A procedure for estimating gestural scores from natural speech

Hosung Nam[1*], *Vikramjit Mitra*[2*], *Mark Tiede*[1,3],
Elliot Saltzman[1,4], *Louis Goldstein*[1,5], *Carol Espy-Wilson*[2], *Mark Hasegawa-Johnson*[6]

[1] Haskins Laboratories, USA
[2] Institute for Systems Research, Dept. of Electrical & Comp. Eng., University of Maryland, USA
[3] Research Laboratory of Electronics, MIT, USA
[4] Department of Physical Therapy, Boston Univ., USA
[5] Department of Linguistics, Univ. of Southern California, USA
[6] Department of Electrical & Computer Engineering, Univ. of Illinois, Urbana-Champaign, USA

nam@haskins.yale.edu, vmitra@umd.edu, tiede@haskins.yale.edu,
esaltz@bu.edu, louisgol@usc.edu, espy@umd.edu, jhasegaw@uiuc.edu

Abstract*

Speech can be represented as a constellation of constricting events, *gestures*, which are defined at distinct vocal tract sites, in the form of a *gestural score*. Gestures and their output trajectories, *tract variables*, which are available only in synthetic speech, have recently been shown to improve automatic speech recognition (ASR) performance. In this paper we propose an iterative analysis-by-synthesis landmark based time-warping architecture to obtain gestural scores for natural speech. Given an utterance, the Haskins Laboratories Task Dynamics and Application (TADA) model was used to generate its prototype gestural score and the corresponding synthetic acoustic output. An optimal gestural score was estimated through iterative time-warping processes such that the distance between original and TADA-synthesized speech is minimized. We compared the performance of our approach to that of a conventional dynamic time warping procedure using Log-Spectral and Itakura Distance measures. We also performed a word recognition experiment using the gestural annotations to show that the gestural scores are suitable for word recognition.

Index Terms: Articulatory Phonology, gestures, vocal tract variables, X-ray microbeam data, TADA model, time warping

1. Introduction

Currently, most ASR systems use tri-phone [1] or quin-phone units to model contextual influence, *coarticulation*. However, such models may suffer from data sparsity as some of the tri- or quin-phone units are very rarely observed in a given database. Such units may fail to capture the coarticulation appropriately because the span of a given tri- or quin-phone's contextual influence is not flexible. On the other hand, Articulatory Phonology views an utterance as a constellation of constricting events (e.g. narrowing lips for /b/ and raising tongue tip for /d/), *gestures*, along the human vocal tract. Gestures are fundamental units that compose an utterance and their arrangement in time can be represented in a gestural score [2]. Each gesture is defined as a critically damped system with a target for a given constricting organs (lip, tongue tip, tongue body, velum and glottis) in the vocal tract [3]. The gesture tasks for each organ are geometrically represented by its tract variables (TV) (see Figure 1 and Table 1 for details). Gestures are allowed to overlap with one another in time and words are distinguished by how the gestures are temporally coordinated as well as what gestures are used. Our recent studies [4, 5, 6] have demonstrated strong potential for using gestural units in robust ASR. However, due to the lack of any speech database with proper gestural annotation, our previous studies have been mostly limited to synthetic data, generated by the Haskins Laboratories Task Dynamic model of speech production, also known as TADA [7]. TADA is a computational model of Articulatory Phonology, which performs the following: (a) creates gestural score for a given word (or phone sequence), (b) computes tract variable time function from the gestural score input, and (c) generates acoustic signal through HLsyn™, a parametric quasi-articulatory synthesizer (Sensimetrics Inc.) [8]. Initial exploration using synthetic speech has shown that: (1) gestures and TVs can be estimated relatively accurately from speech; (2) estimated gestures produce a word recognition accuracy of around 82%; (3) gestures and TVs have both been shown to improve noise robustness of ASR systems [4, 5].

Table 1. *Constriction organs and tract variables*

Constriction organs	Tract Variables (TVs)
Lip	Lip Aperture (**LA**)
	Lip Protrusion (**LP**)
Tongue Tip	Tongue tip constriction degree (**TTCD**)
	Tongue tip constriction location (**TTCL**)
Tongue Body	Tongue body constriction degree (**TBCD**)
	Tongue body constriction location (**TBCL**)
Velum	Velum (**VEL**)
Glottis	Glottis (**GLO**)

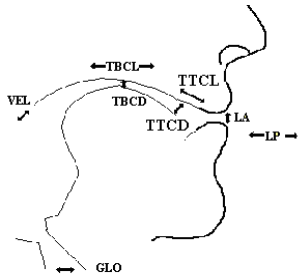

Figure 1. *Geometric schematization of tract variables*

The articulatory information including gestural annotation (gestural scores) on a large speech database would be highly

* The first two authors contributed equally to this work

beneficial not only to speech technology but also to various speech related areas such as phonological theories, phonetic sciences, speech pathology, etc. Several efforts have been made to retrieve such gestural information. Atal [9] showed a way of estimating gestural activation from acoustic signal using temporal decomposition but his method was only focused on gestural activation detection rather than gestural dynamic parameter detection, and his approach was too sensitive to model parameters to obtain a robust estimation. Sun *et al.* [10] proposed an automatic annotation model of gestural scores. Their approach requires manual gestural annotation to train the model, in which potential annotation errors might be introduced by different annotators. Zhuang *et al.* [11] showed that gestural activation intervals and dynamic parameters could be estimated from TVs using a TADA-generated database. Tepperman *et al.* [12] used an HMM-based iterative bootstrapping method to estimate gestural scores but their approach was limited to a small dataset. The fact that no speech database exists with such gestural information despite all these efforts shows how challenging the gestural annotation task is. In this study, we introduce a novel method to annotate gestural scores and estimate the corresponding tract variable time functions from a natural speech database.

2. Database

The *a priori* assumption of our proposed architecture is that the natural speech database upon which it can be implemented should have the phones and words delimited in advance. Our approach can therefore be applied to any speech database with phone-delimited labeling (e.g. TIMIT, Switchboard, Buckeye, etc.). We have chosen the University of Wisconsin X-ray microbeam database (XRMB) [13] for our initial study. The XRMB database contains the time functions of pellets tracked during speech production, which may help in cross-validating the articulatory information generated by our proposed approach. The database includes speech utterances recorded from 47 different American English speakers (25 females and 22 males). Each speaker produces at most 56 tasks, each of which can be either read speech containing a series of digits, TIMIT sentences, or even as large as reading of an entire paragraph from a book. The sampling rate for the acoustic signals is 21.74 kHz. For our study, XRMB utterances were phone-delimited by using a Forced Alignment technique (performed by Yuan, see [14] for more details).

3. Architecture

Manually annotating gestural markups in natural speech is a very difficult task. Compared to phone markups, gestural onsets and offsets are not always aligned with acoustic landmarks. Further, the articulatory gestures, which are constricting actions, do not occur as a seamless sequence but are rather allowed to temporally overlap with one another. This is the major strength of the gestural representation in the sense that it can account for the complex phenomenon of coarticulation, but is also the major obstacle to delimiting gestural onsets and offsets. This is the reason why we aimed for an automated procedure to perform gestural annotation for natural speech.

Given the phone transcript of a natural utterance, $s(t)$ from the XRMB database, the TADA model generates its set of gestures and the corresponding synthetic speech $s_{syn}(t)$ based on its model-driven intergestural timing. $s_{syn}(t)$ will therefore differ from $s(t)$ in time, that is both in rate of speech as well as the individual phone durations. The phone content of $s_{syn}(t)$ and $s(t)$ will be identical as the phone content of $s(t)$ has been used as an input to TADA to create $s_{syn}(t)$. Suppose $D(x, y)$ gives a distance measure between $x(t)$ and $y(t)$, we define a warping scale W on $s_{syn}(t)$ such that $D(s,s_{syn})$ is minimized, bringing $s_{syn}(t)$ closer to $s(t)$, as shown in (1).

$$W_{opt} = \arg\min\left[D(s(t), W(s_{syn}(t))) \right] \quad (1)$$

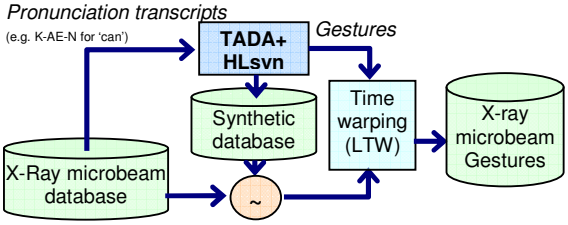

Figure 2. *Block diagram of the overall warping architecture for gesture specification*

Hence, for W_{opt} we can expect the signals $s_{syn}(t)$ and the $s(t)$ to be very similar to each other and how similar they are should be reflected by any distance measure $D(x, y)$. Now, if $G(s_{syn}(t))$ is the gestural score for $s_{syn}(t)$, then $W_{opt}(G(s_{syn}(t)))$ is an optimal gestural score estimated for the $s(t)$ and this is the basic principle behind our iterative time-warping procedure to obtain gestures for natural speech. The overall architecture is shown in Figure 2, which shows that given an utterance $s(t)$ from the XRMB corpus, TADA and HLsyn generate the corresponding synthetic speech $s_{syn}(t)$ and its gestural score, $G(s_{syn}(t))$ based on the phone sequence for $s(t)$. $s(t)$ is then compared with $s_{syn}(t)$ and the comparison information is used to perform non-linear time-warping of the synthetic gestures $G(s_{syn}(t))$ which generates the gestures $\hat{G}(s(t))$, where

$$\hat{G}\left(s(t)\right) = W_{opt}\left(G\left(s_{syn}(t)\right)\right) \quad (2)$$

The time warping shown in Figure 2, is different from the traditional dynamic time warping (DTW) algorithms [15 16]. For the synthetic speech obtained from TADA, $s_{syn}(t)$, the phone boundaries are initially approximated based on its underlying gestural on/offset times. Given the phone boundary sets of both signals, our *landmark-based timing warping* (LTW) is performed by the following three steps. First, comparing the phone boundaries of the natural XRMB utterances $s(t)$ and those approximated for the $s_{syn}(t)$, a time-warping scale W_i (where i represents the warping scale obtained at the i^{th} iteration) is obtained by aligning TADA's phone boundaries to those of natural speech. Second, the time-warping scale is projected onto gestural score $W_i(G(s_{syn}(t)))$; and third, the new gestural score is used by TADA to yield a new acoustic signal $s_{syn,i}(t)$.

From the second iteration onwards, the initial TADA phone boundaries are piecewise modulated (similar to [19]) in steps of 10ms (to a max. of at ± 20 ms), hence for N phones, $4N$ is the maximal number of possible iterations. The synthetic speech $s_{syn,i}(t)$ is generated through the LTW procedure at each iteration i, where only the modulated phone boundaries that result in minimization of the distance measure $D(s,s_{syn,i})$ are accepted. The detailed implementation of the iterative warping strategy is shown in Figure 3, which represents the internal architecture of the "Time Warping (LTW)" block in Figure 2. Figure 3 shows that the proposed iterative LTW procedure is in fact an analysis-by-synthesis approach where the warping scale used to time-warp the gestures (analysis stage) and the time-warped gestures are used to re-synthesize the speech waveform (synthesis stage). Figure 4 compares the XRMB (top), prototype TADA (mid), and time-warped TADA (bottom) utterances for 'seven' from the task003 of XRMB speaker 11, in which each utterance shows the corresponding waveform and spectrogram. Figure 4 also compares the

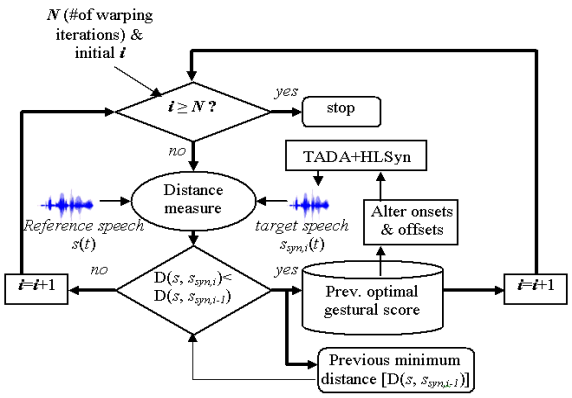

Figure 3. *Internal block diagram of the iterative LTW module*

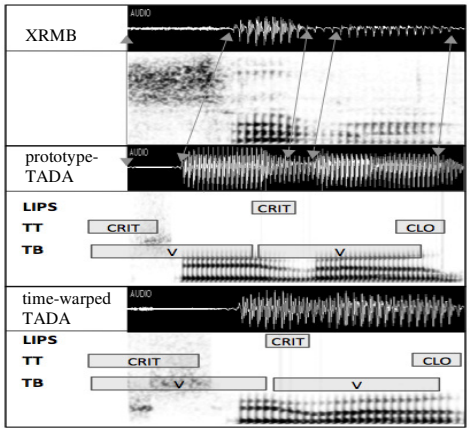

Figure 4. *Waveform and spectrogram of XRMB, prototype TADA, and time-warped TADA speech for 'seven'*

gestural scores between the prototype and time-warped utterances (with lips, tongue tip (TT), and tongue body (TB) gestures as gray blocks overlaid on the spectrogram) and shows how the gestures are modulated in time through time-warping procedure(s). The warping procedure is performed on a word-by-word basis and when all the words in the reference utterance are exhausted, the individual word-level gestural score are concatenated with each other and the overall gestural specification for the entire utterance is obtained. Once the concatenated gesture scores for the whole utterance is obtained, TADA is executed on the entire concatenated gesture to generate the corresponding TVs.

We have implemented the proposed algorithm across 27 tasks and all speakers for those tasks and we name this dataset as XRMB-Gv1, where G stands for gestures and v1 implies that this is the first version of this dataset. The results and analysis presented in the following section are based on the XRMB-Gv1 dataset.

4. Analysis and results

The annotated gestures and TVs for a part of the natural utterance from XRMB-Gv1 database are shown in Figure 5. The top two panels in Figure 5 show the waveform and the spectrogram of the utterance "eight four nine five" whereas the lower eight panels show each gesture's activation time functions and their corresponding TV trajectories, obtained from the proposed method.

We performed two tasks to evaluate our methodology. First, we compared the proposed time-warping strategy with respect to the standard DTW [15] method. To compare the effectiveness of those two warping approaches, we used an

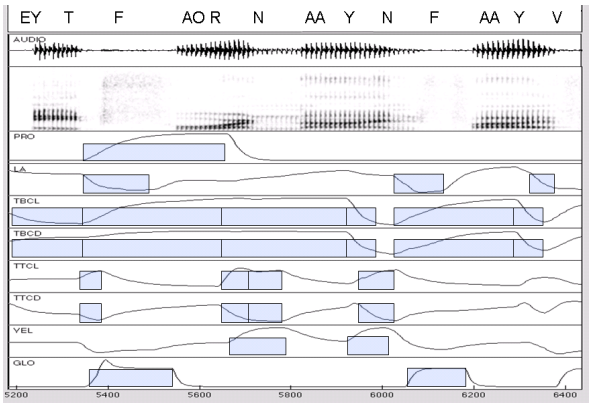

Figure 5. *Annotated gestures (gestural scores) and TVs for a snippet from an utterance from task003 in XRMB*

acoustic distance measure between the XRMB natural speech $s(t)$ and the TADA speech (i) after DTW only, $s_{syn\_DTW}(t)$ vs. (ii) our iterative warping method, $s_{syn,N}(t)$. We used three distance metrics (*a*) Log-Spectral Distance (D_{LSD}), (b) Log-Spectral Distance using the Linear Prediction spectra ($D_{LSD\text{-}LP}$) and the (c) Itakura Distance (D_{ITD}). D_{LSD} is defined as

$$D_{LSD} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{S(\omega)}{\hat{S}(\omega)}\right]^2 d\omega} \quad (3)$$

where $S(\omega)$ and $\hat{S}(\omega)$ are the spectra of the two signals to be compared. In $D_{LSD\text{-}LP}$ the spectra $S(\omega)$ and $\hat{S}(\omega)$ are replaced with their respective LP spectra that were evaluated using a 25ms window with 15ms overlap. D_{ITD} is defined as

$$D_{ITD} = \ln\left[\frac{1}{2N}\left(\sum_{\omega=-N}^{N}\frac{P(\omega)}{\hat{P}(\omega)}\right)\right] - \frac{1}{2N}\left[\sum_{\omega=-N}^{N}\ln\left(\frac{P(\omega)}{\hat{P}(\omega)}\right)\right] \quad (4)$$

where $0 \le \omega \le \pi$

Twelve different tasks (including all available speakers) were selected randomly from the XRMB-Gv1 database to obtain the distance measure between the natural and synthetic speech. Table 2 presents the average distances obtained from using DTW and the iterative time-warping approach presented here.

The next task we performed was to evaluate how effective the obtained gestures are for speech recognition. We selected 835 utterances from the XRMB-Gv1 dataset for training and 400 utterances for testing. The training set consisted of speakers 11 to 46 whereas the testing set consisted of speakers 48 to 63 (speakers 17, 22, 23 38, 47 and 50 did not exist in the XRMB database that we used, and for the remaining speakers not all the tasks were always performed). The training and the testing utterances have no overlap with themselves. Table 3 gives detailed information about the training and the testing sets. For the word recognition experiments, we converted the sequence of overlapping gestures into an instantaneous "gestural pattern vector" (GPV) as proposed in [11]. The realization of a GPV

Table 2. *Distance measures between the warped signal and the XRMB signal from using (i) DTW and (ii) proposed iterative warping strategy*

	D_{LSD}	$D_{LSD\text{-}LP}$	D_{ITD}
DTW	3.112	2.797	4.213
Iterative warping	2.281	2.003	3.834

Table 3. *Details of the train & test data of XRMB-Gv1*

	Train	Test
Number of utterances	835	400
Number of speakers	32	15
Total number of words	22647	10883
Number of unique words	269	225

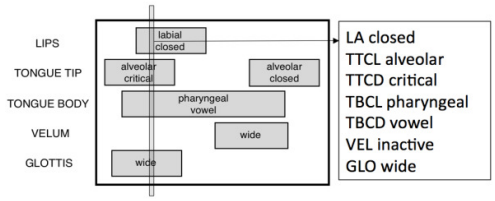

Figure 6. *Gestural score for the word "span". Constriction organs are denoted on the left and the gray boxes at the center represent corresponding gestural activation intervals. A GPV is sliced at a given time point of the gestural score.*

is demonstrated in Figure 6. From the XRMB-Gv1 train set we observed that altogether 1580 unique GPVs are possible, which indicates that theoretically 1580 × 1579 = 2494820 unique GPV bigram sequences are possible. However from the training set we observed that only 3589 unique GPV bigram sequences are observed. Hence for the training and test set we created a 3589 dimensional GPV-bigram histogram for each word. Given a word, only a few GPV bigrams will be observed, hence the word dependent GPV-bigram histogram will be largely a sparse vector. To address that we interpolated the word-dependent GPV bigrams with word-independent GPV bigrams (similar to [11]) using a ratio 5000:1, we observed this ratio to be optimal in terms of the word error rates (WER). We realized two different versions of the word recognizer using (1) Kullback-Leibler divergence (KLD) and (2) a three hidden layer neural network (NN). For the KLD based approach, word level probability mass function ($pmf_{\text{w\_train}}$, for word *w_train*, where *w_train* = 1:269, refer to Table 3) was created. For each word in the test set, the KLD between the pmfs, $pmf_{\text{w\_train}}$ and $pmf_{\text{w\_test}}$ was evaluated. The word model *w_train* that gave the least KLD was identified as the recognized word for *w_test*. KLD is defined as

$$D_{KL}\left[\, pmf_{w\_test} \,\|\, pmf_{w\_train} \,\right] = \sum_{i \in N} pmf_{w\_test,i} \log\left[\frac{pmf_{w\_test,i}}{pmf_{w\_train,i}}\right] \quad (5)$$

as $N \to \infty$ a link between the likelihood ratio (L) and KLD can be established [17] as

$$D_{KL}\left[\, pmf_{w\_test} \,\|\, pmf_{w\_train} \,\right] = -\log_2(L) \quad (6)$$

which indicates that if $pmf_{\text{w\_train}}$ and $pmf_{\text{w\_test}}$ are identical, then L = 1 and D_{KL} = 0. Hence word recognition using KLD can be formulated as

$$W = \arg\min_{w\_train} D_{KL}\left[\, pmf_{w\_test} \,\|\, pmf_{w\_train} \,\right] \quad (7)$$

For the 3-hidden layer NN approach, we used a simple feedforward network with tan-sigmoid activation function, having 400-600-400 neurons in the three hidden layers, trained with scaled-conjugate gradient. The WER obtained are shown in Table 4, which demonstrate that going from a phone to a gestural transcription preserves enough discrete structure to allow word recognition. Once we have realized a corpus with transcribed gestures we can obtain gestural score automatically [18] from a given speech in way that preserves lexical information more robustly than does a derived phone string from the audio.

Table 4. *WER obtained for XRMB-Gv1*

	KLD	NN
WER	1.41	3.72

5. Conclusion and future direction

We proposed an iterative time-warping based architecture that can annotate speech articulatory gestures potentially to any speech database that contains word and phone transcriptions. The strength of this approach is the fact that the articulatory information it generates is speaker independent, hence ideal for ASR applications. Word recognition experiments indicate that the gestures are a suitable unit-representation for speech recognition and can offer WER as low as 1.41% for a multi-speaker word recognition task. We are currently in the process of generating the gestural annotation for the whole of XRMB database and aim to implement our automated gestural annotation procedure to other speech recognition databases containing spontaneous utterances.

Acknowledgements

This research was supported by NSF Grant # IIS0703859, IIS0703048, and IIS0703782. We acknowledge the help from Dr. Jiahong Yuan for providing us the forced-aligned phones and word transcripts for Wisconsin X-Ray Microbeam database.

6. References

[1] F.J. Huang, E. Cosatto and H.P. Graf, "Triphone based unit selection for concatenative visual speech synthesis", Proc. of ICASSP, Vol.2, pp.2037-2040, Orlando, FL, 2002.

[2] C. Browman and L. Goldstein, "Articulatory Phonology: An Overview", Phonetica, 49: 155-180, 1992

[3] E. Saltzman and K. Munhall, "A Dynamical Approach to Gestural Patterning in Speech Production", Ecological Psychology, 1(4), pp.332-382, 1989.

[4] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, L. Goldstein, "Tract variables for noise robust speech recognition", under review IEEE Trans. on Audio, Speech and Language Processing.

[5] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, L. Goldstein, Noise Robustness of Tract Variables and their Application to Speech Recognition, Proc. of Interspeech, pp. 2759-2762, 2009.

[6] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein and E. Saltzman, "Articulatory Phonological Code for Word Classification", Proc. of Interspeech, pp.2763-2766, UK, 2009.

[7] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable task dynamics model in MATLAB", J. Acoust. Soc. Am, Vol. 115, No.5, 2, pp. 2430, 2004.

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

[9] B.S. Atal, "Efficient coding of LPC parameters by temporal decomposition", Proc. of ICASSP, pp.81-84, 1983.

[10] J.P. Sun, X. Jing and L. Deng, "Annotation and Use of Speech Production Corpus for Building Language-Universal Speech Recognizers", Proc. of International Symposium on Chinese Spoken Language Processing, Vol.3, pp.31-34, Beijing, 2000.

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

[12] J. Tepperman, L. Goldstein, S. Lee and S. Narayanan, "Automatically rating pronunciation through articulatory phonology", Proc. of Interspeech, pp.2771-2774, 2009

[13] J. Westbury "X-ray microbeam speech production database user's handbook", Univ. of Wisconsin, 1994.

[14] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," Proceedings of Acoustics, 2008.

[15] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. on Acoust., Speech & Signal Process., 26(1), pp.43-49, 1978.

[16] L. Rabiner, A. Rosenberg and S. Levinson, "Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition", IEEE Trans. on ASSP., 26(6), pp.575-582, 1978.

[17] T. M. Cover and J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.

[18] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, L. Goldstein, "Estimating Gestures from Speech: a Neural Network approach", under 1st revision in the J. of Acoust. Soc. of Am.

[19] E. Bresch, L. Goldstein and S. Narayanan, "An analysis-by-synthesis approach to modeling real-time MRI articulatory data using the Task Dynamic Application frame- work," 157th Meeting of the Acoustical Society of America, Portland, Oregon, May 2009.