Articulatory Phonological Code for Word Classification

Xiaodan Zhuang[1], *Hosung Nam*[2],
Mark Hasegawa-Johnson[1], *Elliot Saltzman*[23], *and Louis Goldstein*[24]

[1]Beckman Institute, ECE Department, University of Illinois at Urbana-Champaign, U.S.A.
[2]Haskins Laboratories, New Haven, U.S.A.
[3]Department of Physical Therapy and Athletic Training, Boston University, U.S.A.
[4]Department of Linguistics, University of Southern California, U.S.A.

xzhuang2@uiuc.edu, nam@haskins.yale.edu
jhasegaw@ad.uiuc.edu, esaltz@bu.edu, louisgol@usc.edu

Abstract

We propose a framework that leverages articulatory phonology for speech recognition. "Gestural pattern vectors" (GPV) encode the instantaneous gestural activations that exist across all tract variables at each time. Given a speech observation, recognizing the sequence of GPV recovers the ensemble of gestural activations, i.e., the gestural score. For each word in the vocabulary, we use a task dynamic model of inter-articulator speech coordination to generate the "canonical" gestural score. Speech recognition is achieved by matching the ensemble of gestural activations. In particular, we estimate the likelihood of the recognized GPV sequence on word-dependent GPV sequence models trained using the "canonical" gestural scores. These likelihoods, weighted by confidence score of the recognized GPVs, are used in a Bayesian speech recognizer.

Pilot gestural score recovery and word classification experiments are carried out using synthesized data from one speaker. The observation distribution of each GPV is modeled by an artificial neural network and Gaussian mixture tandem model. Bigram GPV sequence models are used to distinguish gestural scores of different words. Given the tract variable time functions, about 80% of the instantaneous gestural activation is correctly recovered. Word recognition accuracy is over 85% for a vocabulary of 139 words with no training observations. These results suggest that the proposed framework might be a viable alternative to the classic sequence-of-phones model.

Index Terms: speech production, speech gesture, tandem model, artificial neural network, Gaussian mixture model

1. Introduction

Current state-of-the-art speech recognition systems adopt the assumption that speech is a sequence of phones. These systems work much better for carefully articulated speech, such as broadcast news, than for conversational speech, which has more significant coarticulation and reduction. Basic recognition units designed for coarticulation might be more efficient than the traditional phones. Articulatory phonology represents speech as an ensemble of gestures. This representation is relatively invariant as the gestures can generate coarticulated or reduced speech when they overlap in time[1, 2]. This work proposes a speech recognition framework motivated by articulatory phonology.

Several methods exist for speech recognition using speech production knowledge [3, 4, 5]. King et al. [6] gave a comprehensive review. There have been studies on recovering gestural activation intervals from acoustic signals or articulatory movements using the temporal decomposition method [7, 8]. Livescu et al.[5] proposed recovering gesture ensembles as the state variables in a dynamic Bayesian network. Our recent work that proposed the instantaneous "gestural pattern vector" (GPV) [9] is similar in philosophy with [5], but different in model design and all other computational details. Each GPV encodes gestural activation information across tract variables in the gestural score at a given time.

In this work, we propose a speech recognition framework using GPVs instead of phones. Given a speech observation, recognizing the GPV sequence recovers the intervals of gesture activations together with the target and stiffness of their control regimes, i.e., the complete gestural score. Word recognition is achieved by matching the recognized GPV sequence to canonical gestural scores for each vocabulary word, generated by a task dynamic model of inter-articulator speech coordination [10, 11].

In particular, the recognized GPV sequence is scored by word-specific bigram GPV sequence models, each trained using the GPV sequence converted from the "canonical" gestural scores. Similar to [9], we use an artificial neural network and Gaussian mixture tandem model to estimate the likelihood of speech observation given each GPV. Both the GPV sequence score and the observation likelihood for individual GPVs are used in a Maximum a Posteriori speech recognition framework.

We carry out a pilot experiment of gestural score recovery and word classification on synthesized data from one speaker. Previous work has reported successful recovery of the tract variable time functions from speech acoustics[6, 12]. In this work, we use tract variable time functions in the place of speech observation, since they are better correlated with the gestural activation than acoustic features. About 80% of the instantaneous gestural activation is correctly recovered for speech content unseen in training. Word classification accuracy is over 85% for a vocabulary of 139 words with no training observations.

2. Articulatory phonology and gestural pattern vectors

Articulatory phonology employs constriction gestures as basic units. Gestures as constriction actions are defined by 8 vocal tract variables at 5 constricting devices along the vocal tract – 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 (onset and offset times) and dynamic parameters (target/stiffness/damping) are represented in a gestural score. A gesture defined for a vocal tract variable involves its corresponding articulators and some articulators can be shared by different gestures [1]. The task-dynamic speech production model [13] provides a mathematical implementation of the gesture-to-articulator mapping, and generates vocal tract (constriction) variables and articulator time functions from the gestural score for a given utterance. The tract variable time functions, which shape the acoustics of speech, are regulated by time-varying gestural dynamics parameterized by the target and stiffness parameters of the constriction gestures.

Speech recognition systems employing traditional phones as basic units suffer from the failure to capture direct relations to the corresponding phonetic variations such as coarticulation and reduction. We previously proposed "gestural pattern vector" (GPV) to encode discretized instantaneous gestural activation across tract variables [9]. A GPV (Figure 1) contains the discretized dynamic parameters (constriction target, stiffness) for existing gesture activation at each time frame. The speech gestural score, which is an ensemble of gestures distinctive to speech content, can be approximated by a sequence of GPVs.

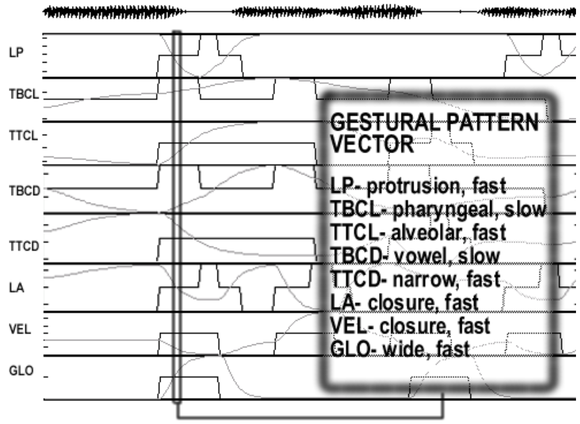

Figure 1: Tract variable time functions (the curves), gestures (the steps) and the gestural pattern vector defined on one frame (5ms) of the utterance "affirmative".

3. GPV-based speech recognition

We propose a speech recognition framework, illustrated in Figure 2, as an alternative to the classic sequence-of-phones model. The proposed framework uses speech gestures as the invariant representation of human speech. Although the detailed timing of gestural activation changes with context, the set of involved gestures is relatively invariant. The ensemble of gestures is approximated by a sequence of GPVs. To classify recognized GPV sequences into words, we use GPV sequence models trained on canonical GPV sequences created from canonical gestural scores for each vocabulary word.

Speech recognition finds the word, w, with Maximum a Posteriori probability:

$$\begin{aligned} W &= argmax_i P(W_i|O) && (1)\\ &\approx argmax_i p(GPVseq_i, W_i, O)/p(O) && \\ &= argmax_i p(GPVseq_i, W_i, O), && (2)\end{aligned}$$

where $p(GPVseq_i, W_i, O)$ is the joint probability of the i^{th} word, the recognized GPV sequence $GPVseq_i$, and the observation O. $GPVseq_i$ is the hypothesis obtained by Viterbi decoding using the GPV sequence model $GPVSeqMdl_i$ for the vocabulary word W_i.

If the priors for different words are assumed to be uniform, i.e., $p(W_i) = p(W_j)$,

$$W \approx argmax_i p(O, GPVseq_i|W_i), \qquad (3)$$

where $p(O, GPVseq_i|W_i)$ is the joint likelihood of the observation and the GPV sequence recognized using the tandem model and the GPV sequence model for the i^{th} word,

$$\begin{aligned} & p(O, GPVseq_i|W_i) && (4)\\ = \; & p(O|GPVseq_i, W_i) * p(GPVseq_i|W_i) && \\ = \; & \prod_{n=1}^{N} p(O_n|GPV_n) * p(GPVseq_i|W_i), && \end{aligned}$$

where $GPV_n,\ n \in \{1, .., N\}$ constitute the GPV sequence $GPVseq_i$.

Equations 3 and 4 indicate that word classification has been converted to a series of GPV sequence recognition problems. Each uses a word-specific GPV sequence model, and results in a score that can be decomposed into confidence scores of recognized GPVs and likelihoods of the GPV sequence on the particular word.

3.1. Hybrid ANN-GMM likelihood model for GPV

Previous work has reported successful recovery of the tract variable time functions from speech acoustics[6, 12]. Here we briefly present the models used to recognize individual GPVs from the tract variable time functions near the local time of interest. For details, please refer to our earlier work [9].

Classification of the GPVs is achieved using an artificial neural network and Gaussian mixture (ANN-GMM) hybrid model, given the tract variable time functions in local time windows centered at the concerned GPV. The hybrid model uses a discriminatively trained artificial neural network (ANN) to estimate posterior probabilities across all GPVs, $\vec{P}(GPV|O)$, which are then subject to a two-step transform F and used as input features to Gaussian mixture models (GMM). The transform F first applies $log\left(\frac{1-\vec{P}(GPV|O)}{1+\vec{P}(GPV|O)}\right)$ and then adopts Principal Component Analysis (PCA) for decorrelation and dimension reduction.

The GMM estimates the following likelihood for use in the proposed framework,

$$p(O_n|GPV_n) \approx p(F(\vec{P}(GPV_n|O_n))|GPV_n). \qquad (5)$$

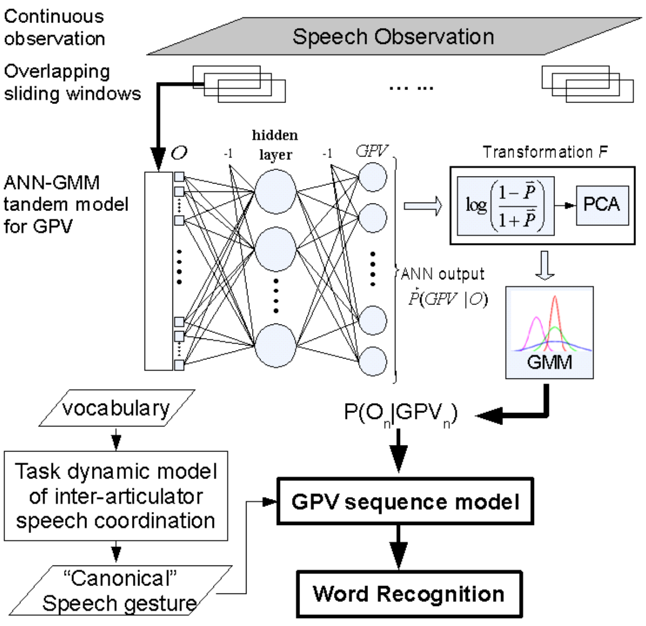
Continuous observation
Speech Observation
Overlapping sliding windows
… …
O
-1
hidden layer
-1
GPV
Transformation F
ANN-GMM tandem model for GPV
$\log\left(\frac{1-\vec{P}}{1+\vec{P}}\right)$
PCA
ANN output $\vec{P}(GPV|O)$
GMM
vocabulary
$P(O_n|GPV_n)$
Task dynamic model of inter-articulator speech coordination
GPV sequence model
"Canonical" Speech gesture
Word Recognition

Figure 2: Speech recognition framework based on GPV.

3.2. GPV sequence recognition

Although the ensemble of gestural activations tends to be distinctive to words, their timing, both intergestural and intragestural, can vary as a function of prosody or performance (e.g., rate, casualness). The recovery of the gestural score is therefore of more interest than the classification of individual GPVs. gestural score recovery can be tackled as GPV sequence recognition. As discussed earlier in this section, word classification is also a series of GPV sequence recognition problems. In particular, these are GPV sequence recognizers informed of word-specific characteristics in sequential information.

Given the characteristics of the gestural score, we need a GPV sequence model that can capture statistics about gestural activations, but is not too sensitive to errors in the GPV sequence output by the hybrid model. N-gram has been widely used as the language model for speech recognition as well as other natural language processing applications. Different from its usual usage, we adopt N-gram as the GPV sequence model to approximate the joint probability of GPV sequences. While there are other options for sequence modeling, using an N-gram GPV sequence model has its own merits. First, it captures frequencies of different GPV types as well as local GPV sequence patterns, while allowing shifting and order swapping of different portions of a GPV sequence. Second, it is computationally inexpensive, and is comparatively robust for training on a small dataset.

The word-independent or word-specific N-gram GPV sequence models are trained using GPV sequences from all training utterances or a particular word respectively. The former captures the general characteristics of a 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, therefore it can be used to distinguish between different words. To maintain robustness and smoothness of the word-specific N-gram GPV sequence models, they are interpolated with the word-independent model.

We use a task dynamic model of inter-articulator speech coordination, implemented in the Haskins Laboratories speech production model TADA [10], to generate gestural scores for speech utterances. In this model, orthographic 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. Using the gestural regimes and intergestural coupling, the intergestural timing model in TADA generates gestural scores including intergestural timing information. These gestural scores are converted to GPV sequences for training the GPV sequence models.

For word recognition (Figure 2), the task dynamic model of inter-articulator speech coordination provides a canonical gestural score for each vocabulary word, which is then used to build the word-specific GPV N-gram model as a way to encode pronunciation.

In the context of gestural score recovery, GPV sequence recognition with the word-independent GPV N-gram model outperforms a set of independent GPV classification tasks, by expressing the relative likelihoods of different GPV sequences.

4. Speech Gesture Dataset

For the pilot experiments reported in this paper, we use a speech dataset synthesized using TADA [10]. This dataset has all the following: acoustics, tract variable time functions, gestures and lexical representation. TADA generates articulatory and acoustic outputs from orthographical input. The gestural score is synthesized in the way described in Section 3.2. The task-dynamic model in TADA takes the gestural score and outputs the tract variable and articulator time functions, which are further mapped to the vocal tract area function (sampled at 200 Hz), and eventually speech acoustics. The dataset contains the same 416 words as in the Wisconsin articulatory database [14] for collaboration reasons.

As mentioned earlier, this work takes tract variable time functions as observations. The synthesized data is used in this pilot experiment to illustrate the concept, and will be used in future work to bootstrap the same models for real speech data.

5. Experiments

5.1. Gestural pattern vectors

Similar to our earlier work [9], we sample the above dataset at 200Hz and obtain the true GPV for each frame, according to the instantaneous activations in the gestural scores. To define a set of frequent GPVs, we randomly split the dataset into three folds and adopt only those GPVs that appear at least 20 times in any two folds. This results in 145 distinctive GPVs, plus a special "unknown" GPV, which accounts for less than 10% of the data that do not correspond to the frequent GPVs.

5.2. Experiment setup

The dataset is 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 test sets respectively. The ANN has 81 hidden nodes and PCA reduces the dimensionality of the transformed features from 146 to 80.

The first experiment is to recover the gestural activation, i.e., the discretized gestural scores. The ANN-GMM tandem models for GPVs and a word-independent GPV sequence bigram model are trained using the training set. These models are used by a Viterbi algorithm as discussed in Section 3. The same experiment is repeated with a uniform ergodic GPV sequence model, in which any GPV can follow any GPV with uniform probability. That is equivalent to classifying each individual GPV independently. The performance of the gestural activation recovery is measured by comparing the recovered GPV sequence with the "canonical" GPV sequence. In particular, we calculate the frame-level F-score for the discretized dynamic parameters used to define the GPVs. Note that only activations in the 145 frequent GPVs are considered.

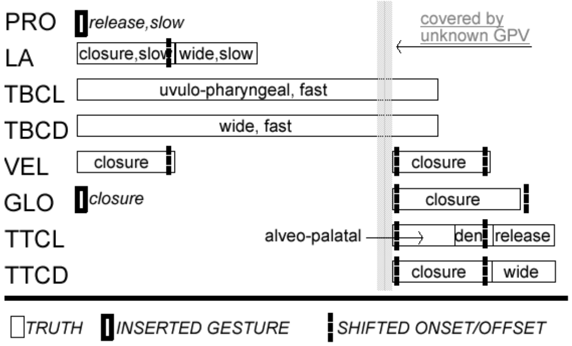

Figure 3: Recovered gestural score for the word "but". (In this example of the recovered gestural score, deviations from the ground truth include insertion of two nonexistent short gesture activations and shift of some onset/offset times.)

The second experiment is to classify the 139 words in the test set, which don't overlap with the 277 words in the training set. Word-specific GPV sequence bigram models are interpolated with the word-independent GPV sequence bigram model with different interpolation weights. To reduce the computational cost of Viterbi decoding using 139 different GPV sequence bigram models, we first generate GPV lattices with two tokens in each step using the word-independent GPV sequence bigram model, and then use different word-specific models to rescore the lattices and extract the best paths with corresponding joint likelihood scores.

5.3. Experiment results

In Figure 3, we present an example of the truth gestural score and the recovered gestural score. Both gestural scores are approximated by GPV sequences. A couple of the truth GPVs are not among the frequent GPVs, therefore assigned to the special "unknown" type.

Table 1: F-score (%) of recovered discretized gestural activation ("Targ": constriction targets; "Stif": constriction stiffness).

Sequence Model		Uniform ergodic	GPV bigram
Targ&Stif		76.38	81.24
Targ		73.51	79.07
Stif		80.79	84.50
Targ	PRO	78.03	84.83
	LA	69.44	77.28
	TBCL	78.56	82.90
	TBCD	83.16	86.01
	VEL	64.79	75.21
	GLO	62.80	72.34
	TTCL	64.14	69.14
	TTCD	63.05	68.44
Stif	PRO	78.46	85.24
	LA	69.99	77.36
	TBCL	83.41	85.90
	TBCD	83.43	85.91

Table 1 presents the F-score of the recovered discretized dynamic parameters, i.e., constriction targets and stiffness, that are used to define the GPVs. The word-independent bigram GPV sequence model outperforms the uniform ergodic sequence model, i.e. independent classification of each GPV in a sequence.

Table 2 presents the word classification accuracy as a function of the N-gram interpolation ratio (the ration between contributions of word-specific and word-independent models). Setting the interpolation ratio to 3:7 yields the best performance. This demonstrates that the word-independent model provides helpful general sequence information about GPVs that is not present in the word-specific training models. As expected, when the ratio grows too high, e.g., to 7:3, the resulting model loses its ability to discriminate between different words.

Table 2: Word classification accuracy(%) with varying interpolation ratio when building the word-specific GPV sequence models.

Interpolation	1:49	3:7	1:1	7:3
Accuracy	84.17	86.33	85.61	48.92

6. Conclusion & Discussion

We propose a framework leveraging articulatory phonology for speech recognition. Given a speech observation, recognizing the sequence of "gestural pattern vectors" recovers the ensemble of gestural activations. Word classification is achieved using an N-gram model of the ensemble of gestural activations. For each word in the vocabulary, a task dynamic model of inter-articulator speech coordination generates a canonical gestural score, used for training word-specific GPV sequence models. These models and an ANN-GMM tandem model for GPVs are used in a Bayesian speech recognizer. A pilot experiment is carried out on synthesized data from one speaker. Given the tract variable time functions, about 80% of the gestures, i.e., the discretized dynamic parameters in the instantaneous gestural activations, are correctly recovered, and word classification accuracy is over 85% for a vocabulary of 139 words not seen in the training data. These results suggest that GPV sequences might be a viable alternative to phone sequences for speech recognition.

Speech gestures, though represented as GPV sequences, are unlike the sequence-of-phones model used in most speech recognizers. The recovered ensemble of gestures is a phonological and phonetic representation distinctive to the content of speech, as the onset and offset times for individual gestural activations naturally encode co-articulation and reduction. Possible improvement over the presented N-gram GPV sequence model could result from collecting statistics separately for subgroups of the tract variables. Deng's lab explored multiple ways to use overlapping articulatory features as sub-word units [4] and to predict spreading of the overlapping features [15]. It would be interesting to see how similar approaches work for matching ensembles of gestures.

Our future work also includes combining this work with recent work by our collaborater [12], and applying the presented framework to real speech data, which will in turn enable more sophisticated GPV sequence models.

7. Acknowledgements

This research is partially funded by NSF grants IIS-0703624, IIS-0703048 and IIS-0703782. The authors would like to thank Vikramjit Mitra for assistance with the dataset.

8. References

[1] 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.

[2] ——, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[3] 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.

[4] 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

[5] 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.

[6] 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.

[7] B. S. Atal, "Efficient coding of lpc parameters by temporal decomposition," in *Proceedings ICASSP*, 1983, pp. 81–84.

[8] T. P. Jung, A. K. Krishnamurthy, S. C. Ahalt, M. E. Beekman, and S. H. Lee, "Deriving gestural scores from articulator-movement records using weighted temporal decomposition," *Ieee Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 2–18, 1996.

[9] 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.

[10] 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.

[11] "TADA: An enhanced, portable task dynamics model in matlab," http://www.haskins.yale.edu/tada_download/index.html.

[12] V. Mitra, I. Y. Ozbek, H. Nam, X. Zhou, and C. Espy-Wilson, "From acoustics to vocal tract time functions," in *Proc. ICASSP*, Taipei, 2009.

[13] 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.

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

[15] 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.