>
<td>3 of 6 (50%)</td>
</tr>
</tbody>
</table>

Table 5.7: Percentage of the deleted phonemes detected by temporal decomposition at each speaking tempo.

that temporal decomposition often locates the deleted phonemes, even at the fastest speaking rate. Figure 5.5 shows the full result of temporal decomposition for the token “a Galatians oʃfʃ” (/a ɡaleʃʃənz ə/) at the normal and fastest speaking rates.
Note that the two figures are plotted on different time scales. Figure 5.7, shown later, displays on one time scale all of the target functions for the five speaking rates of one token. Notice that the target function corresponding to the deleted schwa is still present, even at the fastest speaking rate; it is the third target function at every speaking rate except for the slowest, where it is the fourth target function. It has been marked with an asterisk to aid identification. Figure 5.6 shows the full result of temporal decomposition for the token “a gorilla oʃf” (/ə gəɹɪlə ə/) at the normal and fastest speaking rates. Here, temporal decomposition failed to recover a gesture corresponding to the deleted phoneme at the fastest speaking rate. Instead, the deleted phoneme and following phoneme are spanned by only one gesture. Another error appears to be that the target function corresponding to the stressed /ɪ/ is completely overlapped by the neighbors. Again, note that the two figures are plotted on different time scales.

The second way to test if the present method can model hiding is to investigate if the target function corresponding to the deleted phoneme becomes more overlapped by its neighbors as the speaking rate increases. This is examined by calculating the length of the gesture corresponding to the deleted phoneme at each speaking rate. This length is normalized by the length of the gesture at the normal speaking rate. If the underlying gesture is progressively hidden and not simply shrunken in size, the length of the target function should vary considerably less than that of the underlying acoustic segment.

Table 5.8 shows the normalized lengths of the deleted phonemes at each of the speaking rates. Only tokens in which the deleted phoneme was located by temporal decomposition are considered. The table shows that the underlying gestures are not
Figure 5.5: Temporal decomposition results for “a Galatians o[f]” (/ə gələtənz ə/) at the normal and fastest speaking rates.
Figure 5.6: Temporal decomposition results for “a gorilla o[f] (ə ɡərlə ə/) at the normal and fastest speaking rates.
<table>
<thead>
<tr>
<th>Speaking Rate</th>
<th>Normalized Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slowest</td>
<td>1.08</td>
</tr>
<tr>
<td>Slow</td>
<td>1.18</td>
</tr>
<tr>
<td>Normal</td>
<td>1.00</td>
</tr>
<tr>
<td>Fast</td>
<td>0.88</td>
</tr>
<tr>
<td>Fastest</td>
<td>1.01</td>
</tr>
</tbody>
</table>

Table 5.8: Lengths of the gestures corresponding to deleted phonemes for each of the speaking rates.

simply scaled as the speaking rate increases. The slowest and fastest rates are less than 10% different from the length at the normal speaking rate, and the maximum difference is 18%. This is the first piece of evidence that the gestures are not actually shrunk in length, but are rather more heavily overlapped by the neighboring gestures.

For the second piece of reinforcing evidence, the percentage overlap of the reduced phoneme by its neighbors is shown in Table 5.9. Again, if the underlying gesture is progressively hidden and not simply shrunk in size, the hidden target function should show progressive overlap by its neighbors. As before, only tokens in which the deleted phoneme was located by temporal decomposition are considered. The table shows that the underlying gestures are in fact more heavily overlapped by their neighbors as the speaking rate is increased. Coupled with the results of Table 5.8, this is evidence that supports our second premise that the underlying gestures are hidden, and not simply scaled downward in length. Figure 5.7 displays only the target functions for the same token at each of the five speaking rates. The target function corresponding to the reduced vowel has been marked with an asterisk to aid
<table>
<thead>
<tr>
<th>Speaking Rate</th>
<th>Percentage Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slowest</td>
<td>74%</td>
</tr>
<tr>
<td>Slow</td>
<td>78%</td>
</tr>
<tr>
<td>Normal</td>
<td>81%</td>
</tr>
<tr>
<td>Fast</td>
<td>82%</td>
</tr>
<tr>
<td>Fastest</td>
<td>89%</td>
</tr>
</tbody>
</table>

Table 5.9: Percentage overlap of the gestures corresponding to deleted phonemes for each of the speaking rates.

identification. From this figure, there is a trend that the target function of interest is not shrinking in length, but rather is becoming progressively more overlapped by its neighbors.

These three experiments provide evidence that the generated acoustic gestural scores are capturing the coarticulation effect of deletion for our data. As the data set for this study is extremely small, a larger study should be conducted for stronger conclusions.

5.6 Resynthesis

Despite the fact that previous stages in this ongoing research in generating gestural scores have concentrated on recognition tests, recognition is not the only area of speech research that we expect would benefit from using the gestural score. Synthesis and possibly coding would additionally show benefits. Therefore, as a final test of our system, synthetic tokens were created by manipulating the target functions for a token and then performing resynthesis. The manipulations take on three forms:
Figure 5.7: Target functions for "a Galatians o[f]" (/a galef'sanz a/). Notice that the target function corresponding to the deleted schwa (marked with an asterisk) is still present even at the fastest speaking rate. The acoustic waveform is shown for reference.
• Target functions are removed from an utterance. For example, removing the target functions corresponding to the "r" in "crepe" to form "cape" (/kreip/ → /kelp/).

• Target functions are inserted into an utterance. For example, adding the target function corresponding to a "t" from another utterance to "sade" to form "staid" (/seid/ → /steid/).

• Phoneme deletion is simulated by covering a target function with its neighbors. For example, hiding the first vowel in "gorilla" to make the first part of it sound like "grill" (/gərɪlə/ → /grtl(ə)/).

The resynthesis procedure starts when the acoustic signal is parameterized in the feature extract block of the system for producing acoustic gestural scores. The analysis provides LPC filter coefficients for an all-pole filter describing the acoustic signal. Considering speech as the result of a source (the glottis) and a filter (the vocal tract), the LPC filter values provide the time-varying description of the vocal tract filter. The source, or glottis input to the vocal tract filter, can be inferred by inverse filtering the speech waveform with the LPC filter. The result of the inverse filtering, called the residual, is the signal that generates the original speech waveform when passed forward through the LPC filter.

At this point, the acoustic-to-gestural-score system proceeds normally. The extracted log-area parameters are passed to the seeded temporal decomposition block. After temporal decomposition is performed on the log-area parameters, the residual signal is associated with the resulting target functions for the duration that each target function is active. This residual fragment will stay with the target function as
it is manipulated. Later, the residual fragments will enable a new residual (source) to be created for the manipulated tokens. The target vector for each target function also stays with the target function as it is manipulated.

Once all of the original gestural scores are created from the data, the target functions are manipulated to create the new tokens. New log-area parameter traces are manufactured using the new target function configuration and their target vectors, in the same way that the log-area parameters would normally be reconstructed with unmodified target functions. This inherently provides “blending” in the log-area parameters where the target functions overlap. To construct a new residual, the residual fragments from each of the target functions are added together, using the relative magnitudes of the target functions to weight the residual fragments where the target functions overlap. In this way, the residual is also blended. Our hope is that by creating gradual transitions between each segment, instead of hard boundaries, the resynthesized speech will sound more natural.

Figures 5.8 and 5.9 depict the data for resynthesizing a token with an inserted target function. Figure 5.8 shows the original data for the two tokens, “cape” (/kəp/) and “scape” (/skəp/). The target function corresponding to the /s/ in “scape” is moved to the beginning of the “cape” token, and the resynthesized token is shown in Figure 5.9.

Figures 5.10 and 5.11 depict the data for resynthesizing a token with a removed target function. Figure 5.10 shows the original data for the token “grill” (/grɪl/), along with the original data for a token that is the same as the token that we will create, “rill” (/rl/). The target function corresponding to the /ɡ/ in “grill” is removed to form “rill”, with the resynthesized token shown in Figure 5.11.
Figure 5.8: Temporal decomposition data for the original tokens “cape” (/kelp/) and “scape” (/skelp/). The target function corresponding to the /s/ in “scape” (marked with an asterisk) is moved to the beginning of “cape”, to form a synthesized “scape” (see Figure 5.9).
Figure 5.9: Temporal decomposition data and spectrogram for the synthesized token "scape", an example of resynthesizing a token after inserting a target function.
Figure 5.10: Temporal decomposition data for the original tokens “grill” (/gril/) and “rill” (/rill/). The target function corresponding to the /g/ in “grill” (marked with an asterisk) is removed to form a synthesized “rill” (see Figure 5.11).
Figure 5.11: Temporal decomposition data and spectrogram for the synthesized token “rill”, an example of resynthesizing a token after removing a target function.
As mentioned above, three different manipulations are performed: removing target functions, inserting target functions, and hiding target functions. In addition, each of the original (unmodified) target function configurations are resynthesized as “control” tokens. These original tokens have errors introduced by filtering the log-area parameters and from reconstruction from the temporal decomposition target functions, and therefore provide a best case.

Listeners are provided with each of the tokens, one at a time, and asked merely to record what they perceive. Table 5.10 shows the percent of the tokens correctly recognized, for both the control tokens and the tokens for each of the three types of manipulation used in this experiment.

<table>
<thead>
<tr>
<th>Modification</th>
<th>Subject 1</th>
<th>Subject 2</th>
<th>Subject 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>6 of 7 (86%)</td>
<td>6 of 7 (86%)</td>
<td>6 of 7 (86%)</td>
</tr>
<tr>
<td>Removal</td>
<td>4 of 4 (100%)</td>
<td>3 of 4 (75%)</td>
<td>2 of 4 (50%)</td>
</tr>
<tr>
<td>Insertion</td>
<td>4 of 4 (100%)</td>
<td>4 of 4 (100%)</td>
<td>4 of 4 (100%)</td>
</tr>
<tr>
<td>Hiding</td>
<td>1 of 1 (100%)</td>
<td>1 of 1 (100%)</td>
<td>1 of 1 (100%)</td>
</tr>
</tbody>
</table>

Table 5.10: Listener identification accuracy for resynthesized tokens of each of three types of modifications.

Listener identification accuracy was surprising, as the resynthesized tokens were as recognizable as the control tokens. Despite high recognition accuracy, listener comments indicate that several of the tokens sounded quite distorted. However, listeners reported no perceived increase in distortion for the tokens resynthesized from manipulated target functions. The primary sources of distortion are the low-pass filtering of the log-area parameters and the reconstruction of the log-area parameters.
from the target functions; the additional distortion from manipulating the target functions is small in comparison.

This experiment demonstrates that the acoustics-to-gestural score system could supply a viable method for performing a more natural sounding concatenative-style synthesis.
CHAPTER 6

CONCLUSIONS

Traditional phonetic representations model speech as a concatenation of discrete, static, invariant elements. The reality of speech production disagrees with this model, and as a result, these representations experience difficulty modeling coarticulation phenomena such as assimilation, deletion and epenthesis.

Browman and Goldstein have suggested a description of the speech signal using temporally overlapping gestures on multiple channels. This description is called the gestural score, and it can capture speech phenomena such as those of coarticulation. The gestural score model holds the promise of providing a more unified and parsimonious description of the speech signal, naturally linking the linguistic and physical domains of the signal. However, one of the drawbacks of this model is the lack of methods to infer the abstract gestural score from physical signals that can be easily measured.

The Ohio State University has been conducting an on-going research project to investigate the gestural score as a phonetic representation, and to develop methods to objectively generate gestural scores. It is our hope that systems based on such a model could provide improvements in recognition, synthesis and speech coding.
6.1 Summary of Current Work

The current work is the most recent step in this on-going project. The goal of this research is to investigate the generation of gestural scores from acoustic data, and to ascertain the worth of the resulting gestural scores. The tasks performed in achieving this goal included:

- **Design and implement an acoustic-to-gestural-score system.**

  - A complete acoustic-to-gestural-score system was designed and implemented using MATLAB, C, and ESPS Waves. A block diagram can be found in Figure 4.1.

  - As part of the development, a study was conducted to determine the most appropriate input parameterization. The study found log-area parameters low-pass filtered to 10 Hz to be the most suitable for the task.

  - Also as part of the development, modifications were introduced to temporal decomposition so as to allow for dynamically specified initial window durations. With these modifications, temporal decomposition is provided with "hard" segment boundaries, which it then "softens" into overlapping segments. The performance of the temporal decomposition algorithm was improved, with consequent improvement in the performance of the overall system.

One shortcoming of this system is the crude segmenter used for the "hard" boundary locations. Another is the *a priori* knowledge used for the gesture tier placement. The boundary locations could be easily improved by implementing
a better segmenting algorithm: the tier placement has been left as a future exercise.

- **Verify the correctness of the generated acoustic gestural scores.**

Two tests were conducted to verify the correctness of the generated acoustic gestural scores:

- The first test consisted of comparisons between the acoustic gestural scores, the articulatory gestural scores of the previous works of Jung and Collins, and hand-labeled phoneme locations. The acoustic gestural scores were found to be at least as accurate as their articulatory counterparts.

- The second test consisted of the human recognition tests used in the previous works of Jung and Collins. The acoustic gestural scores again performed extremely well, demonstrating that they contained sufficient information for accurate recognition.

- **Investigate the usefulness of the generated acoustic gestural scores.**

Data was collected and two experiments were conducted to ascertain the usefulness of the generated acoustic gestural scores:

- The data consisted of two speakers reading 27 tokens at five speaking rates. The tokens and speaking rates were chosen to provide examples of the coarticulation effect deletion, and to create a pool of similar tokens between which phonemes could be easily interchanged for resynthesis.
- The first experiment was an investigation into whether the acoustic gestural scores could capture the coarticulation effect of deletion. The acoustic gestural scores located deleted phonemes at least 50% of the time, and the corresponding gestures were found to be relatively unscaled and more greatly overlapped by the neighboring gestures compared to their undeleted counterparts. This is in agreement with the predictions of gestural phonology, and provides evidence that the acoustic gestural scores can capture at least some aspect of coarticulation. Unfortunately, this investigation was hampered by an extremely small data set. Further studies with a larger data set should be conducted to provide stronger conclusions.

- The second experiment consisted of resynthesis of speech tokens from temporal decomposition target functions which had been manipulated. The manipulations involved removing target functions, inserting target functions, and hiding target functions to emulate deletion. Listener tests found the resynthesized tokens to be discernible most of the time.

### 6.2 Future Work

The current work provides a strong first step into generating gestural scores from acoustics. As this work is built upon its predecessors, future work should endeavor to solidify and broaden the current work. Several improvements remain to be investigated, including:

- Improve the crude segmentation, and replace it with any of several algorithms available in the literature that perform segmentation.
• Remove the *a priori* gestural tier information. This problem needs to be studied and a method devised from those studies.

• Extend the operation of the gestural score system from tokens to whole utterances.

• Conduct comprehensive studies of the utility and limitations of the acoustic gestural score using larger data sets.

These tasks are the next steps in achieving the long-term goal of a completely automatic acoustic-to-gestural score system.
BIBLIOGRAPHY


