FSM-BASED PRONUNCIATION MODELING USING ARTICULATORY PHONOLOGICAL CODE

BY

CHI HU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Adviser:

Associate Professor Mark A. Hasegawa-Johnson

Abstract

According to articulatory phonology, the gestural score is an invariant speech representation. Though the timing schemes, i.e., the onsets and offsets, of the gestural activations may vary, the ensemble of these activations tends to remain unchanged, informing the speech content. "Gestural pattern vector" (GPV) has been proposed to encode the instantaneous gestural activations that exist across all tract variables at each time. Therefore, a gestural score with a particular timing scheme can be approximated using a GPV sequence.

In this work, we propose a pronunciation modeling method that uses a finite state machine (FSM) to represent the invariance of a gestural score. Given the "canonical" gestural score of a word with a known activation timing scheme, the plausible activation onsets and offsets are recursively generated and encoded as a weighted FSM. An empirical measure is used to prune out gestural activation timing schemes that deviate too much from the "canonical" gestural score. Speech recognition is achieved by matching the recovered gestural activations to the FSM-encoded gestural scores of different speech contents. In particular, the observation distribution of each GPV is modeled by an artificial neural network and Gaussian mixture tandem model. These models are used together with the FSM-based pronunciation models in a Bayesian framework.

We carry out pilot word classification experiments using synthesized data from one speaker. The proposed pronunciation modeling achieves over 90% accuracy for a vocabulary of 139 words with no training observations, outperforming direct use of the "canonical" gestural score.

To my parents, for their love and support

Acknowledgments

This research is funded by NSF grant IIS-0703624. This project would not have been possible without the support of many people. Many thanks to my adviser, Mark Hasegawa-Johnson, who enlightened me about the idea, helped me with great patience on both algorithm and experiment design, and read my paper several times for revision. Also thanks to my research groups, the Statistical Speech Technology group, and Landmark and Prosody Speech Recognition group. All the members in the groups gave me precious suggestions and advice. Thanks to Vikramjit Mitra and Hosung Nam for assistance with the dataset. And finally thanks to my parents and numerous friends who always offered me support and love.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AF	Articulatory Feature
ANN	Artificial Neural Network
BN	Bayesian Network
CGS	Canonical Gestural Score
DBN	Dynamic Bayesian Network
EM	Estimation Maximization
FSM	Finite State Machine
FST	Finite State Transducer
GLO	Glottis Constriction Degree
GMM	Gaussian Mixture Model
GPV	Gestural Pattern Vector
GR	Gestural Regiment
HMM	Hidden Markov Model
IGR	Intergestural Coupling Relationships
LA	Lip Aperture
LP	Lip Protrusion
MFCC	Mel-Frequency Cepstral Coefficient
MLP	Multilayer Perceptron
PCA	Principal Component Analysis
TBCD	Tongue Body Constriction Degree

TBCL	Tongue Body Constriction Location
TTCD	Tongue Tip Constriction Degree
TTCL	Tongue Tip Constriction Location
VEL	Velum Constriction Degree

Chapter 1

Introduction

The standard approach for automatic speech recognition assumes the speech signal is represented as a concatenation of phones [1]. Under this assumption, current state-of-the-art speech recognition systems work much better for carefully articulated speech, such as broadcast news, than for conversational speech. Speech recognition presents major challenges of not only acoustic variability due to phonetic contexts, but pronunciation variation owing to speech reduction and coarticulation [2]. Different approaches have been proposed for phone-based pronunciation modeling [3, 4], but limited by the coarseness of the phones [1].

1.1 Articulatory Phonology and Speech Gesture

Though relatively less explored for speech recognition, speech production knowledge has been studied to explain speech phonology. In particular, articulatory phonology [5, 6] uses speech gestures as the basic units of phonological contrast, which are characterizations of discrete, physically real events that unfold during the speech production process. Articulatory phonology attempts to describe lexical units in terms of these events and their interrelations, which means that *gestures* are basic units of contrast among lexical items as well as units of articulatory action. Phonology is the set of relations among physically real events, a characterization of the systems and patterns that these events (the gestures) enter into.

The consequences of gestures can be observed in the movements of the speech articulator. Gestures consist of the formation and release of constrictions in the vocal tract, and they are defined as dynamical control regimes for constriction actions at eight different constriction *tract variables* rather than individual articulators. Tract variables dynamically characterize a di-

mension of vocal tract constriction, by which the articulators that contribute to the movement are organized into a coordinative structure. For example, the tract variable of tongue tip constrict location is affected by the action of three articulators: the tongue tip, the tongue body and the jaw.

The tract variables and their component articulators are displayed in Figure 1.1. As shown in Figure 1.1, tract variables consist of five constriction degree variables—lip aperture (LA), tongue body (TBCD), tongue tip (TTCD), velum (VEL), and glottis (GLO)—and three constriction location variables—lip protrusion (LP), tongue tip (TTCL), tongue body (TBCL). Beside the fact that each tract variable involves its corresponding articulators, some articulators can be shared by different gestures as well. For example, lip protrusion, lip aperture, tongue body and tongue tip share "jaw" as their common articulator.

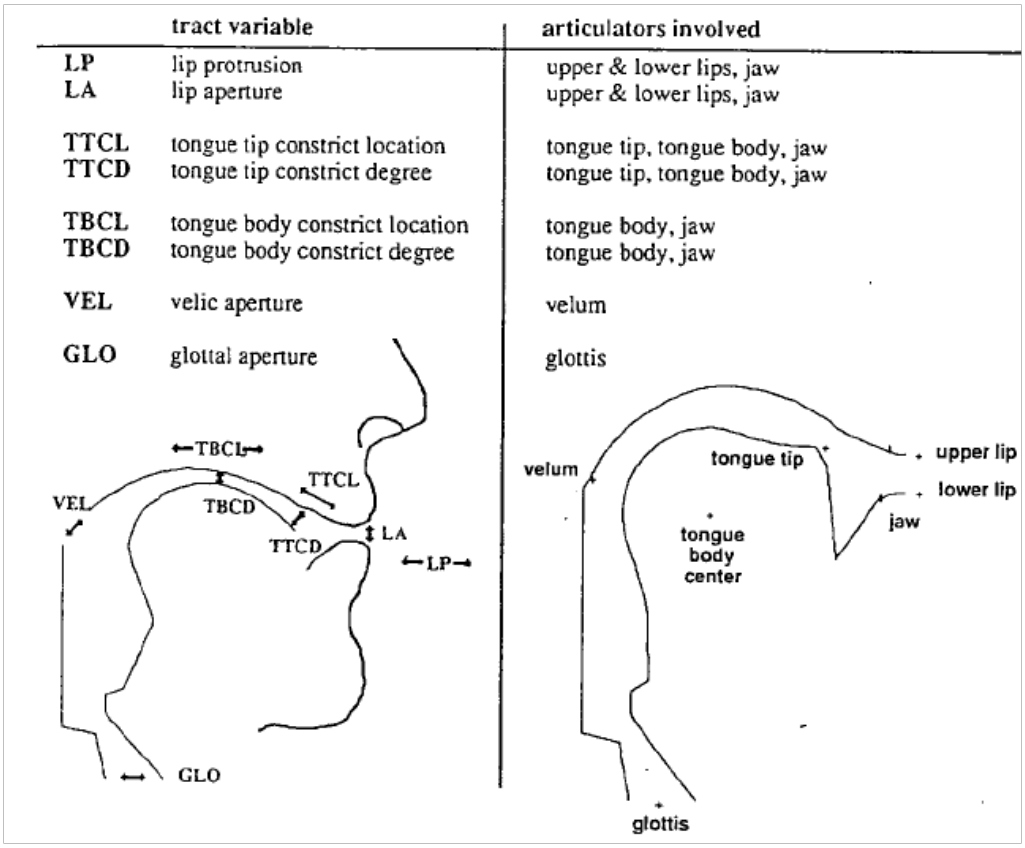

Figure 1.1: Tract variables and associated articulators

For a given constriction gesture, the activation interval, the timing scheme

(onset and offset times) and the dynamic parameters (target/stiffness/damping) are represented in a *gestural score*. Gestural score not only specifies the temporal activation intervals and dynamic parameter specifications for each individual gesture, but also illustrates the overlap timing patterns among the gestures in an utterance as well. An example of this gestural score for "about" is shown in Figure 1.2.

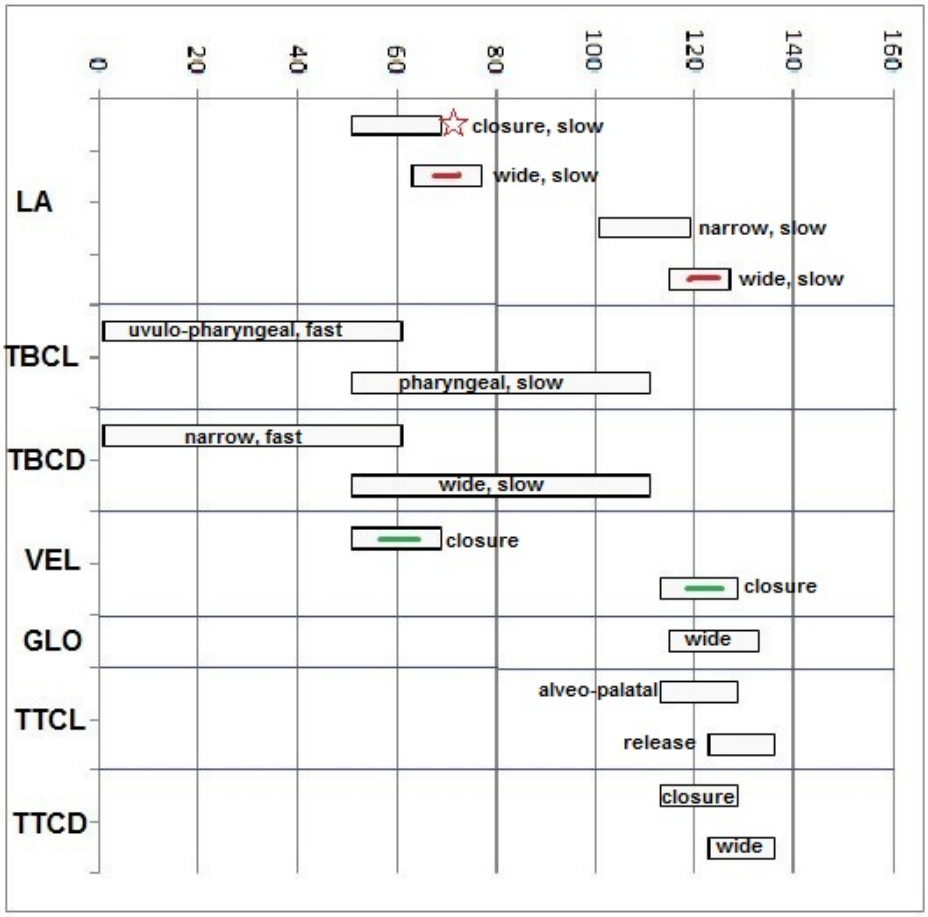

Figure 1.2: CGS of "about" in TADA

As we can see in the above gestural score, the primitive units for characterizing an utterance are no longer phonemes but constricting actions. This also shows the key to articulatory phonology is capturing both cognitive and physical, discrete and continuous features of speech. Unlike traditional units such as segments or phonemes, which occupy pre-allocated time slots, gestures are action units that are intrinsically allowed to overlap with one

another in time.

In addition, the ensemble of the gestures with their dynamic parameters is distinctive to speech content and is referred to as the *invariance of the gestural score*. However, gestures can be modulated in time and space as a function of concurrent gestures or prosodic context, so shifts in the relative timing of the gestures can cause coarticulated or reduced speech when they overlap in time.

1.2 Gestural Pattern Vectors

Zhuang et al. [7] proposed the instantaneous "gestural pattern vector" (GPV) to encode gestural activation information across tract variables in the gestural score at a given time. Shown in Figure 1.3, GPV is defined by the constriction targets and stiffness of gestural activations existing at a particular time across all tract variables. This quasi-atomic unit set is used to represent gestural score, which includes the gestural activation information active at the current time frame. As a result, gestural score can be represented by a sequence of GPVs. Given the invariance of the gestural score, we can obtain the ensemble of gestures for a particular word by recognizing the GPV sequence.

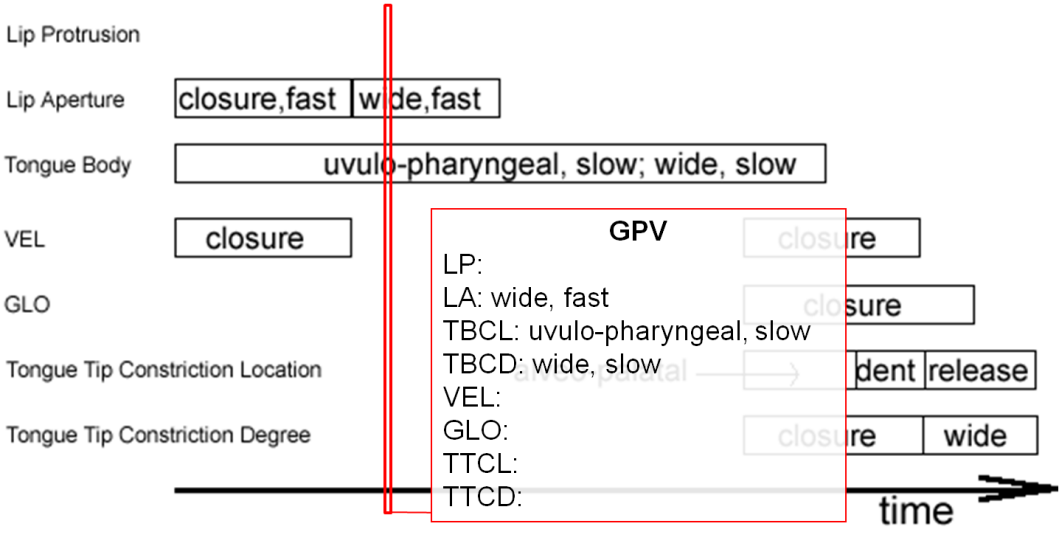

Figure 1.3: GPV of "but" defined on one frame

In order to explicitly modulate the GPV in different types, the cardinalities of the non-null settings of target and stiffness for each tract variable are defined in Table 1.1. There is also a value "null" for each tract variable in the table, which denotes no gestural activation for that tract variable. Note that the number of different GPVs is much smaller than the product of the cardinalities of constriction targets and stiffness in all tract variables, due to the correlation of gestural activation across different tract variables. These cardinalities are determined under the rules that they are not too high to be used in defining a relatively small gestural activation, and they preserve the most important distinctions corresponding to human perception of the language [8].

Table 1.1: Cardinalities of the non-null targets and stiffness of eight tract variables

TV	Target	Stiffness	TV	Target	Stiffness
LP	2	2	TBCD	5	2
LA	5	2	TBCL	4	2
VEL	2	1	TTCD	4	1
GLO	2	1	TTCL	4	1

1.3 Speech Production Based Speech Recognition

Several methods have been proposed for speech recognition using speech production knowledge. King et al. [1] gave a comprehensive overview of speech production knowledge in automatic speech recognition.

In [9], Deng et al. developed an integrative feature-based general statistical framework for automatic speech recognition via minimal units of speech, which is capable of operating on all classes of English sounds. The design process for the recognizer consisted of three elements: the feature-specification system, the probabilistic and fractional temporal overlapping pattern across the features, and the mapping from the feature-overlap pattern to a state-transition graph. They achieved preliminary and effective results in both phonetic classification and phonetic recognition.

Later on, extensive work incorporating high-level linguistic structure constraints in the automatic construction of a feature-based phonological model was reported in [10]. The linguistic information explored in this work included utterance and word boundaries, syllable constituents and word stress. They developed a consistent computational framework based on temporal feature logic for the construction of the phonological model. This model significantly improved earlier versions of the model in [2, 9] by successfully implementing the theoretical constructs in terms of rule formalisms and programs generating state-transition graphs. Experimental results demonstrated that this model would be feasible in the applications in speech recognition.

Markov et al. proposed an approach [11] that combined the acoustic and articulatory information to improve the performance of speech recognition systems. This study integrated features extracted from actual articulatory data with acoustic MFCC features. Without explicitly mapping the acoustic data into articulatory feature space, they used the probabilistic dependency between the two types of speech parameters by learning the hybrid HMM/Bayesian network (BN) model, where acoustic and articulatory data are represented by different BN variables. The evaluation experiments showed that this model, by effectively utilizing the available articulatory information, performed better than the baseline HMM trained only on acoustic features.

Some previous works on production-based speech recognition used dynamic Bayesian networks (DBNs) to recover gestural ensembles.

Livescu et al. [12] reported investigations using articulatory features (AF) for observation models and pronunciation models in speech recognition. Both models were implemented as dynamic Bayesian networks specifically, and tested on experiments for audio-only and audio-visual corpus. They did pronunciation modeling via multiple hidden streams of AFs, but did not outperform phone-based models. The most encouraging recognition results were from the tandem approach of observation modeling by using the outputs of multilayer perceptron (MLP) AF classifiers as part of the observation vector after post-processing. While the hybrid HMM/neural network models had lower accuracy than other models, they required very little training data beyond the MLP training, which might be promising in multilingual scenarios.

In particular, Zhuang et al. [7] proposed the instantaneous "gestural pattern vector" to encode gestural activation information across tract variables in the gestural score at a given time. They used artificial neural network and Gaussian mixture tandem models (ANN-GMM) to recover the instantaneous GPVs from tract variable time functions in local time windows, and achieved classification accuracy up to 84.5% for synthesized data from one speaker. They then used a task dynamic model of inter-articulator speech coordination to generate the "canonical" gestural score (CGS), and estimated the likelihood of the recognized GPV sequence on word-dependent GPV sequence models trained using the CGS. In addition, word-specific bigram GPV sequence models and artificial neural network Gaussian mixture tandem models (ANN-GMM) for the GPVs have been used to distinguish the GPV sequences of different words [13]. The bigram GPV sequence models, however, only leverage the frequencies of GPVs and GPV pairs to classify words, and have not fully explored the invariance of the ensemble of gestural activations.

1.4 Motivation

The finite state machine (FSM) is a compact representation for sequence modeling widely used in phone-based subunit sequence modeling in speech recognition.

In [14], a general framework based on weighted finite automata and weighted finite-state transducers (FSTs) for describing and implementing speech recognizers was presented. This framework could take context-dependent units, pronunciation dictionaries, language models and lattices uniformly as information sources and data structures used in recognitions. Particularly, information sources such as language models and dictionaries could be combined in advance, and further combined with acoustic observations dynamically during recognition using a single composition algorithm. Implementations were performed in various speech recognition and language processing tasks, including continuous speech recognition, isolated word recognition for directory lookup tasks, and segmentation of Chinese text into words.

Hetherington presented a simplified context-dependent phonological rewrite rule system and a technique to efficiently compile the system into FSTs in

[15]. The phonological rules allowed for both input-level and surface-level constraints. He found the rule compilation and application was more than 100 times faster than the nontransducer-based technique he previously used.

In particular, FSM has been used in modeling pronunciation variation to encode lexical pronunciation dictionaries and phonological rewrite rules [3, 16, 17]. In [16], Deng developed the overlapping-feature based phonological model by phonological rules interface with finite-state automata in a phoneme based environment. In [3], each pronunciation component is encoded within an FST representation whose transition weights are probabilistically trained using a modified EM algorithm [18, 19, 20] for finite-state networks. The results from experiment evaluated on the JUPITER [21] weather information conversational system showed that the explicit modeling of phonological effects which cause the deletion or insertion of phonetic events reduced word error rates by 8% on the test set. They also demonstrated that phonological effects which cause allophonic variation without altering the number of phonetic events could be modeled implicitly with context-dependent models to achieve better accuracy and less search space complexity.

As the gestural score can be represented as a GPV sequence [7], the FSM may be an efficient way to encode variations in the GPV sequences, given a particular gestural score. In this work, we propose a pronunciation modeling method that uses an FSM to represent the invariance of a gestural score. For a given word with its "canonical" gestural score, the plausible onsets and offsets of all gestural activations are generated in a recursive process by considering the varying lengths of the gestural activations with the constraint that the ensemble of gestures stay unchanged. These alternative gestural activation timing schemes are encoded as a weighted FSM for GPV sequences. Each path within the FSM represents the gestural score along with the onsets and offsets of all involved gestural activations.

The change of the gestural activations over time is modeled by the transitions between different states, and the length-varying characteristic of the gestural activations modeled by state self-transitions. An empirical measure for a partially generated gestural activation timing scheme is introduced in the recursive process to prune out those that deviate too much from the CGS. Speech recognition is achieved by matching the recovered ensemble of gestural activations to the FSM-encoded gestural scores of different speech

content.

We carry out pilot word classification experiments using synthesized data from one speaker. The proposed pronunciation modeling achieves over 90% accuracy for a vocabulary of 139 words with no training observations, outperforming directly using the CGS or the GPV bigrams. In addition, we tested our pronunciation model using recovered tract variable time functions from speech acoustics in previous work [1, 22], and got some interesting results which need more discussion in the future.

Chapter 2

Speech Recognition Using GPVs

2.1 Recognizing Gestures from Tract Variables

In [7], a speech recognition framework as an alternative to the classic sequence-of-phones model is proposed. The framework uses speech gestures as the invariant representation of human speech. The framework leverages a gestural pattern vector (GPV) representation, which encodes discretized instantaneous gestural activations (constriction target and stiffness) across tract variables at each time frame. A tandem model illustrated in Figure 2.1 is used to recover the instantaneous GPV from tract variable time function in a local time window. Classification accuracy achieved up to 84.5% for synthesized speech, which suggests that the proposed GPV might be a viable unit in statistical models for speech recognition.

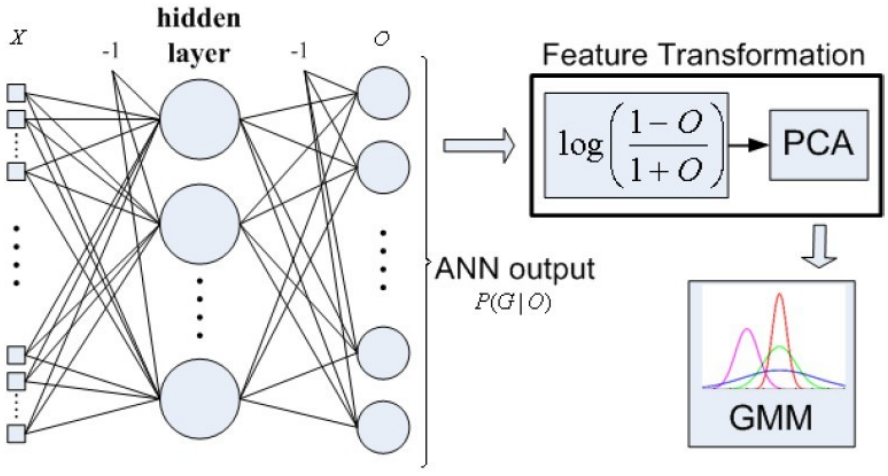

Figure 2.1: ANN-GMM tandem model

Recognition of GPVs is essentially a classification problem, which uses speech gesture activation information including constriction target and stiff-

ness extracted from tract variable time functions in a local time window centered at the target GPV as observations. However, this approach has some challenges: (1) The gestural activation is not perfectly synchronized with tract variable time functions. (2) The dynamic model demands smoothness in some sense so that the tract variable time functions have high correlation within a time-local neighborhood. (3) The tract variable time functions correlate across different tract variables, particularly when gestural activation is absent at some tract variable.

To solve the above problems, a tandem model is introduced. In Figure 2.1, the tandem model uses a discriminatively trained artificial neural network (ANN) to estimate posterior probabilities across all GPVs, which are then used as input features to Gaussian mixture models (GMM). All the observations are concatenated into a single observation vector X as input to the ANN. The output nodes O of the ANN correspond to different GPV types G, each indicating the posterior probability of one GPV type $P(G|X)$, given the current observation X. These ANN outputs are transformed into a new feature using $log(\frac{1-O}{1+O})$ and decorrelated using principal component analysis (PCA). PCA also reduces the dimensionality of the new feature, which is then used in GMM, each for one GPV type.

When testing, the observations X are presented to the input nodes, and classifications are performed by choosing the GPV whose GMM gives the highest likelihood for the new feature obtained by processing the current observation using the ANN and feature transform.

2.2 GPV-Based Word Classification

As an extension of the work described in Section 2.1, Zhuang et al. [13] proposed a speech recognition framework as an alternative to the classic sequence-of-phones model.

Given speech observation, recognizing the GPV sequence recovers the intervals of gestures together with the target and stiffness of their complete gestural score. The ensemble of gestures can be approximated by a sequence of GPVs. To classify recognized GPV sequences into words, we leverage GPV sequence models trained on GPV sequences converted from "canonical" gestural scores for each vocabulary word. Each GPV sequence is also weighted

by the likelihood for all the recognized individual GPVs involved.

Speech recognition based on GPV is formulated according to maximum a posteriori, as follows:

$$\begin{aligned} W &= argmax_i P(W_i|O), && (2.1)\\ P(W_i|O) &\approx p(GPVseq_i, W_i|O) \\ &= \frac{p(GPVseq_i, W_i, O)}{p(O)}, && (2.2) \end{aligned}$$

where $p(GPVseq_i, W_i|O)$ is the posterior of the i^{th} word and the recognized GPV sequence $GPVseq_i$. $GPVseq_i$ is the hypothesis by Viterbi decoding using the GPV sequence model for the vocabulary word W_i. From Equations 2.1 and 2.2,

$$\begin{aligned} W &\approx argmax_i p(GPVseq_i, W_i, O), && (2.3)\\ p(GPVseq_i, W_i, O) &= p(O, GPVseq_i|W_i) * p(W_i). && (2.4) \end{aligned}$$

Assuming equal prior for different speech content—in particular, words W—recognizing speech is formulated as follows:

$$W \approx argmax_i p(O, GPVseq_i|W_i), \quad (2.5)$$

where $p(O, GPVseq_i|W_i)$ is the joint likelihood of the observation O and the GPV sequence recognized using the recognizer for the i^{th} word. This joint likelihood can be formulated as follows:

$$\begin{aligned} & p(O, GPVseq_i|W_i) \\ = \; & p(O|GPVseq_i, W_i) * p(GPVseq_i|W_i) \\ = \; & \prod_{n=1}^{N} p(O_n|GPV_n) * p(GPVseq_i|W_i), && (2.6) \end{aligned}$$

where $GPV_n,\ n \in \{1, .., N\}$ constitute the GPV sequence $GPVseq_i$. $p(O_n|GPV_n)$ is modeled by an artificial neural network Gaussian mixture tandem model (ANN-GMM) as follows:

$$p(O_n|GPV_n) \approx p(F(\overrightarrow{P}(GPV_n|O_n))|GPV_n), \quad (2.7)$$

where $p(GPVseq_i|W_i)$ is modeled using a bigram GPV sequence model [13].

The word-independent or word-specific bigram GPV sequence models are trained using GPV sequences from all training utterances or a particular word, respectively. The former captures the general characteristics of GPV sequence resulting from the physical constraints inherent to consecutive gestural activation. The latter reveals information about the ensemble of gestures in a particular word, and therefore can be used to inform the speech content. To maintain robustness and smoothness of the word-specific bigram GPV sequence models, they are interpolated with the word-independent model.

Although the bigram GPV sequence model has its own merit of simplicity, it cannot explicitly model the temporal overlap and variations in pronunciation caused by overlapping gestures. The constraints in a gestural score are beyond local activation patterns captured by the GPV bigram models. Moreover, Equation 2.6 leverages only one recovered GPV sequence, i.e., one single gestural score (with the timing scheme), and may be vulnerable to noise.

We generalize the GPV-based recognition problem to use alternative recovered GPV sequences for more robust performance:

$$W \approx argmax_i \sum_j p(O, GPVseq_{i,j}|W_i). \tag{2.8}$$

We encode $p(O_n|GPV_n), n = 1, 2, ..., N$ in an FSM converted from a GPV lattice obtained using the Viterbi algorithm. $p(GPVseq_{i,j}|W_i)$ is encoded in a word-specific FSM that encodes the pronunciation of that word, as will be proposed in Chapter 3. Equation 2.8 can be evaluated using FSM composition between the above two FSMs.

We illustrate the speech recognition system in Figure 2.2.

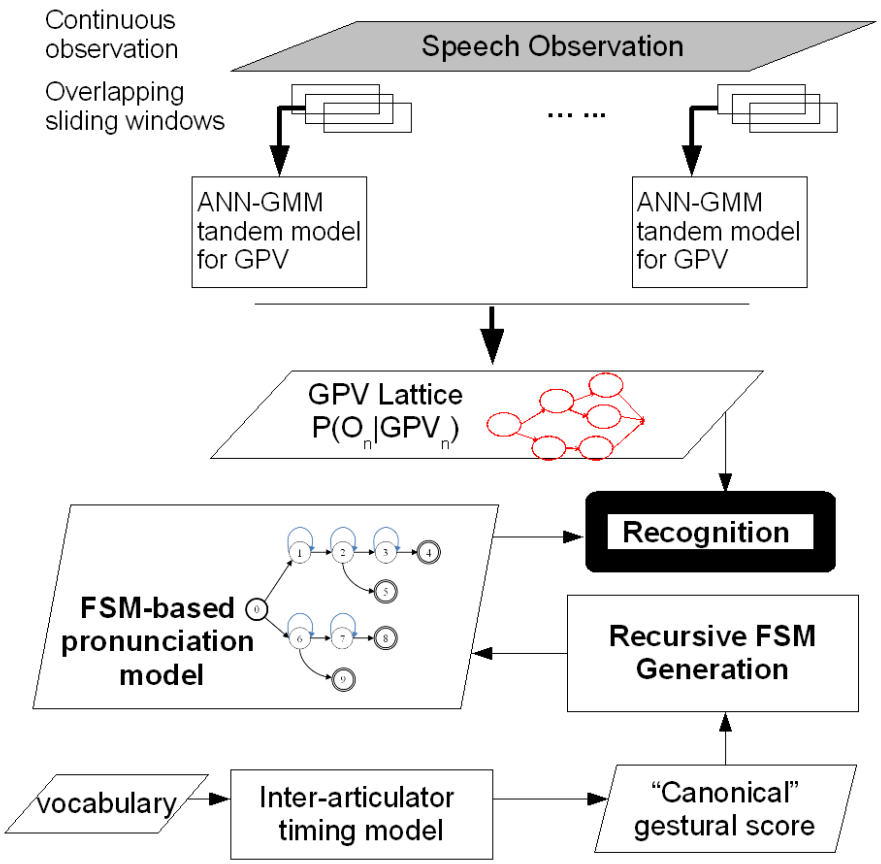

Figure 2.2: FSM-based speech recognition using GPVs

Chapter 3

FSM-Based Pronunciation Model

3.1 Pronunciation Variation

The gestural score encodes both the invariance of the ensemble of gestures and the variability of their onset and offset times. This invariance is possible only with the onsets and offsets of the gestures varying to reflect the variation of the same speech content, such as different speech rates, coarticulation and reduction [5, 6].

Let us take the "coarticulation" of consonants and vowels as an example. Consonantal gestures typically have a greater degree of constriction and a shorter time constant (higher stiffness) than the vocalic gestures [6]. The basic relationship between them is that initial consonants are coordinated with vowel gesture onset, and finally consonants with vowel gesture offset. This results in organizations in which there is substantial temporal overlap between movements associated with vowel and consonant gestures, as was seen in the gestural score of Figure 1.2. The consonant/vowel overlaps illustrate that gestures can give rise to context-dependent articulatory and acoustic trajectories for converting specific invariant units into variable parameters.

Variation also occurs in the case when there are differences in the characteristic patterns of overlapping gestures in syllable-initial and syllable-final positions. For example, in the work-initial case, the end of the velum lowering movement is roughly synchronous with the end of the lip closing movement; however, for the word-final case, the end of velum lowering occurs substantially earlier than the end of lip movement. Syllable-position effects are similar to these word-position effects.

In addition, some kinds of allophonic variation can be shown to result from quantitative variation in a gesture's dynamic parameters as a function of prosodic variables such as stress and position. Gestures shrink in space and

in time in some contexts. This kind of variation scales the metric properties of a gestural activation, but does not alter the composition of articulatory components out of which it is assembled.

Using the gestural approach can effectively analyze phonological and phonetic variation that is attributed to processes occurring during the physical act of talking. It is well known that in connected speech the patterns of gestural overlap may vary. In particular, factors associated with increased fluency result in increasing the temporal overlap among gestures. Further, prosodic boundaries may influence the degree of overlap between neighboring gestures that belong to successive words. Examples in [5] show that there are circumstances in which increased overlap would result in connected speech alternations. One such circumstance is gestural "hiding." For example, two productions of the sequence "perfect memory" were produced—one as part of a word list with an intonation boundary between the two words, and the other as part of a fluent phrase. In the fluent phrase version, the final [t] of "perfect" was not audible, and it would be conventionally analyzed as an example of alveolar stop deletion.

3.2 Finite State Machine Representation

The "canonical" gestural score (CGS) can be obtained for each word using a task dynamic model of inter-articulator speech coordination, implemented in the Haskins Laboratories speech production model of TADA [8]. In this model, orthographical inputs are syllabified by applying the max-onset algorithm to entries in the Carnegie Mellon pronouncing dictionary. The syllabified inputs are parsed into gestural regimes and intergestural coupling relations by gestural dictionary and intergestural coupling principles, respectively. These gestural scores are converted to GPV sequences for training the GPV sequence models.

Zhuang et al. [13] use GPV bigram statistics in the CGS to distinguish different words. However, bigram model has a big limitation in explicitly modeling the overlapping features and pronunciation variation caused by speech reduction and coarticulation. The CGS represents only one sample from the distribution of possible gestural activation timing schemes. Thus the only recovered GPV sequence (gestural score) would be vulnerable to

noise.

We propose a pronunciation model based on finite state machines (FSM) to encode the variance of the gestural activation timing schemes given a particular gestural score, with the explicit constraint that the ensemble of gestures stay invariant. As in [13], we use a GPV sequence to represent a gestural score with a particular timing scheme. The FSM-based pronunciation model encodes the variation of gestural activation timing schemes via a large number of alternative GPV sequences, all containing the same ensemble of gestures.

Each state in the FSM is associated with a particular combination of the instantaneous activations across all tract variables, as will be approximated using a GPV. A new state is introduced only when the set of activations differ from those in its immediate neighboring (connected) state. Therefore, at least one activation onset or offset is observed when the FSM transits from one state to another.

The proposed pronunciation model captures the invariance of the ensemble of gestures and the length-varying characteristic of gestural activations in three ways. First, the transition between different states models the GPV change over time. Second, each GPV path is required to contain one and only one instance of each gesture in the CGS. Third, the self transitions on each state allow varying length of each GPV.

While the gestural activation timing schemes may vary in a gestural score, not all timing schemes are equally plausible. For example, in the word "about", the "release labial closure" gesture should not appear before "initial labial closure" gesture (Figure 1.2). Though determining the particular onsets and offsets of gestural activations is still an open challenge, we take an empirical approach that suppresses options that deviate too much from the CGS.

3.3 Recursive Algorithm

For a particular gestural activation timing scheme, we label all the onsets and offsets according to their order in time $(1, 2, ...)$, referred to as the *order number*. We define the *order number deviation* by calculating the absolute difference ($|O_s - O_{cgs}|$) between the order number (O_s) of an onset or offset,

and its order number in the CGS (O_{cgs}). We denote L_s to be the *pronunciation likelihood* of the GPV on a particular state s. Given the CGS, we estimate L_s to be e^{-n}. Here n is the sum of the order number deviations for all the onsets and offsets observed when transitioning into the concerned state. Note that L_s can be computed as long as all the GPVs before the current states are known.

We use a recursive procedure to produce a FSM-based pronunciation model for each word, which is initialized as an empty FSM with a null state. The reference order numbers for each onset or offset in the CGS, and the gestural activations in the CGS, are also stored in the initialization. The following recursive function generates the states as well as the inter-state transitions of the FSM:

Recursive_FSM_Generation (Existing States, current state $S_{current}$):

If *termination condition* is not satisfied:

1. Propose new states with updated instantaneous gestural activations by adopting at least one of the following changes on $S_{current}$:

   - Introduce the onset of a gesture that has never started.
   - Introduce the offset of a gesture that has started but not yet ended.

2. For each proposed new state S_{new} calculate the likelihood of the state using the order number deviations of the new onsets or offsets:

   - If the likelihood of transition from the initial state to the new state satisfies some *pruning condition*, the new state is discarded.
   - Otherwise, establish a transition from $S_{current}$ to S_{new}, and call Recursive_FSM_Generation ([Existing States; S_{new}], S_{new}).

The *termination condition* of the above recursion is that both onsets and offsets of all gestural activations in the CGS have occurred. This ensures that any path in the FSM will satisfy the invariance of the ensemble of gestures for the concerned word.

The *pruning condition* is introduced for two reasons: First, gestural activation timing schemes that deviate too much from the CGS are not very likely to be justifiable in human speech. Second, the complexity of the pronunciation model should be controlled for practical reasons. We introduce an

empirical measure, *average onset/offset likelihood*: $\sqrt[N]{\prod_s L_s}$, where the product is evaluated from the initial state to the current state $S_{current}$, and N is the number of onsets and offsets in the same span. The pruning condition is satisfied when the average onset/offset likelihood falls below a threshold.

To additionally account for the varying length of each gestural activation, we add a self-transition for each state in the FSM. Each state self-transition is associated with a likelihood, modeling the exponential distribution of the length of the instantaneous gestural activations on this state. The self-transition likelihoods can be predefined or trained using durations of the known ensemble of gestures, using the EM algorithm [3].

Chapter 4

Experiments and Results

4.1 Dataset and Setup

We use a speech dataset synthesized by TADA [8, 23, 24] containing all the following: acoustics, tract variable time functions, gestures and lexical representation. TADA syllabifies the lexical inputs and parses them into gestural regiments (GRs) with intergestural coupling relationships (IGRs). Using the GRs and the IGRs, TADA uses an intergestural timing model to synthesize the gestural scores, which are input to the task-dynamic model to obtain the tract variable time functions. These are mapped to the vocal tract area function (sampled at 200 Hz), and to speech acoustics. These speech acoustics are synthesized by Sensimetrics HLSyn [25], sampled at 10000 Hz. The obtained gestural score is an ensemble of gestures for the utterance, specifying the intervals of time during which particular constriction gestures are active in the vocal tracts. To define a set of frequent GPVs, we randomly split the dataset into three folds and adopt only those GPVs that appear at least ten times in each fold. We get 146 distinctive GPVs, including a special "unknown" GPV, which accounts for less than 10% of the data that do not correspond to the frequent GPVs.

The dataset contains the same 416 words as in the Wisconsin articulatory database [26], which are randomly split into a training set of 277 words and a testing set of 139 words, without word identity overlapping.

The inputs O to the ANN are values of the eight tract variable time functions over a local time window of 15 frames, normalized by the mean and standard deviation within each tract variable in the training and testing sets, respectively. The ANN has 81 hidden nodes and PCA reduces the dimensionality of the transformed features from 146 to 80.

As in [13], artificial neural network Gaussian mixture tandem models

(ANN-GMM) are trained for 145 distinct GPV types. This work takes eight-dimensional tract variable time functions as observations.

The proposed FSM-based pronunciation models are used to classify the 139 words in the test set, which do not overlap with the 277 words used to train the GPV observation models. All the FSMs for the 139 words in the dictionary are unified together to constrain the Viterbi decoding for the GPV lattice. By varying the pruning condition for each word, we can adjust the number of alternative timing schemes of the gestural activations, resulting in different FSM-based pronunciation model sizes. According to our empirical method for pruning, by assigning different threshold values, we can get accordingly different size, i.e., the number of paths of each word, of FSM-based pronunciation models. Each set of the model will be a GPV lattice. The resultant GPV lattices are composed with each word-specific FSM to perform classification according to Equation 2.8.

With the same ANN-GMM tandem models, classification is also performed with two other pronunciation models: (1) The *GPV bigram model* uses frequencies of GPVs and GPV pairs to distinguish different words. The task dynamic model of intergestural timing provides the CGS for each vocabulary word approximated as a GPV sequence, which is then used to build the word-specific GPV bigrams. (2) The canonical GPV sequence model, which is a special case of the FSM-based model such that only one GPV sequence is modeled for each word.

We also conduct a CGS recovery experiment using different *gestural score recovery models*. The so-called gestural score recovery models are the union of the word pronunciation models, without using the identities of the words. These models describe the underlying constraints shared across different gestural scores for different words.

4.2 Results

Figure 4.1 shows the result of FSM-based pronunciation variation modeling for word "the". $A1$ through $A4$ represent the four different gestural activations of this word. Each state encodes instantaneous gestural activations across all tract variables. States 4, 5, 8 and 9 are the terminus of particular paths in the FSM, each describing a complete gestural score with a particu-

lar activation timing scheme. All paths share the same ensemble of gestural activations.

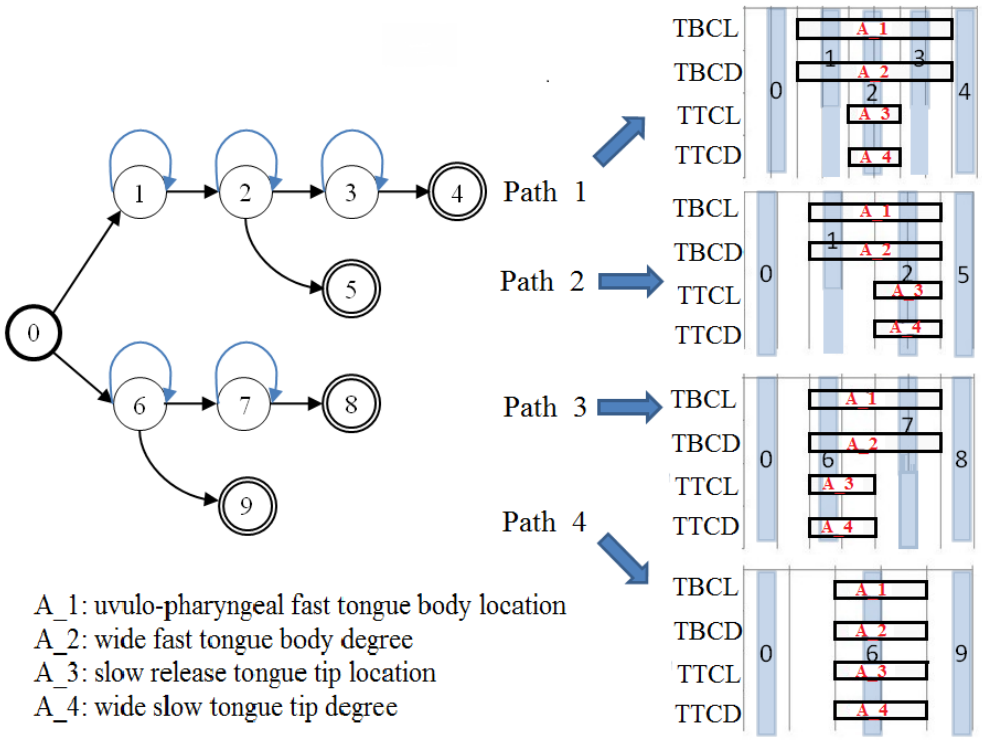

Figure 4.1: Pronunciation variation of "the" as proposed by the FSM

Figure 4.2 illustrates a successful classification of two words. The two words differ in the CGS: "hand" has one additional gestural activation "open velum" which indicates the nasal sound; and they are also different in the target value of tongue body constriction location ("uvulo-pharyngeal" and "pharyngeal"). After recognition using the proposed pronunciation model along with the tandem GPV observation models, the recovered gestural scores for the two words deviate from their CGSs in the small changes of the onsets or offsets of some gestural activations. However, the ensembles of the gestural activations are kept unchanged, resulting in correct classification.

Figure 4.3 gives an example of misclassification from "arm" to "on". The ensemble of gestures recovered during recognition of the utterance "arm" differs from its CGS, by a gestural activation deletion of "labial closure" gesture, a different constriction location of tongue body (from "palatal" to "alveo-palatal"), and a different constriction degree of tongue body (from

"narrow" to "closure"). The resulting gestural score matches the ensemble CGS of "on", thus the misclassification occurs.

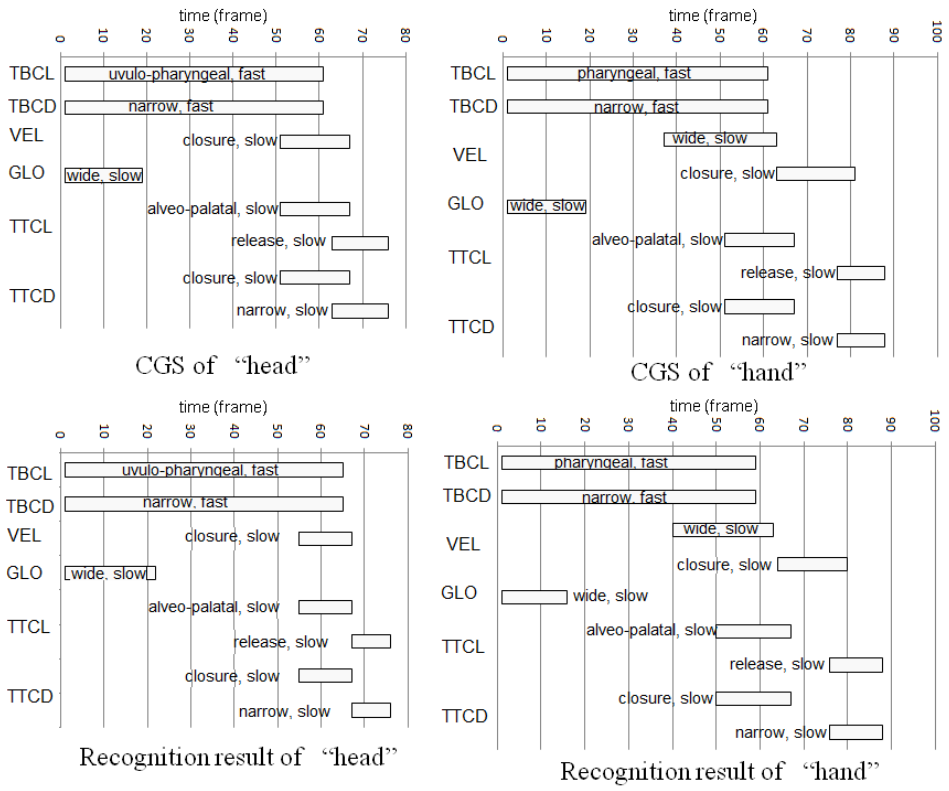

Figure 4.2: Classification of "head" and "hand" (the recovered gestural scores have timing schemes different from the CGSs)

Table 4.1 presents the classification accuracy using different pronunciation models. The proposed FSM-based pronunciation model with each word having over 200 different timing schemes (FSM-based II) achieves the highest classification accuracy of over 90%. The model of a smaller size, which contains fewer than 50 timing schemes for each word (FSM-based I) results in a lower accuracy of almost 90%. They both outperform the GPV bigram model and the canonical GPV sequence method. This suggests that the proper deviation of the timing schemes of the gestural activations from the CGSs leads to more robust pronunciation models.

Table 4.2 presents the F-score of the recovered discretized dynamic parameters, i.e., constriction targets and stiffness, that are used to define the GPVs. We can see that the proposed FSM-based model also achieves the best results in CGS recovery for most of the tract variables.

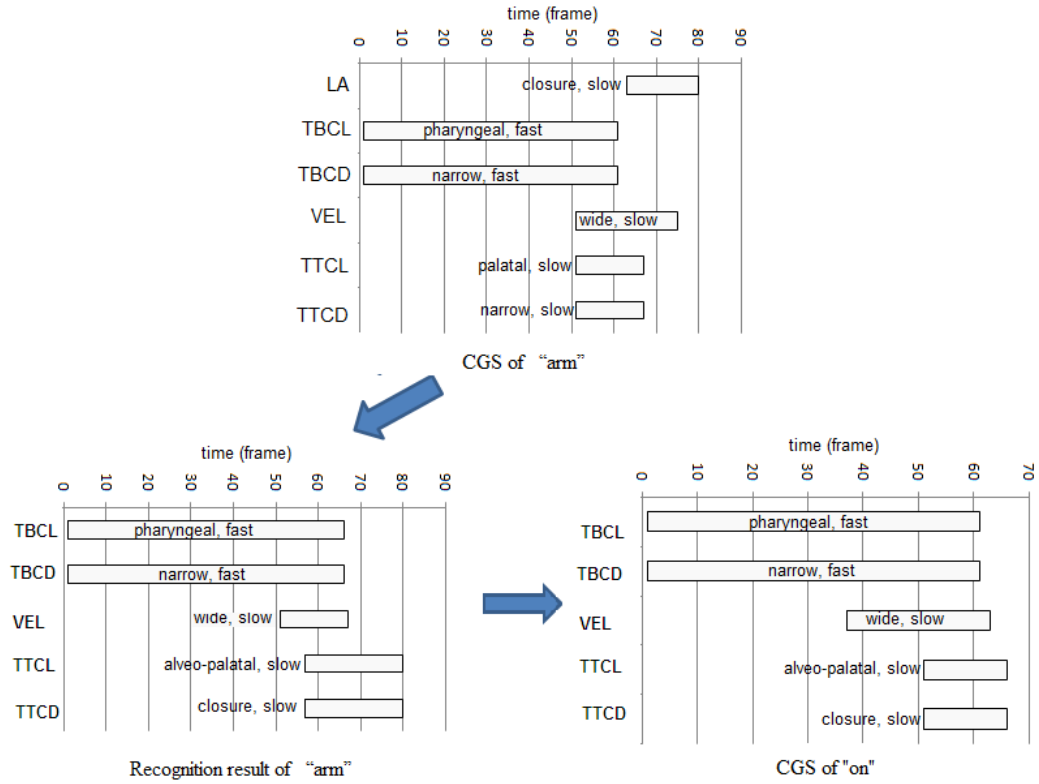

Figure 4.3: Misclassification from "arm" to "on"

Table 4.3 presents the F-score of the above gestural recovery experiment using the tract variable time function estimated from acoustics [22]. Here we see the accuracy much off the ground truth, especially for TTCL and TTCD.

Table 4.1: Word classification accuracy (%) with different pronunciation models

Models	GPV bigram	Canonical GPV sequence	FSM-based I	FSM-based II
Accuracy	84.89	87.77	89.93	90.65

Table 4.2: F-score (%) of recovered discretized gestural activation ("Targ": constriction targets; "Stif": constriction stiffness)

Models		GPV bigram	FSM-based II
Targ&Stif		81.35	84.74
Targ		79.23	82.99
Stif		84.56	87.37
Targ	PRO	85.26	84.38
	LA	77.48	80.11
	TBCL	82.98	87.73
	TBCD	86.07	88.51
	VEL	75.50	78.94
	GLO	72.72	76.03
	TTCL	69.32	73.93
	TTCD	68.62	75.56
Stif	PRO	85.66	84.43
	LA	77.41	80.70
	TBCL	85.93	89.06
	TBCD	85.94	89.02

Table 4.3: F-score (%) of recovered discretized gestural activation using estimated tract variable time functions

Models		GPV bigram
Targ&Stif		42.96
Targ		37.04
Stif		52.04
Targ	PRO	30.07
	LA	30.13
	TBCL	40.30
	TBCD	60.29
	VEL	30.84
	GLO	25.51
	TTCL	16.79
	TTCD	14.33
Stif	PRO	30.07
	LA	27.14
	TBCL	59.50
	TBCD	59.39

Chapter 5

Conclusion and Discussion

5.1 Conclusion

According to articulatory phonology, the gestural score is an invariant speech representation. Though the activation interval (timing schemes), i.e., the onsets and offsets, and the dynamic parameter, i.e., target, stiffness, and damping, of the gestural activations may vary, the ensemble of these activations tends to remain unchanged, informing the speech content. The instantaneous "gestural pattern vector" (GPV) is proposed as a sub-word unit for encoding gestural activation information across all tract variables.

In this work, we propose a pronunciation modeling method that uses a finite state machine (FSM) to represent the invariance of a gestural score, as well as to encode the variance of the gestural activation timing scheme given a particular gestural score which is approximated by a sequence of GPVs. Given the "canonical" gestural score (CGS) of a word with a known activation timing scheme, the plausible activation onsets and offsets are recursively generated and encoded as a weighted FSM. Each state in the FSM is associated with a particular type of GPV, and phonological rules are introduced to guarantee that at least one activation onset or offset is observed when the FSM transits from one state to another. An empirical measure is used to prune out gestural activation timing schemes that deviate too much from the CGS. Speech recognition is achieved by matching the recovered gestural activations to the FSM-encoded gestural scores of different speech content.

We carry out pilot word classification experiments using synthesized data from one speaker. The proposed pronunciation modeling achieves over 90% accuracy for a vocabulary of 139 words with no training observations, outperforming direct use of the CGS and the previously proposed GPV bigram model. In addition, we conduct CGS recovery experiments using the same

data. The FSM-based model also achieves the best results for most of the tract variables. Although the same gestural recovery experiment performed on the tract variable time function estimated from acoustics [22] has much lower accuracy compared to the ground truth, we still see the plausibility of applying our system on it by improving the acoustic model as well as the pronunciation model.

5.2 Discussion

There are a few extensions that we consider as possible future research.

First, it may be beneficial to engage a more accurate state likelihood estimation using methods proposed in [27], which measures information including articulatory effort, communicative efficacy (parsing cost for the listener), as well as the overall utterance duration. However, that will probably lead to increased computational cost and demand for more data.

Second, we want to improve the performance of both the gestural recovery and word recognition experiment, which performed on the estimated tract variable time functions from both synthesized speech and real speech.

Third, it is also convenient to adapt synthesized acoustic as the system input to do gestural recovery and word classification.

Finally, we expect to apply the proposed method to real speech, where more pronunciation variation is observed.

References

[1] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *Journal of Acoustic Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[2] L. Deng and D. Sun, "Speech recognition using the atomic speech units constructed from overlapping articulatory features," in *EUROSPEECH'93*, 1993, pp. 1635–1638.

[3] H. S. Timothy J. Hazen, I. Lee Hetherington and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," in *Proc. ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation*, Estes Park, Colorado, 2002, pp. 99–104.

[4] D. McAllester, L.Gillick, F. Scattone, and M. Newman, "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," in *Proc. ICSLP*, Sydney, Australia, December 1998, p. 0986.

[5] J. Kingston and M. E. Beckman, *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, 3rd ed. Cambridge, UK: Cambridge University Press, 1990, ch. 19, pp. 341–376.

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

[7] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein, and E. Saltzman, "The entropy of the articulatory phonological code: Recognizing gestures from tract variables," presented at 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 2008.

[8] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable task dynamics model in MATLAB," *Journal of the Acoustical Society of America*, vol. 115, no. 5, p. 2430, 2001.

[9] L. Deng and D. Sun, "Phonetic classification and recognition using HMM representation of overlapping articulatory features for all

classes of English sounds," in *Proc. ICASSP'94*, Adelaide, Austrailia, 1994. [Online]. Available: citeseer.ist.psu.edu/deng94phonetic.html pp. I–45–I–48.

[10] J. Sun, L. Deng, and X. Jing, "Data-driven model construction for continuous speech recognition using overlapping articulatory features," in *Proc. ICSLP '00*, vol. 1, 2000, pp. 437–440.

[11] K. Markov, J. Dang, and S. Nakamura, "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," *Speech Communication*, vol. 48, pp. 161–175, 2006.

[12] K. Livescu, O. Çetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magimai-Doss, and K. Saenko, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," in *Proc. ICASSP*, Hawaii, U.S.A., 2007, pp. 621–624.

[13] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein, and E. Saltzman, "Articualatory phonological code for word classification," presented at 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 2009.

[14] F. Pereira and M. Riley, "Speech recognition by composition of weighted finite automata," in *Finite-State Language Processing*. Cambridge, MA: MIT Press, 1996, pp. 431–453.

[15] I. L. Hetherington, "An efficient implementation of phonological rules using finite-state transducers," in *Proc. EUROSPEECH*, Aalborg, Denmark, September 2001, pp. 1599–1602.

[16] L. Deng, "Finite-state automata derived from overlapping articulatory features: A novel phonological construct for speech recognition," in *Proceedings of the Workshop on Computational Phonology in Speech Technology*, Association for Computational Linguistics, Santa Cruz, CA, June 1996, pp. 37–45.

[17] J. Sun and L. Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition," *Journal of Acoustic Society of America*, vol. 111, no. 2, pp. 1086–1101, February 2002.

[18] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, June 1977.

[19] J. Eisner, "Parameter estimation for probabilistic finite-state transducers," in *Proc. of the Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July 2002, pp. 1–8.

[20] H. Shu and I. L. Hetherington, "EM training of finite-state transducers and its application to pronunciation modeling," in *Proc. ICSLP*, Denver, CO, September 2002, pp. 1293–1296.

[21] V. Z. et al., "JUPITER: A telephone-based conversational interface for weather information," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 1, pp. 85–96, January 2000.

[22] V. Mitra, I. Y. Ozbek, H. Nam, X. Zhou, and C. Espy-Wilson, "From acoustics to vocal tract time functions," in *Proc. ICASSP*, Washington DC, 2009, pp. 4497–4500.

[23] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, no. 4, pp. 332–382, 1989.

[24] Haskins Laboratories, "TADA: An enhanced, portable task dynamics model in matlab," New Haven, CT, 1997. [Online]. Available: http://www.haskins.yale.edu/tada_download/index.html

[25] Haskins Laboratories, "HLsyn: High-level parametric speech synthesis engine," New Haven, CT, 1998. [Online]. Available: http://www.sens.com/hlsyn/

[26] J. Westbury, "X-ray microbeam speech production database user's handbook," University of Wisconsin Waisman Center, Madison, WI, 1994.

[27] J. Simko and F. Cummins, "Sequencing of articulatory gestures using cost optimization," presented at 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 2009.