INTERSPEECH 2009 BRIGHTON

Estimation of articulatory gesture patterns from speech acoustics

Prasanta Kumar Ghosh[1], *Shrikanth Narayanan*[1], *Pierre Divenyi*[2], *Louis Goldstein*[3], *Elliot Saltzman*[4]

[1]Department of Electrical Engineering, University of Southern California, LA, CA, 90089
[2]EBIRE, Martinez, CA 94553
[3]Department of Linguistics, University of Southern California, LA, CA, 90089
[4]Haskins Laboratories, New Haven, CT 06511

prasantg@usc.edu, shri@sipi.usc.edu, pdivenyi@ebire.org, louisgol@usc.edu, esaltz@bu.edu

Abstract

We investigated dynamic programming (DP) and state-model (SM) approaches for estimating gestural scores from speech acoustics. We performed a word-identification task using the gestural pattern vector sequences estimated by each approach. For a set of 75 randomly chosen words, we obtained the best word-identification accuracy (66.67%) using the DP approach. This result implies that considerable support for lexical access during speech perception might be provided by such a method of recovering gestural information from acoustics.

Index Terms: gestural patterns, acoustic to gesture inversion

1. Introduction

The idea that the perception and production of speech are intricately bound to one another is not new. To be more precise, the motor theory of speech perception, proposed by Liberman and his colleagues [1, 2] hypothesizes that the perceptual code of speech is not acoustic but articulatory: the acoustic signal of speech only helps the listener's auditory-nervous system to uncover the articulatory gesture performed by the talker. Such a viewpoint helps to explain that although a speech sound, such as the phoneme /d/, is always perceived as /d/ no matter how diverse the phonetic context in which it appears is, its acoustic form will exhibit large variations as a function of the context. Although the motor theory has undergone significant evolution since its inception, its basic premises have remained valid (for an excellent review, see [3]), even as speech science and technology evolved rapidly, fueled by the growth of computer technology. The same is true for the concept of analysis by synthesis, first proposed by Halle and Stevens in 1962 [4]. Contemporary cognitive science and neuroscience have validated the existence of links between sensory and motor mechanisms active during speech perception [5] as well as neural activity in pre-motor cortical areas specializing in speech production when someone listens to speech [6, 7].

The link in speech science and technology between the acoustic structure of speech perception and the articulatory patterns that gave rise to these acoustics was first explored in the design of speech synthesizers based on articulatory knowledge. Significant progress in the development of such articulatory synthesizers was made at the Bell Laboratories from the 1960s on [8, 9, 10, 11]. This work successfully employed a mapping from measured [12, 13] speech articulator time-functions to a readily understandable, and often excellent quality, acoustic signal. However, investigators did come to realize that the problem of acoustic-to-articulatory transform is a difficult one. The mapping between formant frequencies and vocal tract area functions is ill-posed, as it is not unique. However as early as in the 1970s attempts have been made to obtain an inverse transform from speech acoustics to the underlying articulatory gestures using additional constraints: articulatory, dynamic, and continuity-mapping [14, 15, 16, 17]. Several methods were explored, such as codebooks pairing speech corpora consisting of parallel acoustic and articulatory representations [18, 19], neural networks [20], and HMMs or other stochastic methods [21].

Previous work [18], which we expand here, is based on a codebook approach that differs from those of other investigators in several ways. The most important difference is that we adopt an analysis-by-synthesis procedure by using the Haskins Laboratories task-dynamics (TADA) articulatory synthesizer [22] to produce the samples in both the training and the test sets. This synthesizer generates gestural motion patterns in a task space of eight vocal tract constriction variables (tract variables). Intergestural coarticulation is produced as the simple consequence of temporal overlap among these motion patterns.

2. Problem Definition

Traditional phonology represents speech as a sequential concatenation of primitive phonological units, phonemes; each phoneme has its own acoustic properties in a given context. On the other hand, articulatory phonology represents speech as an ensemble of gestures [23, 24]. Gestures are defined as dynamical control regimes for constriction actions in 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)). According to articulatory phonology, the tract variable time functions, which shape the acoustics of speech, are regulated by a gestural score, that is composed of (roughly) step-function-like temporal activation intervals of tract variable dynamical control regimes, each parameterized by a constriction target and a stiffness of the activated gesture. The goal of this paper is to estimate these gestural scores. For our purposes, the gestural score is divided into a sequence of 5 ms slices, and for each slice, a gestural pattern vector [25] is calculated, that encodes the instantaneous gestural control parameters for the tract variables controlled at that time slice.

Work supported by NSF and NIH.

Let us consider $\{\underline{A}_k\}_{k=1}^N$ to be a sequence of acoustic feature vectors derived from given speech signal $x(n)$. Let $\underline{A}_k$ be the feature vector of k^{th} frame, where the frame size is N_w and frame shift N_{sh}. Let $\{\underline{G}_k\}_{k=1}^N$ denote the corresponding gestural pattern vectors. The elements of $\underline{G}_k$ are the gestural parameter values at frame k. We assume that the sequence of gestural pattern vectors provides a good approximation of the respective gestural score. Thus the problem is to estimate $\{\underline{G}_k\}_{k=1}^N$ from given acoustic features $\{\underline{A}_k\}_{k=1}^N$.

3. Proposed Approaches

To estimate $\{\underline{G}_k\}_{k=1}^N$ from independent $\{\underline{A}_k\}_{k=1}^N$, we propose two approaches: a) dynamic programming similar to that reported in [18] and b) a state-model approach where $\{\underline{A}_k\}_{k=1}^N$ defines the observation sequence and $\{\underline{G}_k\}_{k=1}^N$ defines the corresponding state sequence. These two approaches are described below.

3.1. Dynamic programming (DP) approach

In this approach, we find $\{\underline{G}_k\}_{k=1}^N$, which maximizes the following likelihood, given $\{\underline{A}_k\}_{k=1}^N$, after [18]

$$\underset{\{\underline{G}_k\}_{k=1}^N}{\arg\max} \prod_{j=1}^{N} p\left(\underline{G}_j|\underline{A}_j\right) p\left(\underline{G}_j|\underline{G}_{j-1}\right) \tag{1}$$

where $p\left(\underline{G}_j|\underline{G}_{j-1}=\underline{g}\right) = \mathcal{N}\left(\underline{g},\Lambda\right)$ is the normal probability density function (pdf) with mean $\underline{g}$ and diagonal covariance matrix Λ. Lammert et al. [18] showed that such criterion is suitable for deriving solutions to the articulatory inversion problem. Since $p\left(\underline{G}_j|\underline{A}_j\right) = \frac{p\left(\underline{G}_j,\underline{A}_j\right)}{p\left(\underline{A}_j\right)}$, the overall likelihood in two spaces (acoustic and gesture) are considered in the optimization. $p\left(\underline{G}_j|\underline{G}_{j-1}\right)$ is used as additional factor so that the gestural pattern vector varies smoothly from one frame to the next. Λ is estimated from the gestural pattern vector sequences in the training set. The diagonal entries are the variances of the components in the gestural pattern vectors.

This optimization problem is solved by dynamic programming (DP). Note that each gesture variable can take any real value. Thus, for the acoustic features at each frame the possible candidates of gestural pattern vectors provided by the DP algorithm are uncountably infinite. Therefore, we restrict the search space by finding the gestural pattern vectors from the training set, whose corresponding acoustic vectors are in the neighborhood of the acoustic vector of the current frame. The best path among these candidates are obtained to maximize the likelihood in eqn. (1).

3.2. State Model (SM) approach

Although each component of a gestural pattern vector can take any real value, we observe that some gestural pattern vectors are very rare in the dataset. Hence, instead of considering the full range of possible gestural pattern vectors, we quantize them. Let all the gestural pattern vectors be clustered into Q clusters with the mean vectors $\underline{G}^1$, ..., $\underline{G}^Q$. The mean of each cluster is chosen to be the representative pattern vector for all members of the cluster; the result is a finite set of quantized gestural pattern vectors. We denote the quantized pattern vector of frame l by $\underline{G}_l^q \in \left\{\underline{G}^1,\ ...,\ \underline{G}^Q\right\}$.

We assume $\underline{G}^1$, ..., $\underline{G}^Q$ are Q gestural pattern states and $\{\underline{A}_k\}_{k=1}^N$ are observed acoustic vectors. We consider speech production to be characterized by a generative model in which acoustic vectors are generated based on the probability density function $p\left(\underline{A}_k|\underline{G}_k^q\right)$ when the state at frame k is $\underline{G}_k^q \in \left\{\underline{G}^1,\ ...,\ \underline{G}^Q\right\}$. The transition probabilities from one state in frame k to another state in frame $k+1$ is given by $p\left(\underline{G}_{k+1}^q|\underline{G}_k^q\right)$ and, thus, the probability of generating independent $\{\underline{A}_k\}_{k=1}^N$ from a state sequence $\{\underline{G}_k^q\}_{k=1}^N$ is given by $\prod_{k=1}^N p\left(\underline{A}_k|\underline{G}_k^q\right) p\left(\underline{G}_k^q|\underline{G}_{k-1}^q\right)$. We estimate both $p\left(\underline{A}_k|\underline{G}_k^q\right)$ and $p\left(\underline{G}_{k+1}^q|\underline{G}_k^q\right)$ from the training data.

Given an observed acoustic sequence $\left\{\underline{A}_j\right\}_{j=1}^N$, the goal is to find the best quantized gestural state sequence that will maximize a likelihood expressed similarly to eqn (1) but using quantized gestural pattern vectors as follows:

$$\underset{\left\{\underline{G}_k^q\right\}_{k=1}^N}{\arg\max} \prod_{j=1}^{N} p\left(\underline{G}_j^q|\underline{A}_j\right) p\left(\underline{G}_j^q|\underline{G}_{j-1}^q\right) \tag{2}$$

Since $p\left(\underline{G}_k^q|\underline{A}_k\right) \propto p\left(\underline{A}_k|\underline{G}_k^q\right) P\left(\underline{G}_k^q\right)$ (by Bayes' Rule), eqn (2) can be written as

$$\underset{\left\{\underline{G}_k\right\}_{k=1}^N}{\arg\max} \prod_{j=1}^{N} p\left(\underline{A}_j|\underline{G}_j^q\right) P\left(\underline{G}_j^q\right) p\left(\underline{G}_j^q|\underline{G}_{j-1}^q\right) \tag{3}$$

where, $P\left(\underline{G}_j^q\right)$ is the probability of $\underline{G}^q$ at the j^{th} frame, which is also estimated from the training corpus.

The gestural pattern vectors in the training set are used to obtain $\underline{G}^1$, ..., $\underline{G}^Q$ by the K-means algorithm. After quantization, $p\left(\underline{G}_k^q|\underline{G}_{k-1}^q\right)$ is estimated from the sequence of quantized gestural pattern vectors. For each quantized gestural pattern, all corresponding acoustic vectors are used to estimate $p\left(\underline{A}_k|\underline{G}_k^q\right)$.

Given a test speech utterance, i.e., $\{\underline{A}_k\}_{k=1}^N$, we perform the best decoding of the state sequence such that the likelihood in eqn (3) is maximized. This is performed using an approach similar to viterbi-decoding.

4. DataSet and Experimental Setup

In this paper, we use a speech dataset synthesized by Haskins Laboratories speech production model TADA [22]. TADA is a MATLAB implementation of the Task Dynamic model of speech articulator coordination. To synthesize speech using TADA we used 213 natural and phonetically balanced sentences, drawn from the Harvard IEEE Corpus. To input the sentences into TADA, we used the programs capability to receive orthographic input. From within TADA, this orthography is then converted into a syllabified phoneme sequence via a syllabified version of the CMU pronouncing dictionary. The phonemes are represented, in turn, as sets of tract variable controls (usually 1-3 per phoneme), and the syllable frames are used to specify a coupling graph [26], from which the gestural activations (constituting the gestural score) are triggered. From the gestural score, we calculated a gestural pattern vector sequence at a 200 Hz frame rate. The gestural pattern vector in our work is defined (somewhat differently than in [25]) as the constriction target value for each tract variable active at that frame: TTCD, TTCL, TBCD, TBCL, LA, PRO, VEL or GLO. When more than one gesture controls a given tract variable at a given frame, the value is the mean of the target values, weighted by the parameter blending weight specified by TADA. The gestural score then controls articulator motion, which is input to a vocal tract model, that also generates acoustic output (using HLsyn [27]). The acoustic output of is transformed

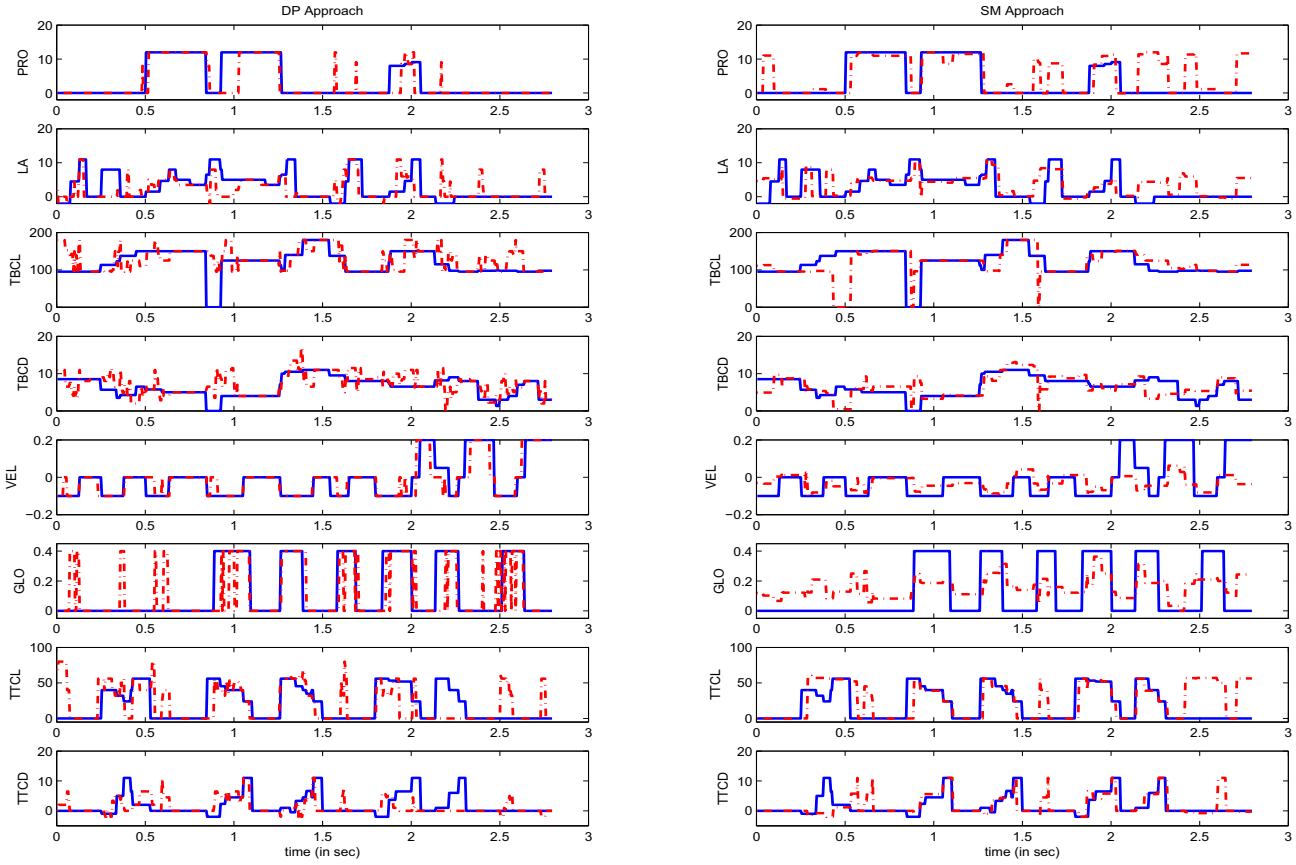

Figure 1: *Illustration of the gestural pattern vector estimates using DP and SM approach. Solid lines are the reference gesture variable trajectories obtained from TADA and dash-dotted trajectories are their estimates using DP approach (left column) and SM approach (right column).*

into 13 Mel Frequency Cepstral Coefficients (MFCCs) [28] using a 10ms window size and 5ms window advance rate. Thus, we obtain 125849 parallel 13-dimensional acoustic vectors and 8-dimensional gestural pattern vectors. Out of these we used 88006 parallel vectors for training (corresponds to 150 sentences) and remaining 37843 vectors for testing (corresponds to 63 sentences).

Among 88006 gestural pattern vectors in the training set, there are 1975 unique gestural pattern vectors and 192 vectors among them cover $\sim$70% of all gestural pattern vectors. Thus, for our experiment we chose Q for SM approach to be 150, 180 and 210.

5. Evaluation and Results

Fig. 1 illustrates the estimated gestural pattern vector sequence obtained using DP and SM approaches. The left column plots the time trajectory of the original gesture variables (solid line) and their estimates using DP approach (dash-dotted line). The right column plots the same but using SM approach. Neither approach performed uniformly for all eight gestural variables. In general, the estimated trajectories follow the basic trend of the reference trajectories for most of the gesture variables. However, for few variables, the estimated trajectories do not match to the reference ones. In particular, due to vector quantization, the estimated GLO using SM approach never matched the reference GLO trajectory, which was not the case for DP approach. Because of vector quantization, the quantized vectors are not necessarily any gestural pattern vectors in the training set and hence estimated trajectories may not match well to the reference trajectories in some cases. To analyze the performance of these two approaches, we derived an evaluation metric over 63 test sentences.

To evaluate the performance of both DP and SM approaches, we manually picked 75 words randomly from the 63 test sentences and performed a word identification experiment based on the estimated gestural pattern vectors for these words.

Let $\mathbf{G}_i$ denote the reference gestural pattern vector sequence for the i-th test word (i=1, ..., 75) and let $\hat{\mathbf{G}}_i$ denote the estimated gestural pattern sequence for the i-th word. For identifying the i-th word, $\hat{\mathbf{G}}_i$ was compared to $\mathbf{G}_j$, j=1, ..., 75, using dynamic time warping (DTW) between the two sequences of gestural pattern vectors using Mahalanobis distance between vectors as distance metric. After obtaining alignment using DTW, the distance between $\hat{\mathbf{G}}_i$ and the reference gestural pattern vector sequence of the j^{th} word was computed, for all $i,\ j \in \{1,\ ...,\ 75\}$. The word for which the reference gesture

pattern was closest to $\hat{\mathbf{G}}_i$ was identified as the i^{th} word. Table 1 shows the percentage accuracy of this word identification task for the different estimation approaches, using MFCC to define the acoustic vectors.

Schemes	% Identification Accuracy
DP	66.67%
SM (Q=150)	52.00%
SM (Q=180)	61.33%
SM (Q=210)	60.00%

Table 1: Word identification accuracy for various gesture pattern estimation approaches.

It can be seen that, in general, the performance of SM approach is worse than that of DP approach. Although gestural activation variables are roughly quantal in nature (i.e., step functions between zero and one), quantization over the time course of the entire gestural score does not necessarily maintain the original quantal nature for each component of the gestural pattern vector sequence. Thus, it appears that the quantization error can diminish the performance of the SM approach. This is evidenced by the fact that, due to quantization error being greater for low Q, the performance of SM drops to 52% for Q=150. On the other hand, although the DP approach is not subject to quantization errors, it constrains the estimated sequence to be as smooth as possible which belies the underlying quantal nature of the gestural activation variables. In spite of that DP approach achieves the best performance (66.67%). This means 50 words out of 75 words were correctly identified. The best performance among SM approaches is obtained for Q=180. Among misidentified words, "friends", "busses" and "sharp" were wrongly identified as "good", "child" and "cars" respectively, for both DP and SM (Q=180) approaches.

6. Conclusions

We presented two approaches for estimating gestural patterns from speech signals. The greatest accuracy (66.67%) in a word identification task was obtained using estimated gestural pattern estimated by the DP approach, suggesting that gestural recovery from acoustics using this method provides considerable information for identifying a word from a set of words.

7. References

[1] Liberman, A. M., "Some results of research in speech perception", *J. Acoust. Soc. Am.*, vol 29, pp 117-123, 1957.

[2] Liberman, A. M., and Mattingly, I. G., "The motor theory of speech revisited", *Cognition*, vol 21, pp 1-36, 1985.

[3] Galantucci, B., Fowler, C. A., and Turvey, M. T., "The motor theory of speech perception reviewed", *Psychonomic Bulletin and Review*, 13(3), pp 361-377, 2006.

[4] Halle, M., and Stevens, K. N., "Analysis by synthesis", *Proceedings on the seminar on speech compression and processing, W. Wathen-Dunn and L. E. Woods, Eds.*, vol 2, paper D7, Cambridge MA: USAF Cambridge Research Center, December 1959.

[5] Galantucci, B., Fowler, C. A., and Goldstein, L., "Perceptuomotor compatibility effects in speech", *Attention, Perception, and Psychophysics*, 71(5), pp 1138-1149, 2009.

[6] Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M., "Linstening to speech activates motor areas involved in speech production", *Nature Neuroscience*, vol 7, issue 7, pp 701-702, 2004.

[7] Pulvermuller, F., Hauk, O., Nikulin, V. V., and Ilmoniemi, R. J., "Functional links between motor and language systems", *European Journal of Neuroscience*, 21(3), pp 793-797, 2005.

[8] Coker, C., "Speech synthesis with parametric articulatory model", *Proc. of the Speech Symposium, Kyoto*, 1968.

[9] Coker, C. H., and Fujimura, O., "A model for specification of vocal tract area function", *J. Acoust. Soc. Am.*, vol 40, pp 1271, 1966.

[10] Coker, C. H., Umeda, N., and Browman, C. P., "Automatic synthesis from ordinary English text", *IEEE Trans. Audio Electroacoust.*, vol AU-21, pp 293-297, 1973.

[11] Olive, J. P., "Rule synthesis of speech from diadic units", *Proc. Int. Conf. Acoust. Speech Signal Procss.*, vol 2, pp 568- 570, May 1977.

[12] Fant, C. G. M., "Acoustic theory of speech production", *The Hague: Mouton*, 1960.

[13] Ladefoged, P., and Harshman, R., "Formant frequencies and movements of the tongue", *UCLA*, vol 45, pp 39-52, 1979.

[14] Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", *J. Acoust. Soc. Am.*, vol 63, pp 1535-1555, 1978.

[15] Ladefoged, P., Harshman, R., Goldstein, L., and Rice, L., "Generating vocal tract shapes from formant frequencies", *J. Acoust. Soc. Am.*, vol 64, pp 1027-1035, 1978.

[16] Wakita, H., "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms", *IEEE Transactions on Audio and Electroacoustics*, vol 21, issue 5, pp 417-427, 1973.

[17] Hogden, J., Rubin, P., McDermott, E., Katagiri, S., and Goldstein, L., "Inverting mappings from smooth paths through Rn to paths through Rm: A technique applied to recovering articulation from acoustics", *Speech Communication*, vol 49, pp 361-383, 2007.

[18] Lammert A., Ellis D. P. W., Divenyi P., "Data-driven articulatory inversion incorporating articulator priors", *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, SAPA 2008, 21 September 2008, Brisbane, Australia.

[19] Richards, H. B., Mason, J. S., Hunt, M. J., and Bridle, J. S., "Deriving articulatory representations of speech with various excitation modes", *Proc. ICSLP*, pp 1233-1236, 1996.

[20] Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zacks, J., and Levy, S., "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data", *J. Acoust. Soc. Am.*, vol, 92, issue 2.1, pp 688-700, 1992.

[21] Dusan S. and Deng L., "Acoustic-to-articulatory inversion using dynamical and phonological constraints", *Proc. 5th Seminar on Speech Production*, pp 237-240, May 2000.

[22] Nam H., Goldstein L., Saltzman E., and Byrd D., "Tada: An enhanced, portable task dynamics model in matlab", *J. Acoust. Soc. Am.*, vol 115, issue 5, pp 2430, 2004.

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

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

[25] Zhuang X., Nam H., Johnson M. H., Goldstein L., and Saltzman E., "The Entropy of Articulatory Phonological Code: Recognizing Gestures from Tract Variables", *Proc. Interspeech*, pp 1489-1492, 2008.

[26] Saltzman, E., Nam, H., Krivokapic, J., and Goldstein, L., "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation", *In P. A. Barbosa, S. Madureira, and C. Reis, (Eds.), Proceedings of the 4th International Conference on Speech Prosody (Speech Prosody 2008)*, Campinas, Brazil, 2008.

[27] Hanson, H.M. and Stevens, K.N., "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn", *J. Acoust. Soc. Am.*, vol 112, pp 1158-1182, 2002.

[28] Ellis D. P. W., "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available:www.ee.columbia.edu/ dpwe/resources/matlab/rastamat/.