Robust Speech Recognition: Articulatory Information to Account for Coarticulation

Rob Bailey, Kossivi Wody Edji, Vikramjit Mitra, Carol Espy-Wilson
rbailey9@umd.edu, kedji@umd.edu, vmitra@umd.edu, espy@umd.edu

MERIT BIEN, University of Maryland, College Park, MD

*Abstract**— Current state-of-the-art phone-based speech recognition systems fail to model coarticulation properly. Tri-phone and bi-phone based approaches can only model coarticulatory effects due to the immediate neighboring phones. Phonetic studies have shown that coarticulatory effects can exist beyond the regime of di-phone or tri-phone based models and have also claimed that articulatory information can help to properly address those effects. The objective of our current study is to retrieve articulatory information from speech. Our task is twofold. First we will develop a recurrent neural network (RNN) architecture that predicts articulatory information given synthetic acoustic speech (generated from the TAsk Dynamic and Applications Model) and compare it with previously proposed systems. To evaluate such a system for natural speech we require a natural speech database containing articulatory information. With this in mind, the second goal of our research is to propose and realize a methodology to specify articulatory information for a natural speech database, in this case the X-ray microbeam corpus.

1. INTRODUCTION

Current state-of-the-art Automatic Speech Recognition (ASR) systems assume that speech is a piecewise stationary signal, and models such stationary regions as phones. Such models primarily rely upon the distinctiveness of such stationary regions while creating the phone based models. Coarticulation is a speech production effect which results in assimilation of the place of articulation of one speech sound due to that of another. As a result, the distinctiveness of the phones often gets lost so that phone based acoustic models, particularly for spontaneous speech, do not fare well. To address this short coming, di-phone or tri-phone based acoustic models are often developed, but they assume that coarticulatory effects only impact immediate phones which is not always true, as there are many instances where coarticulation extends beyond the immediate neighbors. In order for speech variability to be accurately modeled, limitations such as, clearly articulated speech or limited vocabulary for the ASR system must be imposed. Several studies [1, 2] have suggested that articulatory information can model coarticulation effectively, which can efficiently address the pitfalls of the phone-based ASR architecture. To efficiently incorporate articulatory information in an ASR system, one needs to obtain such information from the acoustic signal. This research aims to analyze the feasibility of estimating such articulatory information from the speech signal, with a goal to exploit such information in an ASR system.

Two forms of articulatory information were considered in this study, the pellet trajectories and tract variables. The pellet trajectories provide absolute articulatory motion information in a Cartesian plane, and are obtained by placing pellets (or transducers) on different articulators in the vocal tract. Unfortunately being an absolute measure, the pellet information can be inconsistent and may suffer from variability. The tract variables are measures of the various constriction locations and their degrees in the vocal tract. The tract variables are relative measures and hence should be invariant. Obtaining articulatory information from speech is commonly known as the 'speech-inversion' problem and such a problem suffers from non-uniqueness, which stems from the fact that many different articulatory configurations result in similar acoustic properties. McGowan [3] stated that the tract variables, being a relative measure, should be less prone to non-uniqueness problems than the pellet trajectories. Based on these facts we expect the tract variables will be better estimated than the pellet trajectories from the speech signal. This research uses eight tract variables and seven pellet locations (specified by x and y –axis locations in a Cartesian plane) as listed in Table 1.

* The first two authors have contributed equally to this paper and are in alphabetical order.

Table 1: Description of Tract Variables and Pellets

Tract Variables	
LA	Lip Aperture
LP	Lip Protrusion
TTCD	Tongue Tip Constriction Degree
TTCL	Tongue Tip Constriction Location
TBCD	Tongue Body Constriction Degree
TBCL	Tongue Body Constriction Location
VEL	Velum
GLO	Glottis
Pellets	
TD	Tongue Dorsum
TT	Tongue Tip
TR	Tongue Rear
TB	Tongue Body
UL	Upper Lip
LL	Lower Lip
JAW	Jaw

The initial results reported in this study are obtained from using a synthetic speech database generated by the TAsk Dynamics Applications (TADA) model [4]. The synthetic database consists of utterances along with their groundtruth articulatory information, which helped in training the speech-inversion models. Unfortunately no natural speech databases contain tract variable information, even though some of them contain pellet trajectories. To realize a speech inversion module for natural speech we are proposing a methodology to create tract variable specifications for the X-ray microbeam database (XRMB) [5], which contains natural speech utterances.

2. DATABASES

2.1 TADA

TADA is a speech production model generated at Haskins Laboratory. TADA takes either the phonetic transcription or the orthography of a word and outputs tract variables, the pellet trajectories and the gestures for that word along with acoustic parameters such as the amplitudes and bandwiths of formants. The acoustic parameters are used by HLSyn to generate the synthetic acoustic waveform. Using TADA, a synthetic word corpus of 420 utterances was generated. Eighty five percent of the data were used as training samples and the remaining were used for testing.

2.2 X-ray Microbeam

The University of Wisconsin's X-ray Microbeam Speech Production database, used in this study, contains naturally spoken utterances both as isolated sentences and short paragraphs. The speech data were recorded from 32 female speakers and 25 male speakers, where each speaker completed 118 tasks. The data comes in three forms: text data consisting of the orthographic transcripts of the spoken utterances, digitized waveforms of the recorded speech and simultaneous X-ray trajectory data of articulator movements obtained from transducers (pellets) placed on the articulators. The trajectory data are recorded for the individual articulators: Upper Lip, Lower Lip, Tongue Tip, Tongue Blade, Tongue Dorsum, Tongue Rear, and Jaw.

3. METHODOLOGY

3.1 Gestures for Natural Speech

According to gestural-phonology, gestures [6] are constriction actions along the vocal tract which are defined by dynamic parameters. The Task Dynamic model of speech production assumes that the articulatory movements are the results of the gestures, i.e. the gestures are the action units and the tract variables are their manifestations in time. According to gestural phonology, a given word can be represented as a constellation of gestures and coarticulation can be modeled as gestural overlap in time and reduction in space.

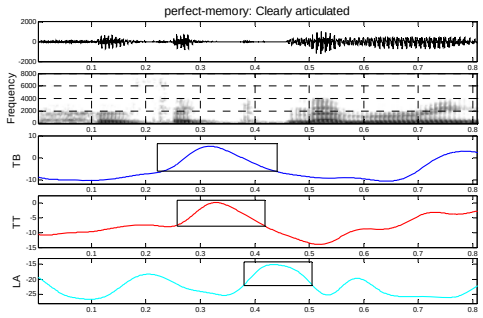
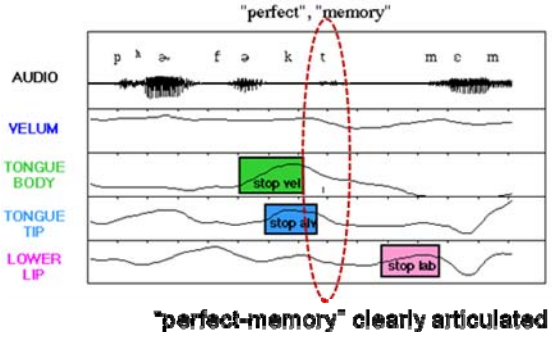
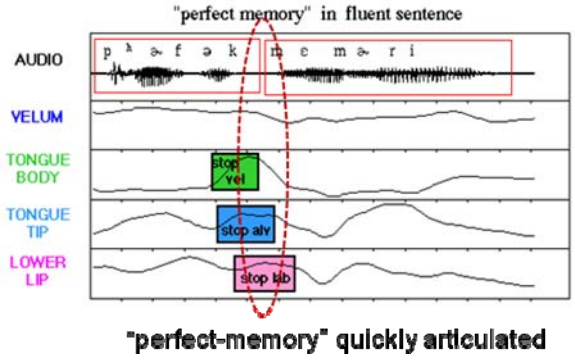


Figure 1: An example of gestures. It shows the movement or trajectories of three articulators while pronouncing the phrase “perfect memory”

Figure 2 shows two instances of the utterance ‘perfect memory’ from a male speaker; where the first one is well articulated and the second one is quickly pronounced.



(a)



(b)

Figure 2: Illustration of coarticulation- phrase "perfect memory" pronounced clearly (a), and quickly (b)

In Figure 2(a), the words 'perfect' and 'memory' are uttered with a slight pause between them, i.e. as isolated words. In Figure 2(b), 'perfect memory' is uttered more fluently with no pause between the words. A comparison of the waveforms at the end of the word 'perfect' shows that the /t/ burst of the more carefully articulated utterance in part (a) is absent from the more casually spoken utterance in part (b). This apparent "deletion" of the phone /t/ is due to cross word-boundary coarticulation in the more casual utterance. That is, the speaker starts to articulate /m/ in the word "memory" before he has finished articulating /t/ in the word "perfect". This coarticulation is evident from the articulatory information (the corresponding gestures are shown as blocks) displayed beneath the waveforms for the articulators tongue body, tongue tip and lower lip. The curves show how the vertical displacement for these articulators, which can be understood as the reverse of the constriction degrees of the relevant gestures, changes as a function of time. While the constriction degrees for the different articulators are similar for these two utterances during the /k/ and /t/ at the end of "perfect" and during the /m/ at the beginning of "memory, the timing is substantially different. For the more fluently spoken utterance, the closure gesture for the /m/ (labeled as "stop lab") overlaps with the tongue tip constriction gesture for the /t/ (labeled as "stop alv"). However, this overlap does not occur for the utterance in part (a). What is most important to note is that, although the physical waveform for the more fluent utterance does not show a /t/ burst because of the overlapping gesture for the /m/, the closure gesture for the /t/ is still made by the speaker. Thus, the large variability that can occur in the physical signal is reduced at the gestural level. This stability at the gestural level is attributed to the 'invariance property' of the gestures.

In an articulatory-gesture based ASR system, gestures will be used as the sub-word level representation of the speech signal. To realize such an ASR system we need a natural-speech database that contains articulatory gestural specifications along with its associated tract variables. For this task, we have chosen to work with the X-ray microbeam corpus. The procedure involved in specifying the articulatory gestures for XRMB utterances are as follows- (1) create a pronunciation dictionary containing words with their possible phone sequences, and (2) create a synthetic speech database given the pronunciation dictionary using TADA. TADA generated the ground truth gestural specifications and their corresponding tract variables as well. To infer the gestures for XRMB speech from the synthetic database we obtain the dynamic time warping scale by comparing the synthetic words with their corresponding natural speech counterparts. The obtained scale is then used to perform dynamic time warping of the associated gestures. The warped gestures can be interpreted as the articulatory gestures corresponding to the natural speech of the X-ray microbeam corpus. Finally, we evaluated the quality of the DTW performed on the word and the gesture.

Since we know that articulatory gestures are action units and the articulatory motions are their results, hence to estimate the gestures, we must first obtain the articulatory information using speech inversion.

3.2 Speech Inversion

We have performed speech inversion through training Artificial Neural Networks (ANNs) and optimized the network by observing the error surface obtained from varying the number of hidden layers, the number of neurons in each of those hidden layers and the contextual information of the input acoustic feature. Neural networks are composed of simple elements operating in parallel. These elements are inspired by the biological nervous systems. As in nature, the connections between elements largely determine the network function. You can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements. Neural networks have been trained to perform complex functions in various fields to solve problems in function fitting, pattern recognition, and clustering [7]. During the training of ANNs, the weights and biases tied to each neuron are updated until the specified configuration provides the least error (typically mean squared error) or satisfied some convergence requirements. Typically an ANN architecture consists of an input layer (whose size is determined by the dimensionality of

the input features), one or more hidden layers (the number of hidden layers are user specified), and an output layer whose size depends upon the dimensionality of the target set.

We have used a nonlinear autoregressive network that has a feedback loop connecting the output with the input layer. The feedback loop acts as a low pass filter, yielding smoother trajectories for both the pellets and the tract variables. It should be noted that articulatory information such as the pellet trajectories or the tract variables are inherently low pass in nature and hence the low-pass constraint of the autoregressive network helps to capture this property. Speech inversion is inherently a non-linear mapping problem. To induce non-linearity in the network, tan-sigmoid activation functions were used.
In this research the speech signal is parameterized as Mel-Frequency Cepstral Coefficients (MFCCs) and Acoustic Parameters (APs) [8,9,10] (e.g. formant information, mean Hilbert envelope, energy onsets and offsets, periodic and aperiodic energy in subbands [11] etc.). The ANN inputs are the contextualized parameter vector generated from the speech signal. The contextualized parameter vector can be created as follows: given a frame in the signal, we consider n frames before the main frame as well as n frames after the main frame. We concatenate the feature vectors derived from these frames to create the contextualized feature vector, our input. The number n is an integer between five and eleven and this integer value is what we refer to as our context value.

We implemented the process of ANN optimization through five different trials while training. Several steps were taken to optimize the network. With a single hidden layer architecture having 25 neurons and a feedback loop of unit delay, we iteratively varied the context. Once we obtained the optimal context we held that constant for the rest of the trials. We then varied the number of neurons in the first hidden layer, from twenty-five to two hundred in increments of twenty-five and the optimal number of neurons was selected based upon the obtained error. Once we obtained the optimal number of neurons in the first hidden layer we held that number constant and incremented the delay in the feedback loop from one to five. The optimal delay was found and held constant and we then added a second hidden layer and varied the number of neurons in it, in the same fashion as that for the first hidden layer. Once we have found the optimal context, delay and number of neurons in the first and second hidden layers, we then increased the number of iterations from five thousand to eight thousand and trained the network one last time.

We used several variations of neural networks in our research based on the number of outputs from the network and the type of parameterization used. We first trained fifteen individual networks; each one gave a single output, the trajectory for a specific tract variable or pellet. Next, we trained one single network which output all of the tract variables and another separate network which output all of the pellet trajectories. We used both of these methods for MFCCs and for APs. Therefore, for clear reference from now on we will refer to the results from these configurations: (a) individual networks estimating tract variables or pellet trajectories using MFCCs or APs with a single output and (b) a single network estimating tract variables or pellet trajectories using MFCCs or APs with multiple outputs.

4. RESULTS

The articulatory information must be known *a priori* to estimate gestures; therefore we first present the results from speech inversion and end with specifying gestures for natural speech.

4.1 Neural Network Training Results

Throughout the process of optimization, the accuracy of the estimated trajectory of the articulators or tract variables relative to the target trajectories improved. The performance metric used to evaluate the accuracy of the estimated articulatory trajectories (tract variables or pellets) was the Pearson product moment correlation (PPMC). PPMC can be computed using the following equation:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Where r = Pearson correlation, x_i = actual trajectory, y_i = estimated trajectory, n= number of elements in x_i

a) Training using one ANN for each TV or Pellet (individual networks with a single output)

In Table 2, we present the optimal PPMC of the different pellets and tract variables using MFCCs and APs

Table 2: ANN correlation for TV and pellet for individual networks with a single output where TV = Tract Variable, r= Pearson Correlation, Pel = Pellet.

MFCC				AP			
TV	r	Pel	r	TV	r	Pel	r
GLO	0.98	LL	0.64	GLO	0.99	LL	0.60
VEL	0.90	UL	0.41	VEL	0.73	UL	0.63
LA	0.85	JAW	0.85	LA	0.76	JAW	0.83
LP	0.52	TD	0.93	LP	0.69	TD	0.88
TTCD	0.93	TF	0.89	TTCD	0.90	TF	0.82
TTCL	0.93	TR	0.93	TTCL	0.86	TR	0.88
TBCD	0.91	TT	0.84	TBCD	0.83	TT	0.75
TBCL	0.91			TBCL	0.88		
Avg	0.87	Avg	0.78	Avg	0.83	Avg	0.77

Using both MFCCs and APs, the correlation values of the tract variables were found to be higher than the pellet trajectories. The plots of the estimated trajectories are shown in Figure 3.

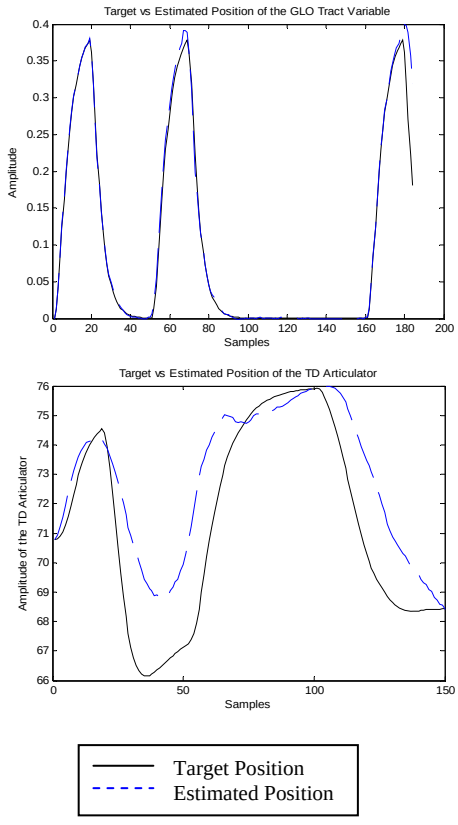


Figure 3: Estimated vs. target trajectory of the TD pellet and GLO tract variable using MFCCs.

b) Training using one ANN for all TVs or Pellets (one network with multiple outputs)

Similarly to part a), we present in Table 3 the optimal correlation of the tract variables and pellets.

Table 3: ANN correlation for TV and pellet for a network with one input and multiple outputs where TV = Tract Variable, r= Pearson Correlation (%), Pel = Pellet.

MFCC				AP			
TV	r	Pel	r	TV	r	Pel	r
GLO	0.80	LL	0.52	GLO	0.91	LL	0.39
VEL	0.44	UL	0.46	VEL	0.33	UL	0.31
LA	0.50	JAW	0.61	LA	0.56	JAW	0.40
LP	0.39	TD	0.83	LP	0.58	TD	0.62
TTCD	0.68	TF	0.83	TTCD	0.68	TF	0.62
TTCL	0.63	TR	0.83	TTCL	0.70	TR	0.62
TBCD	0.63	TT	0.66	TBCD	0.68	TT	0.53
TBCL	0.66			TBCL	0.63		
Avg	0.59	Avg	0.68	Avg	0.63	Avg	0.50

Comparing the correlations, we can state that the estimation of the tract variables is more accurate than those of the pellets when using APs but not when using MFCCs.

The approach used in b) yielded poor estimations, with a correlation average of 0.59 for tract variables compared to 0.87 for tract variables in part a). Therefore, for the rest of the project we used individual networks with a single output to estimate a given pellet trajectory or tract variable.

For some articulators and tract variables APs yielded better results, while MFCCs gave more accurate results for others. We will continue to use both parameters until further experiments confirm the ideal parameter.

The different experiments using artificial neural networks have proven that tract variables gave more accurate estimations than articulators, as shown in Figure 3. We will focus the rest of the speech inversion task in obtaining articulatory information from tract variables.

4.2 Dynamic Time Warping Results

We used DTW to perform a non-linear mapping of synthetic speech signals to natural speech signals. Figure 4 shows different signals of the word “problem”. Wave 1 represents the synthetic signal, wave 2 is the natural signal and wave 3 is the warped signal.

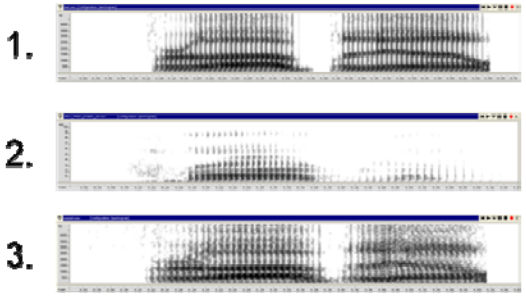


Figure 4: DTW performed on the word “problem” 1- synthetic signal, 2- natural signal, 3-warped signal

Initially, signal 1 and 2 differed from each other in terms of their corresponding phone durations. After the DTW of signal 1 was performed, we can see that both the natural signal and the warped signal line up properly (in terms of frication and duration). Once we warped the word itself, we also warped the gestures of synthetic speech in order to obtain the gestures for natural speech.

5. CONCLUSION

Through our research we were able to successfully train different autoregressive neural network architectures for the prediction of tract variables and pellet trajectories. The results show that individual ANNs with a single output provided a more accurate estimate than the one with multiple outputs. Furthermore, we observed that for both the acoustic features, MFCCs and APs, the tract variables were estimated more accurately, in most of the cases, than the pellet trajectories. The most accurate network configuration was the one in which we had different networks estimating each tract variable individually. We were not able to make such a claim regarding the performance of the MFCCs against the APs; some networks were more accurate with MFCCs as the input and others with APs.

We were able to use dynamic time warping to warp synthetic speech to the natural speech of the same word and phonetic transcription. We were also able to use that warping in order to obtain the gestures for the natural speech from the gestures of the synthetic speech.

This research is a preliminary step in designing an ASR system which uses gestures to account for coarticulation. Future work would be directed toward realizing better and more efficient strategies to derive gestural information for natural speech. Once the gestural specifications and articulatory information are generated in full for a natural speech corpus, the proposed speech inversion models should be implemented for the natural speech corpus to see the feasibility of the speech inversion task in a real-world scenario.

6. ACKNOWLEDGMENT

This research was supported by National Science Foundation CISE award #0755224

7. REFERENCES

[1] J. Frankel and S. King, "ASR - Articulatory Speech Recognition", In proceedings of Eurospeech, pp. 599-602, Aalborg, Denmark, September 2001.

[2] J. Frankel and S. King, "A Hybrid ANN/DBN Approach to Articulatory Feature Recognition", in Proceedings of Eurospeech, Interspeech-2005, pp.3045-3048, Lisbon, Portugal, 2005.

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

[4] 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, pp. 2430, 2004.

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

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

[7] H. Demuth, M. Beale and M. Hagan "Neural Network Toolbox 6 User's Guide" The MathWorks, Natick, MA, 2008.

[8] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks", *PhD thesis*, University of Maryland College Park, December 2004.

[9] K. Stevens, S. Manuel and M. Matthies, "Revisiting place of articulation measures for stop consonants: Implications for models of consonant production". Proceedings of *International Congress of Phonetic Science*, Vol-2, pp. 1117-1120, 1999.

[10] S. Chen and A. Alwan, "Place of articulation cues for voiced and voiceless plosives and fricatives in syllable-initial position", Proceedings of *ICSLP*, vol.4, 113-116, 2000.

[11] O. Deshmukh, C. Espy-Wilson, A. Salomon and J. Singh, "Use of Temporal Information: Detection of the Periodicity and Aperiodicity Profile of Speech", *IEEE Trans. on Speech and Audio Process*ing, Vol. 13(5), pp. 776-786, 2005.