Robust Word Recognition using articulatory trajectories and Gestures

Vikramjit Mitra[1*], *Hosung Nam*[2*], *Carol Espy-Wilson*[1], *Elliot Saltzman*[23], *Louis Goldstein*[24]

[1]Institute for Systems Research, Dept. of Electrical & Comp. Eng., University of Maryland, USA
[2]Haskins Laboratories, New Haven, USA
[3]Department of Physical Therapy and Athletic Training, Boston University, USA
[4]Department of Linguistics, University of Southern California, USA

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

Abstract*

Articulatory Phonology views speech as an ensemble of constricting events (e.g. narrowing lips, raising tongue tip), *gestures*, at distinct organs (lips, tongue tip, tongue body, velum, and glottis) along the vocal tract. This study shows that articulatory information in the form of gestures and their output trajectories (tract variable time functions or TVs) can help to improve the performance of automatic speech recognition systems. The lack of any natural speech database containing such articulatory information prompted us to use a synthetic speech dataset (obtained from Haskins Laboratories TAsk Dynamic model of speech production) that contains acoustic waveform for a given utterance and its corresponding gestures and TVs. First, we propose neural network based models to recognize the gestures and estimate the TVs from acoustic information. Second, the "synthetic-data trained" articulatory models were applied to the natural speech utterances in Aurora-2 corpus to estimate their gestures and TVs. Finally, we show that the estimated articulatory information helps to improve the noise robustness of a word recognition system when used along with the cepstral features.

Index Terms: Noise Robust Speech Recognition, Articulatory Phonology, Speech gestures, Tract Variables, TADA Model Neural Networks, Speech Inversion.

1. Introduction

Speech inversion or acoustic-to-articulatory inversion of speech has been widely researched in the last three decades. Various factors have stimulated research in this area, the most prominent being the failure of the current state-of-the-art phone-based automatic speech recognition (ASR) to account for variability in speech, for example, *coarticulation*. There are several strong arguments for considering articulatory information in ASR systems. First, it may help to account for coarticulation and reduction in a more systematic way. Second, articulatory information has been shown [2] to be more robust to speaker variations and signal distortions. Finally, it has been demonstrated [2, 3, 4] that articulatory information can significantly improve the performance of an ASR system in noisy environments.

The motivation for the current study is that an overlapping gesture-based architecture inspired by Articulatory Phonology (AP) [9, 10] can overcome the limitations of phone-based units in addressing coarticulation. AP defines speech as a constellation of articulatory gestures (known as gestural score) [9, 10], where the gestures are invariant action units that define the onset and the offset of constriction actions by various constricting organs (lips, tongue tip, tongue body, velum, and glottis) and a set of dynamic parameters (e.g., target, stiffness etc.) [9]. In AP, the intra and inter-gestural temporal overlap accounts for acoustic variations in speech originating from coarticulation, reduction, rate variations etc. Gestures are defined by relative measures of the constriction degree and location at distinct constriction organs, that is, by one of tract variables in Table 1. The gestures' dynamic realizations are the tract variable trajectories, named as TV here. Note TV does not stand for a tract variable itself but its time function output. Hence, the gestures and TVs are characterized to be less variant than x-y pellet information found in electro-magnetometer [6] and X-ray Microbeam data [7], which have been widely used for articulatory-to-acoustic speech inversion research. Using such articulatory data may be problematic as they are often contaminated with measurement noise and may suffer from inconsistency in transducer placements across speakers during the measurement procedure. Furthermore, use of absolute flesh-point information can aggravate the non-uniqueness in the acoustic-to-articulatory mapping [8]. Haskins Laboratories TAsk Dynamic Application (TADA) model is a mathematical implementation of Articulatory Phonology, which generates TVs and acoustics signals from gestural input for a given utterance [1]. Using this model, we generated a set of synthetic data completely free from measurement noise and inconsistencies.

While investigating the realizability of such articulatory gesture based ASR architecture, we observed that the estimated TVs can improve noise robustness of word recognition systems [3, 4]. In this study, we aim to investigate whether a word recognition system can be improved in various noise environments when estimated static gestural information are used (or combined with TVs). We first trained two inversion models to estimate gestures and TVs from acoustic signals using TADA-generated synthetic database. The "TV-estimator" is a feedforward (FF) artificial neural network (ANN) architecture, which estimates the TVs (shown in Table 1 and Figure 1) given acoustic features. The "Gesture-recognizer" is a cascade of (a) autoregressive (AR) ANN for gestural activation detection and (b) FF-ANN for gestural parameter detection. Second, we applied the "synthetic-data trained" TV-estimator and Gesture-recognizer to the natural speech in Aurora-2 corpus. Third, we evaluated the role of the articulatory information and its noise robustness in a Hidden Markov Model (HMM) based word recognition system. We expect various types of speech variability to be minimized by employing a novel set of articulatory information: TVs and gestures. We demonstrate that this novel set of articulatory information when estimated from the speech signal can improve the noise robustness of a

* The first two authors contributed equally to this work

natural speech word recognizer when used in conjunction with the standard cepstral features.

2. The Data

To train the TV-estimator and the Gesture-recognizer we needed a speech database that consists of groundtruth TV trajectories and gestures; unfortunately no database with such annotations exists at present. For this reason, TADA along with HLsyn [11] (a parametric quasi-articulator synthesizer developed by Sensimetrics Inc.) is used in our work to generate a database that contains synthetic speech along with their articulatory specifications. We randomly selected 960 utterances from the Aurora-2 training data. For a given utterance, the digit sequence, mean pitch and gender information were input to TADA. TADA then generated the gestures at relevant tract variables (see Table 1), output TVs, vocal tract area function, and formant information. Finally, the pitch, formant, and vocal tract constriction information were input to HLsyn which generated the synthetic waveforms. Note that the TVs and gestures generated by TADA are based upon the default speaker model defined in TADA, and hence they are speaker-independent. The sampling rate of the synthetic speech and the TV time functions are 8 kHz and 200 Hz respectively. We named this dataset as AUR-SYN, where 70% of the files were randomly selected as the training-set, 10% as the development set, and the rest as the testing-set.

Table 1. *Constriction organ, vocal tract variables*

Constriction organ	Tract Variables
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)

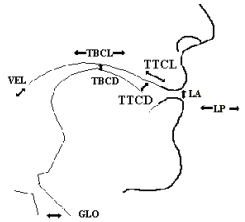

Figure 1. *Constriction organs and associated tract variables.*

The other database used in our work is the Aurora-2 [12] dataset which consist of connected digits spoken by American English talker. The TV-estimator and the Gesture-recognizer were trained with the synthetic dataset AUR-SYN and then executed on the training and testing set of Aurora-2. This was performed to observe if the estimated TVs and the recognized gestures for the natural utterances help in improving the noise robustness of the word recognition task of Aurora-2. The Speech data in Aurora-2 are sampled at 8 kHz and are in binary raw format. There are three test sections in Auroa-2, A, B and C, where test-set A and B each have four subparts representing four different noise types, hence together they have eight different noise types. Section C involves channel effects; and as our work focuses on noise-robustness only, we have ignored test-set C. Training in clean and testing in noisy scenario is used in all the experiments reported here.

For both the TV-estimator and the Gesture-recognizer, the speech signal was parameterized as Mel-frequency cepstral coefficients (MFCCs), where 13 cepstral coefficients were extracted at the rate of 200Hz with an analysis window of 10ms. The MFCCs and the target articulatory information (TVs) were z-normalized and scaled before ANN training such that their dynamic range is confined to [-0.95, +0.95]. The cepstral observations were contextualized before being sent to the TV-estimator and the Gesture-recognizer. The feature contextualization is defined by the context-window parameter $\hat{C}$, where the current frame (with feature dimension 13) is concatenated with $\hat{C}$ frames from before and after the current frame (with a frame shift of 2 or time shift of 10ms), generating a concatenated feature vector of size $(2\hat{C}+1)\times13$ and covering a temporal context window of $(2\hat{C}+1)\times10$ ms. From our prior research [13], we noticed that the optimal context window for the TV-estimator was of 170 ms and the optimal context windows for the Gesture-recognizer models [14] ranged from 90 to 190ms for gestural activation detection and 210 to 290ms for gestural parameter estimation.

3. The TV estimator

Artificial Neural Networks (ANNs) have been used in several studies [15, 16] for speech inversion. Compared to other architectures, ANNs are efficient both in terms of memory and execution speed [16]. We have used a 3-hidden layer feed-forward (FF) network, where the number of neurons for the three layers was selected to be 150-100-150 based on our observation in [13]. The dimension of the contextualized MFCCs were $(2\times8+1)\times13 = 221$, (where $\hat{C} = 8$ was found to be the optimal), requiring the ANN to have 221 inputs. The number of ANN outputs was equal to the number of TVs which is equal to 8 (as seen from Table 1). The ANN was trained with a back-propagation algorithm with scaled conjugate gradient as the optimization rule and trained for 5000 epochs. A tan-sigmoid function was used as the excitation for all of the layers. The predicted TVs were smoothed using a Kalman smoother to retain the inherent smoothness characteristics of the TVs [13].

4. The Gesture Recognizer

The TADA model uses the discrete gestural scores to generate continuous time-varying constriction distance and location trajectories which are known as the TVs. These TVs, along with other vocal tract related information are used by HLsyn to generate the synthetic speech acoustics. In this procedure, speech is synthesized from the knowledge of the articulatory configurations. TVs are output trajectories derived from gestural input for a given a speech signal. Therefore, the estimated TVs from the input speech might help to recognize the corresponding gestural scores. We formulated three approaches using the estimated TV information to recover gestural scores from speech as shown in Figure 2: (1) perform TV estimation first and use the acoustic observation (in the form of MFCCs) along with the estimated TVs to recognize the gestural scores, (2) Obtain gestural scores directly from the acoustic observation, and (3) perform TV estimation first and use only the estimated TVs for performing gestural score recognition. For all of the three approaches we adopted a 2-stage cascade model of ANNs (shown in Figure 3), where gestural activation (onset and offset) information is obtained in the first stage using a non-linear AR-ANN, and gestural parameter estimation (gestures' target and stiffness parameters) is performed in the second stage using an FF-ANN. For a given tract variable (e.g. LA, TTCD, etc), a single cascaded model was trained for each of the three approaches (shown in Figure 2), hence altogether 3 cascade models were trained for each tract variable.

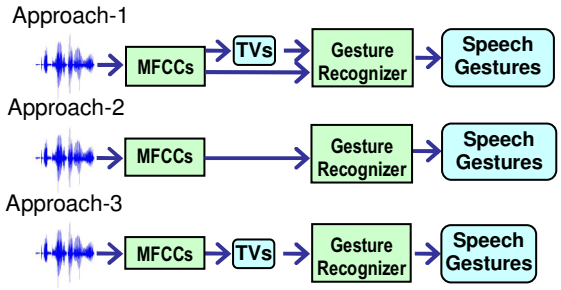

Figure 2. *The three approaches for Gesture recognition.*

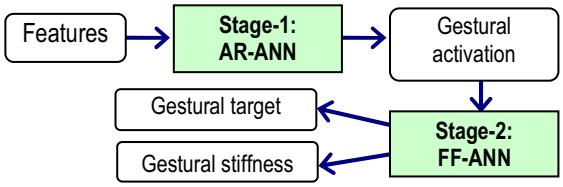

Figure 3. *The 2-stage cascaded ANN architecture for the Gesture recognizer*

The gestural activation is discrete and quasi-stationary in nature, that is, the activation can have only two states with value = 1 when active, and = 0 when inactive. Once the gesture is active or inactive it stays in that state for a given interval of time (at least 50ms and at most 300ms), which means that the activation does not toggle between the states instantaneously. The quasi-stationary nature of the activations are better captured by the feedback loop in the AR-ANN, which prevents the ANN to make instantaneous switching between the states and helps to stay in a given state once it has switched. The nonlinear AR-ANN used is a recurrent network with feedback connection from the last layer of the network, and was trained using 'dynamic back-propagation'.

The second stage uses an FF-ANN to predict gestural dynamic parameters: constriction target and gestural stiffness [10, 17], during the gestural active intervals. We considered 10 different tract variable types for the Gesture-recognizer model: LP, LA, TTCL, TTCD, TBCLC, TBCLV, TBCDC and TBCDV, VEL and GLO. Note that, since tongue body gestures are shared by velar consonants and vowels, TBCL and TBCD tract variables were split into consonant (TBCLC and TBCDC) and vowel sub-tract variables (TBCLV and TBCDV). The feature inputs to the 2-stage cascaded ANN architecture (Figure 3) in each of the three approaches are (1) estimated TVs + MFCCs, (2) MFCCs and (3) estimated TVs, respectively. The features were temporally contextualized (as specified in section 2) and the optimal context windows for each stage were found to vary for different tract variables.

5. ASR experiments

The TV-estimator and the Gesture-recognizer were trained with the synthetic speech from AUR-SYN and then used to predict the TVs and gestural scores for the natural speech of the Auroa-2 database. The estimated TVs and the recognized gestures were used along with the cepstral features MFCC and RASTA-PLP [18] to perform word recognition experiments on Aurora-2. This was done to ascertain if the articulatory information in the form of TVs and gestural scores helps to improve the noise robustness of the conventional word recognition systems. We used the HTK-based speech recognizer distributed with the Aurora-2 corpus [12] to perform the ASR experiments discussed in this paper. Test section A and B of Aurora-2 were used. The ASR experiment was based on training with clean and testing with noisy data. Note that the TV estimator uses the MFCCs obtained directly from acoustic signal, and hence has no explicit capability to deal with noisy speech.

6. Results

TV-estimator: We begin our experiments by evaluating the performance of the TV-estimator. We have used the Pearson product-moment correlation (PPMC) coefficient (equation (1)), to compare the estimated TVs with their groundtruths. The PPMC gives a measure of amplitude and dynamic similarity between the estimated and the groundtruth TVs.

$$r_{PPMC} = \frac{N\sum_{i=1}^{N} e_i t_i - \left[\sum_{i=1}^{N} e_i\right]\left[\sum_{i=1}^{N} t_i\right]}{\sqrt{N\sum_{i=1}^{N} e_i^2 - \left(\sum_{i=1}^{N} e_i\right)^2}\sqrt{N\sum_{i=1}^{N} t_i^2 - \left(\sum_{i=1}^{N} t_i\right)^2}} \quad (1)$$

where e and t represent the estimated and the groundtruth TV vectors respectively, having N data points. Table 2 shows the PPMC obtained for the 8 TV estimates and Figure 4 shows the estimated and the groundtruth TVs for utterance 'ground'.

Table 2. *PPMC for the estimated TVs*

GLO	VEL	LA	LP	TBCL	TBCD	TTCL	TTCD
0.988	0.990	0.973	0.984	0.997	0.991	0.983	0.991

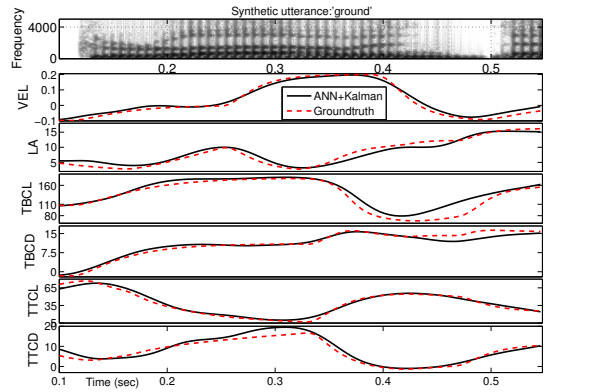

Figure 4. *Groundtruth and estimated TVs from the TV-estimator*

Gesture-recognizer: The three different Gesture-recognition approaches described in Section 4, Figure 2, was constructed for each of the 10 gestures. The network configurations (i.e., input contextual information, number of neurons) for the AR-ANN and the FF-ANN in the cascaded architecture were optimized using the development set of AUR-SYN. The gesture-recognition accuracy is obtained using equation (1)

$$\text{Rec. Acc.} = \frac{M - S}{M} \times 100 \quad (2)$$

where, M is the total number of frames and S is the number of frames having at least one of the gestural parameter (amongst gestural activation, target and stiffness) wrongly recognized. Table 3 presents the overall gesture recognition accuracy (averaged across the 10 different gestures) obtained from the three approaches. It can be observed from Table 3 that approach-1 gave the best accuracy and hence the Gesture – recognizer based on approach-1 will be used in the ASR experiments presented in this paper. Approach-1 has the added advantage of using multiple context windows and multiple streams of information (i.e., the MFCCs and the TVs). Typically a larger context window is necessary for gestural parameter estimation while a smaller context window is necessary for TV estimation and gesture activation detection. The multi-resolution and multi-stream information may be the reason behind approach-1's superior performance.

Noise-robust word recognition: The estimated TVs and the recognized gestures were concatenated with the (a) 39 dimensional MFCCs and (b) 39 dimensional RASTA-PLP for

Table 3. *Overall Gesture Recognition accuracy*

	Approach-1	Approach-2	Approach-3
Rec. Acc.	93.66	89.62	90.20

performing the clean-train word recognition task of the Aurora-2 database. The backend uses eleven whole word HMMs ('zero' to 'nine' and 'oh') and two silence/pause models 'sil' and 'sp', each with 16 states and the number of Gaussian mixtures were optimized for each input feature set. Table 4 shows the word recognition results for MFCC and RASTA-PLP with and without the TVs and the gestures. Table 4 and Figure 5 show that the estimated TVs and the recognized gestures helped to improve the noise robustness of the word recognition system. Table 4 shows that the estimated TVs and the gestures by themselves were not sufficient for word recognition, which indicate that the acoustic features (MFCC/RASTA-PLP) and the articulatory information (TVs & Gestures) are providing complementary information, hence neither of them alone is offering better result than when they are used together.

Table 4. *Overall Gesture Recognition accuracy*

	Clean	0-20dB	-5dB
MFCC	99.00	51.04	6.35
MFCC+TV	98.82	70.37	10.82
MFCC+TV+Gestures	98.56	73.49	16.36
RASTA-PLP	99.01	63.03	10.21
RASTA-PLP+TV	98.96	68.21	12.56
RASTA-PLP+TV+Gestures	98.66	75.47	19.88
TV	72.47	42.07	10.06
TV+Gestures	82.80	47.50	9.48

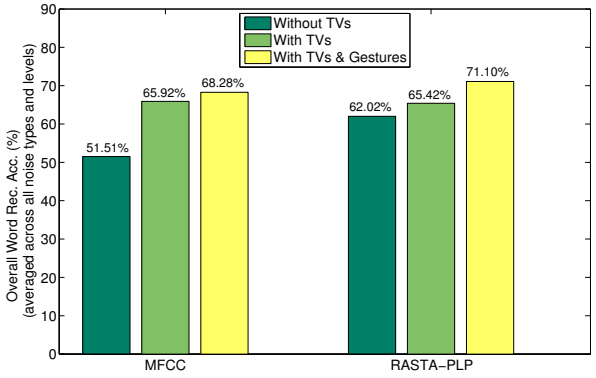

Figure 5. *Overall word recognition accuracy using MFCC and RASTA-PLP with and without the TVs and Gestures.*

7. Conclusion and Discussion

The major challenge of the TV-estimator and Gesture-recognizer was that they have been trained with clean synthetic speech and then executed on clean and noisy natural speech from different speakers, which probably introduce severe acoustic mismatch to the estimator and recognizer. They also suffer from limited training data as the AUR-SYN only consists of 960 utterances, which is roughly 11% of the entire 8440 utterances in Aurora-2 training corpus. Training the models with the whole Aurora-2 training corpus may significantly improve the accuracy of the TV-estimator and gesture-recognizer which may result in showing further improvement in the recognition accuracies. Unfortunately Aurora-2 does not come with groundtruth TV specification. Presently we are working on realizing a natural speech database with TV and gestural information [19], which in turn would help us to build a more robust TV-estimator and gesture-recognizer for natural speech.

Acknowledgements

This research was supported by NSF Grant # IIS0703859, IIS0703048, and IIS0703782

8. References

[1] 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.

[2] K. Kirchoff, "Robust Speech Recognition Using Articulatory Information", PhD Thesis, University of Bielefeld, 1999.

[3] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, L. Goldstein, "Tract variables for noise robust speech recognition", under review *IEEE* Trans. on Aud., Speech & Lang. Processing

[4] 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, Brighton, UK, 2009.

[5] J. Ryalls and S. J. Behrens, "Introduction to Speech Science: From Basic Theories to Clinical Applications", Allyn & Bacon, 2000.

[6] A. A. Wrench and H. J. William, "A multichannel articulatory database and its application for automatic speech recognition", In 5th Seminar on Speech Production: Models and Data, pp. 305–308, Bavaria, 2000.

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

[8] R.S. McGowan, "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: preliminary model tests", Speech Comm., Vol.14, Iss.1, pp. 19-48, Elsevier Science Publishers, 1994.

[9] C. Browman and L. Goldstein, "Articulatory Gestures as Phonological Units", Phonology, 6: 201-251, 1989

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

[11] H.M. Hanson, R.S. McGowan, K.N. Stevens and R.E. Beaudoin, "Development of rules for controlling the HLsyn speech synthesizer", Proc. of ICASSP, Vol.1, pp.85-88, 1999

[12] H.G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", In Proc. ISCA ITRW ASR2000, pp. 181-188, Paris, France, 2000.

[13] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, L. Goldstein, "Retrieving Tract Variables from Acoustics: a comparison of different Machine Learning strategies", To appear in the IEEE Journal of Selected Topics on Signal Processing, Special Issue on Statistical Learning Methods for Speech and Language Processing, 2010.

[14] 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.

[15] G. Papcun, J. Hochberg, T.R. Thomas, F. Laroche, J. Zachs and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data", J. of Acoust. Soc. Am, 92(2), pp. 688-700.

[16] K. Richmond, "Estimating Articulatory parameters from the Acoustic Speech Signal", PhD Thesis, Univ. of Edinburgh, 2001.

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

[18] H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Trans. Speech Aud. Proces., 2, pp.578–589, 1994.

[19] H. Nam, V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C. Espy-Wilson and M. Hasegawa-Johnson, "A procedure for estimating gestural scores from articulatory data", 159th meeting of the Acoust. Soc. of Am., Baltimore, 2010.