interspeech 2008 incorporating SST 08

The Entropy of the Articulatory Phonological Code: Recognizing Gestures from Tract Variables

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

[1]Beckman Institute, Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, U.S.A.
[2]Haskins Laboratories, Yale University, U.S.A.

xzhuang2@uiuc.edu,nam@haskins.yale.edu
jhasegaw@ad.uiuc.edu,goldstein@haskins.yale.edu,saltzman@haskins.yale.edu

Abstract

We propose an instantaneous "gestural pattern vector" to encode the instantaneous pattern of gesture activations across tract variables in the gestural score. The design of these gestural pattern vectors is the first step towards an automatic speech recognizer motivated by articulatory phonology, which is expected to be more invariant to speech coarticulation and reduction than conventional speech recognizers built with the sequence-of-phones assumption.

We use a tandem model to recover the instantaneous gestural pattern vectors from tract variable time functions in local time windows, and achieve classification accuracy up to 84.5% for synthesized data from one speaker. Recognizing all gestural pattern vectors is equivalent to recognizing the ensemble of gestures. This result suggests that the proposed gestural pattern vector might be a viable unit in statistical models for speech recognition.

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

1. Introduction

A sequence of segmental phonological units, e.g., phones, is the most widely used representation of speech. Since these abstract units are not allowed to temporally overlap with each other, they have difficulty accounting for the phonetic variations such as coarticulation and reduction. In contrast, articulatory phonology represents speech as an ensemble of gestures, which are relatively invariant units that can generate coarticulated or reduced speech when they overlap in time[1, 2].

Current state-of-the-art speech recognition systems adopt the sequence-of-phones representation of speech. These systems work much better for clearly articulated speech, such as broadcast news, than for conversational speech, which sees frequent reduction and coarticulation. Various approaches have been proposed to offset these problems, usually adding complexity to the system design. One alternative is to use more efficient basic recognition units instead of the traditional phones.

The research community has reported works on speech recognition using knowledge about speech production [3, 4, 5]. See [6] for a comprehensive review. There have been studies on recovering gestural activation intervals from acoustic signals or articulatory movements using temporal decomposition method [7, 8], but only intervals of gestures are recovered without the target and stiffness of their control regimes. Livescu et al.[5] proposed recovering gesture ensembles as the state variables in a dynamic Bayesian network. Our work is similar in philosophy with [5], but different in model design and all other computational details.

In this paper, we describe preliminary steps toward developing an automatic speech recognizer that is motivated by articulatory phonology, using so-called "instantaneous gestural pattern vectors" rather than phones, as its sub-word units. Each gestural pattern vector encodes instantaneous information across various tract variables in the gestural score, and includes the constriction targets and stiffnesses associated with the gestural activations existing at a particular time in all tract variables. These sub-word units together recover the ensemble of gestures, which tends to be distinctive to the uttered lexical items. This paper presents the design of these sub-word units.

A tandem model is used to recover the instantaneous gestural pattern vector from tract variable time functions in a local time window. By using the "gestural pattern vector", defined with gesture activations in all tract variables, our statistical models learn an irreducibly multivariate transformation between tract variable observations and their underlying gesture targets and stiffnesses. We achieve classification accuracy up to 84.5% for synthesized data from one speaker. This result suggests that the proposed gestural pattern vector might be a viable unit for statistical models of speech recognition.

2. Articulatory phonology and speech gesture

Traditional phonology has viewed segments/phonemes as primitive phonological units and speech as their sequential concatenation. Since these units are symbolic and are not allowed to temporally overlap with each other, they can not appropriately account for phonetic variations such as coarticulation and reduction in various supra-segmental contexts. On the other hand, articulatory phonology, in which constriction gestures along vocal tract are invariant units, represents speech as an ensemble of gestures[1, 2]. 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). The activation interval (onset and offset times) and dynamic parameter specifications of constriction gestures and intergestural timing patterns are represented in a gestural score

(Figure 1). Each gesture involves its corresponding articulators and some articulators can be shared by different gestures. For example, jaw is an articulator shared by LP, LA, TTCL, TTCD, TBCL and TBCD gestures. The task-dynamic speech production model[9] provides a mathematical implementation of the gesture-to-articulator mapping, and generates constriction tract variable and articulator time functions from gestural score input for a given utterance.

The key to articulatory phonology is that the theory simultaneously captures both cognitive/discrete and physical/continuous characteristics of speech by posing constricting actions as primitive units. Since gestures are action units, they are intrinsically allowed to overlap with one another in time (Figure 1) unlike traditional units (segments/phonemes) occupying pre-allocated time slots. In addition, gestures can be modulated in time and space as a function of concurrent gestures or prosodic context while maintaining their intrinsic invariance.

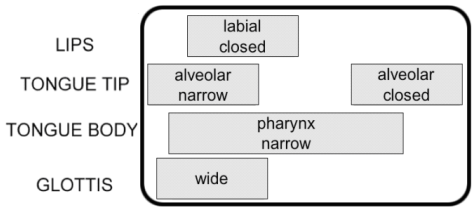

Figure 1: Gestural score for the word "spad".

Low performance of speech recognition systems using a set of traditional phonemes can be attributed to inefficiency of the basic recognition units, which intrinsically fail to capture direct relations to the corresponding phonetic variations. Employing articulatory gestures as sub-word units in recognition would allow us to account for such variations as natural outcomes of simple modulations of gestural patterns or parameter values, maintaining the units' invariance and lexical distinctiveness.

3. Gestural pattern vectors

We propose to use a "gestural pattern vector" to encode instantaneous gestural information across various tract variables in the gestural score. A gestural pattern vector (Figure 2) is defined by the constriction targets and stiffnesses associated with the gesture activations existing at a particular time across all tract variables.

According to articulatory phonology, the tract variable time functions, which shape the acoustics of speech, are regulated by time-varying articulatory dynamics parameterized by the constriction targets and stiffnesses of gesture activations. Although the ensemble of gesture activations tends to be distinctive to words, their timing, both intergestural and intragestural, can vary as a function of prosodic or performance (e.g., rate, casualness) context. This results in significant variation in speech gestural score as well as tract variable time functions. To find a quasi-atomic unit set with reasonable size, instead of defining these units on the whole speech gestural score, we define gestural pattern vectors only using gesture activation information existing at the current time frame. In this framework, the speech gestural score is represented by a sequence of gestural pattern vectors. By recognizing the gestural pattern vector sequence, we can obtain the speech gestural score, in particular, the ensemble of gesture activations, which is distinctive to words.

Different tract variable time functions are correlated, particularly when there is no gesture activation in some tract variable.

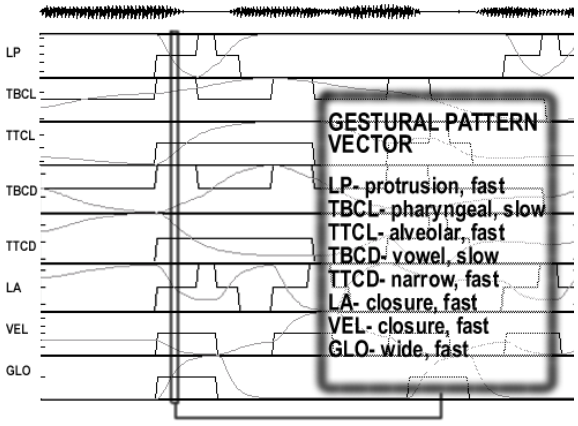

Figure 2: Tract variable time functions(the curves), gestures(the steps) and the gestural pattern vector defined on one frame(5ms) of the utterance "affirmative".

Table 1: Cardinalities of the non-null targets and stiffnesses of the eight tract variables.

	Target	Stiffness		Target	Stiffness
LP	1	2	TBCD	5	2
VEL	2	1	TBCL	4	1
GLO	2	1	TTCD	3	1
LA	5	2	TTCL	4	1

Speech gesture activation across all tract variables is used to define gestural pattern vectors. Note that because of the correlation of speech gesture activation across different tract variables, the number of different gestural pattern vectors is much smaller than the product of the cardinalities of constriction targets and stiffnesses in all tract variables.

Overlapping gesture activations within one tract variable is equivalent to a gesture activation, with target and stiffness being the weighted average of their counterparts in the original set of overlapping gestures. Since the resultant tract variable time functions are identical, we construct gestural pattern vectors from the equivalent gesture activation with the averaged parameters.

The cardinalities of the non-null settings of target and stiffness for each tract variable are shown in Table 1. Each variable can also take a value of "null", which means no active gesture for that tract variable. The cardinalities are determined such that 1) they are not too high to be used in defining a relatively small gesture unit set, and 2) they preserve the most important distinctions corresponding to human perception of the language [10].

4. Recognition of gestural pattern vectors

Articulatory phonology understands speech as an ensemble of gestures. Although the detailed timing of gesture activation change with context, the set of involved gestural pattern vectors is invariant. With successful recognition of gestural pattern vectors, we might be able to recover the ensemble of gestures, which could further be used to reveal the content of the utterance, even in challenging situations such as speech reduction and coarticulation.

Previous work has worked on recovering the tract variable time functions from speech acoustics[6]. However, no work has been done to recognize the speech gestures from the tract variable time functions. In this work we design various statistical

models to recognize gestural pattern vectors from tract variable time functions.

Recognition of gestural pattern vectors is essentially a classification problem in which we attempt to extract the gestural pattern vector, therefore the speech gesture activation information, including constriction target and stiffness, in all tract variables at any particular time, from the values of the tract variable time functions near that time. We use, as observation, tract variable time functions in a local time window centered at the target gestural pattern vector. The challenges are at least three-fold: 1) The speech gesture activation is not perfectly synchronized with tract variable time functions. The speech gesture activation at an immediately preceding time often has a strong impact on the articulator dynamics, therefore the tract variable time functions at the current and following times. 2) The dynamic model demands smoothness in some sense, so that the tract variable time functions have high correlation within a time-local neighbourhood.3) The tract variable time functions correlate across different tract variables, particularly when gesture activation is absent at some tract variable. Also, noncritical articulators tend to have higher associated variance.

We use a tandem model to classify the gestural pattern vectors, given the observed tract variable time functions in local time windows. As shown in Figure 3, a tandem model uses a discriminatively trained artificial neural network (ANN) to estimate posterior probabilities across all gestural pattern vectors, which are then used as input features to Gaussian mixture models (GMM). All the observations from the time functions across all tract variables from the time window are concatenated into a single observation vector X, as the input to the ANN. The observations are normalized with the global mean and variance of the corresponding tract variable time functions. The output nodes O of the ANN correspond to different gestural pattern vector types G, each indicating the posterior probability of one gestural pattern vector type $P(G|X)$, given the current observation X. These ANN outputs are transformed into a new feature using $log\left(\frac{1-O}{1+O}\right)$ 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 gestural pattern vector type.

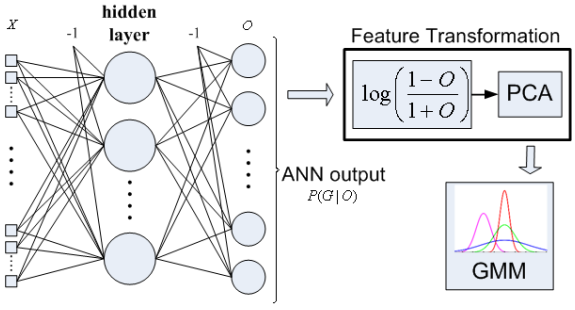

Figure 3: Classification using a tandem model(ANN+GMM).

When testing, we present the observation X to the input nodes, and perform classification using two different approaches. The first approach only uses the ANN output directly: the gesture pattern vector type that corresponds to the output node with highest posterior probability $P(G|X)$ is the classification output. The second approach uses the whole tandem model by choosing the gestural pattern vector whose GMM gives the highest likelihood for the new feature obtained by processing the current observation using the ANN and feature transform.

5. Speech Gesture Dataset

In this study, we use a speech dataset synthesized by Haskins Laboratories speech production model, TADA[10]. This dataset provides reasonable synthesized speech with all the following corresponding information: acoustics, tract variable functions, gestures and lexical representation. TADA generates articulatory and acoustic outputs from orthographical input. In the 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. Using the gestural regimes and intergestural coupling, the intergestural timing model in TADA generates gestural scores including intergestural timing information. 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 (synthesized by Sensimetrics HLSyn[11], 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. TADA was used to synthesize 363 words in Wisconsin articulatory database.

6. Experiments

6.1. Gestural pattern vectors

We go through the above dataset frame by frame at 200Hz and label each frame with gestural pattern vector defined in Section 3, i.e., each frame being an instance of a gestural pattern vector type. In the current dataset, 380 different gestural pattern vector types show up.

Some gestural pattern vector types in the dataset are very rare. We exclude frames labeled as rare gestural pattern vector types for the following reasons: 1) Most rare types do not bear significant speech production implications compared to the frequent types. 2) Rare types could hardly be learned by a statistical model. For the classification experiment, from all the 380 gestural pattern vector types, only those types with at least 30 instances are considered, and only those frames labeled as these considered gestural pattern vector types are used. This results in 181 gestural pattern vector types (37689 frames), reduced from all the 380 types (39882 frames) that are observed at least once in the dataset. Only less than 6% frames are excluded from the dataset.

6.2. Experiment setup

The frames of the 181 gestural pattern vector types in the dataset are divided into training set (about 12,600 frames)and testing set(about 25,000 frames) without overlapping, both having similar gestural pattern vector type distributions. The neural network is trained on the training set using the back propagation algorithm. The gestural pattern vector GMMs are trained on the neural network outputs on the training set. One GMM is first trained on the whole training set, then adapted to different gestural pattern vector GMMs respectively according to Maximum A Posterior criteron. Both approaches in Section 4 are applied to the classification task on the testing set.

The observations, i.e., input X to the neural network, are values of the eight tract variable time functions over a local time window of w frames, normalized by the mean and standard deviation within each tract variable. Different window

lengths $w \in \{5, 7, 9\}$ are applied in this study. The number of hidden nodes in the neural network is chosen from $\{36, 45, 54, 63, 72, 81, 90, 99\}$. PCA reduces the dimensionality of the new features from 181 to 80.

6.3. Experiment results

In Figure 4, We present the performance of recognizing gestural pattern vectors from tract variables only using the neural network, with different local time window lengths w and different numbers of hidden nodes h.

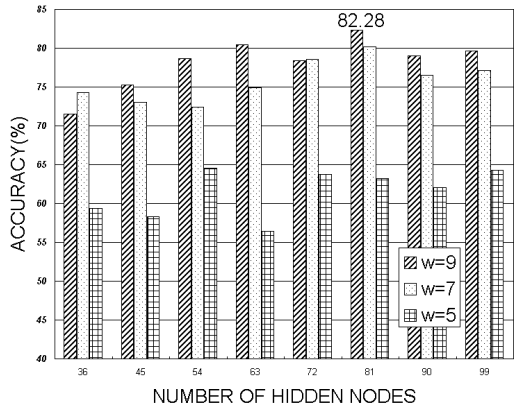

Figure 4: Recognizing gestural pattern vectors using ANNs.

Among the different local time window lengths used, window length of nine gives the best performance. Performance improves with increased number of hidden nodes, which corresponds to the complexity of the neural network, until the number of hidden nodes go beyond 81, when the neural network is too complex to be robustly trained on the current dataset.

In Figure 5, We present the classification performance using the complete tandem models with different numbers of Gaussian components ($Tandem1$: 1000, $Tandem2$: 1500). It is observed that the tandem model gives additional improvement over the best neural network performance.

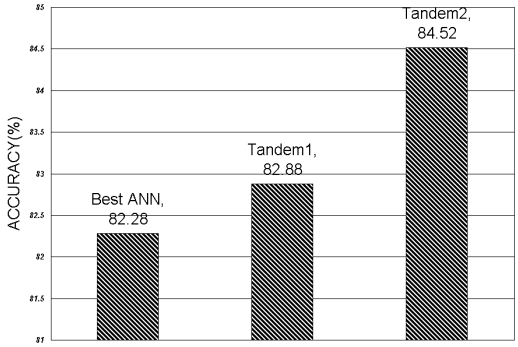

Figure 5: Recognizing gestural pattern vectors using tandem models.

7. Conclusion & Discussion

The instantaneous "gestural pattern vector" is proposed as a sub-word unit for encoding gesture activation information across tract variables. We use a tandem model combining artificial neural network and Gaussian mixture models to recover the instantaneous gestural pattern vectors from tract variable time functions in local time windows, and achieve classification accuracy up to over 84.5% for synthesized data from one speaker. This result suggests that the proposed gestural pattern vector might be a viable sub-word unit for statistical models of speech recognition.

The design of these gestural pattern vectors is the first step towards an automatic speech recognizer motivated by articulatory phonology. In such a speech recognizer, speech would be recognized by recovering the ensemble of gestures.

The ensemble of speech gestures is recognized by classifying each frame into a gestural pattern vector type. This has different implications than the sequence-of-phones model used in most current state-of-the-art speech recognition systems, in that our recovered ensemble of gestures is a phonological and phonetic representation distinctive to the content of speech. Although the ensemble of gestures is recognized via a sequence of gestural pattern vectors, we may view the recognized gestures in a way different from this, taking advantage of the invariant properties as suggested by articulatory phonology. We plan to pursue this issue in future study.

8. Acknowledgements

This research is partially funded by NSF grant IIS-0703624, NIH grant DC-02717 and NSF grant IIS-0703782.

9. 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] 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.

[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] Sensimetrics Corp, "High level parameter speech synthesis system," `http://sens.com/hlsyn/`.