FSM-Based Pronunciation Modeling using Articulatory Phonological Code

Chi Hu, Xiaodan Zhuang, and Mark Hasegawa-Johnson

Beckman Institute, Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, U.S.A.
chihu1@illinois.edu, xzhuang2@illinois.edu, jhasegaw@illinois.edu

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.

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 (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. 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 contents.

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.

Index Terms: articulatory phonology, speech production, speech gesture, finite state machine

1. Introduction

The standard approach for automatic speech recognition assumes the speech signal is represented as a concatenation of phones [1]. 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 [15, 16], but limited by the coarseness of the phones [10]. 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. The ensemble of gestures is invariant, except for some types of historical alternation, but shifts in the relative timing of the gestures can cause coarticulated or reduced speech when they overlap in time. *Gestural score* specifies the temporal activation intervals for each gesture in an utterance.

Several methods have been proposed for speech recognition using speech production knowledge [7, 8, 9]. King et al. [1] gave a comprehensive review. Some previous works on production-based speech recognition [9, 11] used dynamic Bayesian networks (DBNs) to recover gestural ensembles. In particular, Zhuang and Nam et al. [11] proposed the instantaneous "gestural pattern vector" (GPV) to encode gestural activation information across tract variables in the gestural score at a given time. 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 [12]. 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.

The finite state machine (FSM) is a compact representation for sequence modeling widely used in phone-based subunit sequence modeling in speech recognition [13, 14]. In particular, FSM has been used in modeling pronunciation variation to encode lexical pronunciation dictionaries and phonological rewrite rules [3, 4, 15]. In [3], Deng developed the overlapping-feature based phonological model by phonological rules interface with finite-state automata in a phoneme based environment. As the gestural score can be represented as a GPV sequence [11], 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 (CGS), 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 are too much deviated 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.

2. Speech recognition using GPVs

Articulatory phonology describes speech in terms of gestures. Gestures are defined as dynamical control regimes for constriction actions at eight different constriction tract variables consisting 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). For a given constriction gesture, the activation interval or the timing scheme (onset and offset times) and dynamic parameters (target/stiffness/damping) are represented in a gestural score. The ensemble of the gestures with their dynamic parameters is distinctive to speech content which is inferred to as the *invariance of the gestural score*.

Zhuang, Nam et al [12] propose a speech recognition framework as an alternative to the classic sequence-of-phones model. The proposed framework uses speech gestures as the invariant representation of human speech. The framework leverage a *gestural pattern vector* (GPV) representation, which encodes discretized instantaneous gestural activations (constriction target and stiffness) across tract variables at each time frame [11]. The speech gestural score can be approximated by a sequence of GPVs.

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

$$W \approx argmax_i p(O, GPVseq_i|W_i), \tag{1}$$

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,

$$\begin{aligned} & p(O, GPVseq_i|W_i) \\ = \; & p(O|GPVseq_i, W_i) * p(GPVseq_i|W_i) \\ = \; & \prod_{n=1}^{N} p(O_n|GPV_n) * p(GPVseq_i|W_i), \end{aligned} \tag{2}$$

where $GPV_n,\ n \in \{1, .., N\}$ constitute the GPV sequence $GPVseq_i$ and $p(O_n|GPV_n)$ is modeled by an artificial neural network Gaussian mixture tandem model (ANN-GMM). $p(GPVseq_i|W_i)$ is modeled using a bigram GPV sequence model [12].

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 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:

$$W \quad \approx \quad argmax_i \sum_j p(O, GPVseq_{i,j}|W_i), \qquad (3)$$

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 Section 3. Equation 3 can be evaluated using FSM composition between the above two FSMs.

We illustrate the speech recognition system in Figure 1.

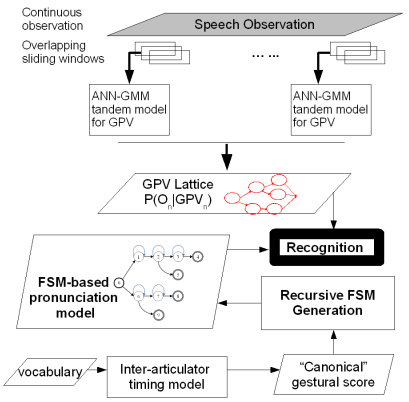

Figure 1: FSM-based speech recognition using GPVs

3. FSM-based pronunciation model

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, coarticulartion and reduction [5, 6].

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 [17]. Zhuang, Nam et al. [12] use GPV bigram statistics in the CGS to distinguish different words. However, the CGS represents only one sample from the distribution of possible gestural activation timing schemes.

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 [12], 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 neighbouring (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. 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.

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 [15].

4. Experiments and Results

4.1. Dataset and Setup

We use a speech dataset synthesized by TADA [17] 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.

The dataset contains the same 416 words as in the Wisconsin articulatory database [18], which are randomly split into a training set of 277 words and a testing set of 139 words, without word identity overlapping.

As in [12], 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 don't 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. The resultant GPV lattices are composed with each word-specific FSM to perform classification according to Equation 3.

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 model provides the CGS for each vocabulary word approximates 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 2 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 particular activation timing scheme. All paths share the same ensemble of gestural activations.

Figure 3 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, which result in correct classification.

Figure 4 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.

Table 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

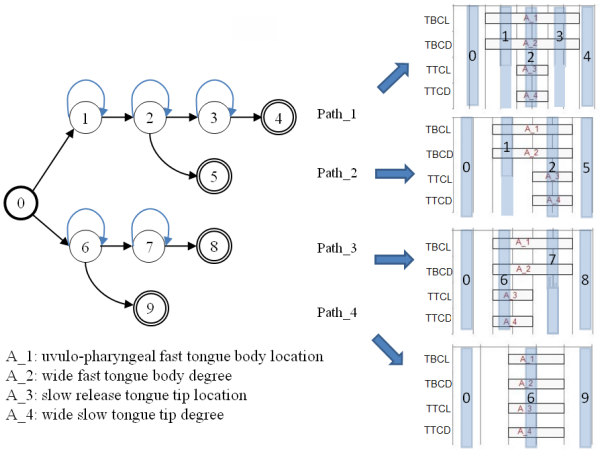

Figure 2: Pronunciation variation of "the" as proposed by the FSM

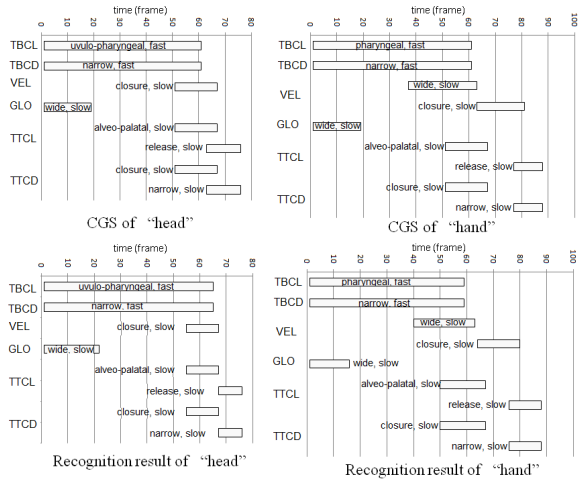

Figure 3: Classification of "head" and "hand" (The recovered gestural scores have activation timing schemes different from the CGSs.)

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 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.

5. Conclusion & Discussion

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.

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 (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. An empirical measure is used to prune out gestural activation timing schemes that deviate too much from the CGS. Speech recognition is achieved by

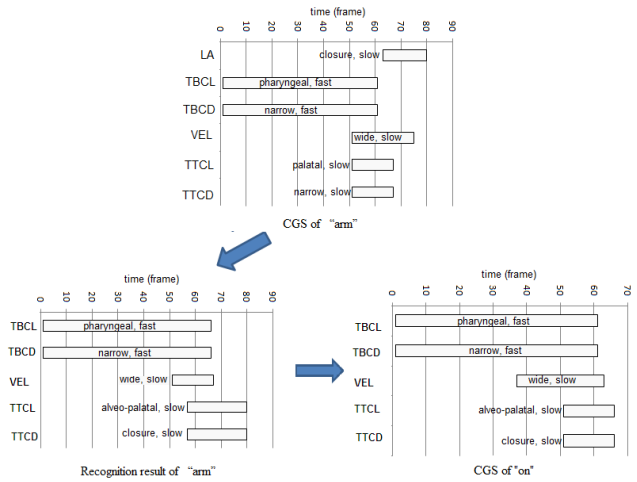

Figure 4: Misclassification from "arm" to "on"

Table 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

matching the recovered gestural activations to the FSM-encoded gestural scores of different speech content.

We carry 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 perviously 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.

There are a few extentions that we consider as possible future research. First, it may be beneficial to engage a more accurate state likelihood estimation using methods proposed in [15], though that will probably lead to increased computational cost and demand for more data. Second, we expect to apply the proposed method on real speech, where more pronunciation variation is observed.

6. Acknowledgements

This research is funded by NSF grant IIS-0703624. The authors would like to thank Vikramjit Mitra and Hosung Nam for assistance with the dataset.

7. 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 *3rd European Conference on Speech Communication and Technology*, 1993.

[3] 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, 28 June 1996, pp. 37–45.

[4] J. Sun and L. Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition," *Journal of Acoustic Society of America,* 111(2):1086–1101, February 2002.

[5] C. P. Browman and L. Goldstein, "Tiers in articulatory phonology, with some implications for casual speech," *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech, Kingston, J., and Beckman, M. E. [Eds], Cambridge U Press*, pp. 341–376, 1991.

Table 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

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

[7] 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.

[8] L. Deng and D. Sun, "Phonetic recognition using HMM representation of overlapping articulatory features for all classes of english sounds," in *Proc. ICASSP '94*, Adelaide, Austrailia, 1994, pp. I–45–I–48. [Online]. Available: citeseer.ist.psu.edu/deng94phonetic.html

[9] K. Livescu, O. Cetin, 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.

[10] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition",*Journal of the Acoustical Society of America*, October 31, 2006.

[11] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein, and E. Saltzman, "The entropy of the articulatory phonological code: Recognizing gestures from tract variables," in *Proc. Interspeech*, Brisbane, Australia, 2008.

[12] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein, and E. Saltzman, "Articulatory phonological code for word classification," in *Proc. Interspeech*, Brighton, United Kingdom, 2009.

[13] F. Pereira and M. Riley, "Speech recognition by composition of weighted finite automata," *Finite-State Language Processing*, pp. 431–453, 1997.

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

[15] 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.

[16] 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.

[17] 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. 5,2, p. 2430, 2004.

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