Analysis of coarticulated speech using estimated articulatory trajectories

Ganesh Sivaraman [1], Vikramjit Mitra[2], Mark Tiede[3],
Elliot Saltzman[4], Louis Goldstein[5], Carol Espy-Wilson[1]

[1]University of Maryland College Park, MD, [2]SRI International, Menlo Park, CA,
[3]Haskins Laboratories, New Haven, CT, [4]Boston University, Boston, MA,
[5]University of Southern California, Los Angeles, CA

`{ganesa90, espy}@umd.edu, vmitra@speech.sri.com, tiede@haskins.yale.edu,`
`esaltz@bu.edu, louisgol@usc.edu`

Abstract

Speech acoustic patterns vary significantly as a result of coarticulation and lenition processes that are shaped by segmental context or by performance factors such as production rate and degree of casualness. The resultant acoustic variability continues to offer serious challenges for the development of automatic speech recognition (ASR) systems. Articulatory phonology provides a formalism to understand coarticulation through spatiotemporal changes in the patterns of underlying gestures. This paper studies the coarticulation occurring in certain fast spoken utterances using articulatory constriction tract-variables (TVs) estimated from acoustic features. The TV estimators are trained on the University of Wisconsin X-ray Microbeam (XRMB) database. The utterances analyzed are from a different corpus containing simultaneous acoustic and Electromagnetic Articulograph (EMA) data. Plots of the estimated TVs show that the estimation procedure successfully detected the articulatory constrictions even in the case of highly coarticulated utterances that a state-of-the-art phone recognition system failed to detect. These results highlight the potential of TV trajectory estimation methods for improving the performance of phone recognition systems, particularly when sounds are reduced or deleted.

Index Terms: Coarticulation, Speech production, Speech inversion, articulatory phonology

1. Introduction

Conversational speech exhibits significant variability due to speaking rate, accents, cognitive load, etc. Coarticulation and reduction are common phenomena that occur in fast rate speech, especially affecting the acoustic properties of the speech signal that relate to manner and place of articulation. The resulting acoustic variability continues to offer serious challenges for the development of automatic speech recognition (ASR) systems that can perform well with minimal constraints. For example, conventional ASR systems attempt to account for coarticulatory effects through tri- or quin-phone and cross-word models; however it is inherently difficult to quantify a fixed scope for coarticulatory effects. Articulatory Phonology (AP) [1] provides a unified framework for understanding how spatiotemporal changes in the pattern of underlying speech gestures can lead to corresponding changes in the extent of inter-gestural temporal overlap and the degree of gestural spatial reduction. In turn, these changes in overlap and reduction create acoustic consequences that are typically reported as assimilations, insertions, deletions and substitutions. The Task Dynamics and Applications (TADA) model of speech production represents speech production actions as a set of vocal-tract constriction degree and position variables (TVs) [2]. It has been shown [3] that the performance of acoustic-to-articulatory speech inversion systems can be improved when training procedures incorporate synthetic acoustic and articulatory (TV) data generated by the TADA model. Such procedures have been shown to improve the noise robustness of ASR systems [4]. However, synthetic speech and articulatory trajectories do not display all the variability observed in natural speech, and, hence, it is essential to incorporate acoustic and articulatory data from actual speakers when training speech inversion systems to improve their generalizability. In this paper, all speech inversion systems estimate articulatory data as TVs; thus, we will often refer to them as TV estimators.

Fast rate speech leads to significant coarticulation and reduction phenomena. For example, in "perfect memory", the '/t/' often appears to be deleted acoustically due to the overlap of the lip closure for '/m/' with the tongue tip constriction for '/t/'; examination of the TV trajectories, however, shows that the underlying gestures persist. To obtain data to investigate such contexts we recorded speech at normal and fast rates concurrently with Electromagnetic Articulograph (EMA) data, using the IEEE sentences [5] as the corpus for this task. We will refer to this dataset as the EMA-IEEE database. A complete description of the recordings of the EMA-IEEE sentences is given in section 2.

Recording of articulatory data is an expensive and time consuming process. Since it is not feasible to record data from a large number of subjects, we augmented recorded data with models for estimating articulatory trajectories from acoustics. In this paper, we trained artificial neural networks (ANNs) for acoustic-to-articulatory speech inversion using speech and articulatory data obtained from the U.W. X-ray Microbeam (XRMB) database [6]. A description of these speech inversion systems is given in section 3. The trained speech inversion systems were used to estimate TVs for specific fast and normal rate utterances from the EMA-IEEE database. TVs were estimated from the sensor positions recorded using EMA, and were compared to the actual TVs obtained from the Electromagnetic Articulograph (EMA) recordings of the same IEEE sentences. A speech inversion system was also trained on the articulatory (EMA) data recorded for this experiment.

This paper compares the ability of various speech inversion systems to detect an utterance's underlying gestures given the significant coarticulation effects of fast speech.

Section 4 outlines the experimental procedures. The analysis and discussion of the selected fast utterances is presented in section 5.

2. The EMA-IEEE dataset

2.1. Dataset description

A female native speaker of American English in her mid-twenties with no self-reported speech or hearing deficits produced the 720 IEEE sentences at 'normal' and 'fast' production rates, where normal was her preferred rate (approximately 2.9 syllables/sec), and fast was produced approximately 20% more quickly. A WAVE EMA system (Northern Digital) was used to observe the trajectories of sensors placed midsagittally on the speaker's tongue (dorsum, blade, and 1 cm posterior from apex), jaw (lower incisors), lips (upper and lower vermillion border, and left mouth corner), together with reference sensors placed on the upper incisors, nose, and mastoid processes used to correct for head movement. The movement data were sampled at 100 Hz together with synchronized audio at 22050 Hz. In post-processing, movement data were aligned to the speaker's occlusal plane and low-pass filtered at 20 Hz, providing the anterior/posterior, inferior/superior, and lateral positions of each sensor relative to an origin centered on the upper incisor reference.

2.2. Conversion of EMA data to TVs

The sensors described above along with the palate trace of the female speaker were used to estimate constriction degree (TTCD, TBCD) TVs from the tongue tip (TT) and tongue body (TB) EMA sensor positions by computing the minimum distance between the pellets and the palate along the midsaggital plane. Lip Aperture (LA) was computed as the distance between the Upper Lip (UL) and Lower Lip (LL) sensors. The TBCD, TTCD and LA TVs were computed by the following formulae:

$$LA = (UL_x - LL_x)^2 + (UL_y - LL_y)^2 + (UL_z - LL_z)^2 \quad (1)$$

$$TBCD = Min\{Distance(TB, palate)\} \quad (2)$$

$$TTCD = Min\{Distance(TT, palate)\} \quad (3)$$

Figure 1 shows the three constriction degree TVs and the EMA sensor positions superimposed on the anterior vocal tract.

3. Speech inversion systems

ANNs were used to estimate TV trajectories [3] from the speech signal. An ANN can have M inputs and N outputs; hence, a nonlinear complex mapping of M vectors into N different functions can be achieved. In such an architecture, the same hidden layers are shared by all N outputs, giving the ANN the implicit capability to exploit any correlation that the N outputs may have amongst themselves. The feed-forward ANN used in our study to estimate the TVs from speech were trained with back propagation using a scaled conjugate gradient algorithm.
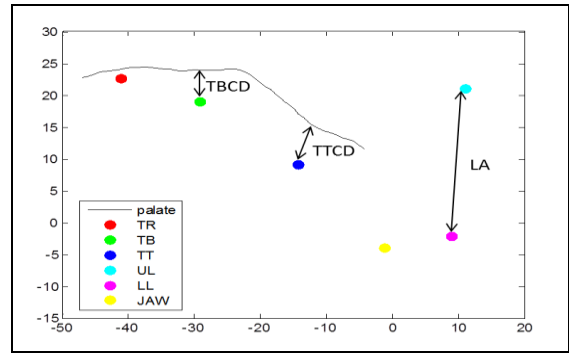

Figure 1: *EMA sensors and associated TVs*

3.1. Articulatory datasets

Speech Inversion (SI) systems were trained using two different sets of acoustic and TV data (see Table 1).

Table 1: *Description of different articulatory datasets used for training speech inversion systems*

Dataset	Description
XRMB natural speech	Complete U.W. X-ray microbeam database with pellet trajectories converted to TVs using the method outlined in [7]
EMA-IEEE converted to TVs	EMA articulatory data described in Section 2.1 converted to TVs using the method outlined in Section 2.2

3.2. Data preparation and feature extraction

The XRMB database [6] consists of continuous speech data along with horizontal and vertical displacements of 8 pellets placed on the speaker's lips, tongue and jaw. The X-Y pellet displacement measures were converted into a set of 6 constriction degree and location TV trajectories using a geometric transformation as outlined in [7]: Lip Aperture and Protrusion (LA, LP), Tongue Body Constriction Degree and Location (TBCD, TBCL), and Tongue Tip Constriction Degree and Location (TTCD, TTCL). The transformed XRMB database consisted of 1720 sentences across 46 different speakers, constituting of a corpus of 'groundtruth' TVs.

The speech signal, downsampled to 8 kHz, was parameterized as MFCCs where 13 cepstral coefficients were extracted using a Hamming analysis window of 20ms with a frame rate of 10ms. The TVs and MFCCs were mean and variance normalized to have zero mean and a variance of 0.25. The mean and variance normalization was performed separately for every speaker in the database. This ensured some normalization of inter-speaker variations in measurements of acoustics and articulations. The MFCCs were then contextualized by concatenating every other feature frame within a 160ms window on either side of each frame.

3.3. ANN Training

For the ANN-based TV estimator, the input dimension was 221 (= 13 MFCCs x 17 frames) and the output dimension was 6 (= 6 TVs). Eighty percent of the data was used for training, and 10% each was used for cross validation and testing. A 2 hidden layer neural network was trained in a greedy layer-wise manner. Networks with different numbers of hidden-layer neurons (100 to 500) were trained, and among them the best performing network was chosen for training the 2nd hidden layer. Network size was not increased to include further hidden layers as the performance improvement of the second

over the single hidden layer network was marginal. The performance of the TV estimator was measured by computing the Pearson Product Moment Correlations (PPMC) of the estimated TVs with the groundtruth TVs on the test set.

3.4. TV estimator training results

Four different TV estimators were trained using the two datasets described in section 3.1. A TV estimator was trained on the complete XRMB database. This estimator is referred to as X_NORM. To normalize gender specific acoustic variations, the XRMB database was divided into male and female speaker utterances and a TV estimator was trained on each of these subsets. The systems trained on these gender specific subsets are referred to as XF_NORM and XM_NORM. Another TV estimator was trained using the EMA-IEEE dataset. This system estimates only 3 TVs (LA, TBCD, and TTCD) as the other TVs were not computed from EMA trajectories. We refer to this estimator as E_IEEE. Table 2 summarizes these TV estimators.

Table 2: *Summary of different speech inversion (TV estimator) systems*

TV estimator name	Training dataset
X_NORM	XRMB utterances converted to TVs
XF_NORM	Female speakers' utterances from XRMB database converted to TVs
XM_NORM	Male speakers' utterances from XRMB database converted to TVs
E_IEEE	Single female speaker EMA data converted to TVs

The trained TV estimators were tested on 10% of their respective datasets where the sentences were chosen randomly. The performance of the TV estimator was measured by the Pearson Product Moment Correlation (PPMC) between the estimated and ground-truth TVs using the test set. The results for the different TV estimators are given in Table 3.

Table 3: *PPMC results of trained TV estimators on their respective test data sets. (NA: TVs were not estimated)*

TV estimator name	LA	TBCD	TTCD	LP	TBCL	TTCL
X_NORM	0.66	0.59	0.76	0.56	0.78	0.65
XF_NORM	0.72	0.66	0.79	0.62	0.82	0.66
XM_NORM	0.68	0.64	0.78	0.57	0.83	0.72
E_IEEE	0.64	0.80	0.72	NA	NA	NA

4. Experimental procedure

Effects of coarticulation and reduction can be expressed in many forms in fast rate speech including deletion, assimilation and substitution. In this paper, we selected two utterances from the EMA-IEEE dataset and one utterance from an earlier study [8] illustrating coarticulation effects. Both fast rate and normal rate utterances of these selected sentences were analyzed. Articulatory data was converted to TV representation using the

same method described in section 2.2. The following are the three sentences chosen for analysis.

1. The empty **flask stood** on the tin tray.
2. The beam dropped down on the **workman's head**.
3. She had a **perfect memory** for details. (from [8])

The words in bold contain the clusters of interest. None of the above utterances were included in any of the TV estimators trained. Each of these utterances was analyzed using the TV estimators described in section 3. We analyzed only the LA, TBCD, and TTCD TVs.

5. Results and analysis

The average correlations of the estimated TVs with actual TVs for the three selected utterances are shown in Table 4.

Table 4: *Correlations (PPMC) of estimated TVs from different TV estimators for the selected sentences (n = normal rate, f = fast rate)*

TV estimator name	flask stood		workman's head		perfect memory	
	n	f	n	f	n	f
X_NORM	0.56	0.59	0.61	0.75	0.40	0.51
XF_NORM	0.56	0.59	0.55	0.72	0.28	0.55
XM_NORM	0.56	0.59	0.59	0.63	0.44	0.58
E_IEEE	0.86	0.82	0.75	0.79	0.18	0.44

From Table 4, we can see that the E_IEEE system has the highest correlations for sentences 1 and 2 since those utterances were produced by the same speaker (note that these utterances were not included in the training of this system). Hence, we plotted the estimated TVs from E_IEEE_ system for analysis of sentences 1 and 2.

5.1. Analysis of "flask stood"

Figure 2 shows spectrograms and the TVs for the normal-rate and fast-rate productions of sentence 1. In the case of the normal-rate production, the consonant cluster /sk/ at the end of "flask" and the /st/ at the beginning of "stood" are clearly seen in the acoustics and both the actual and estimated TVs show constrictions in the right regions.
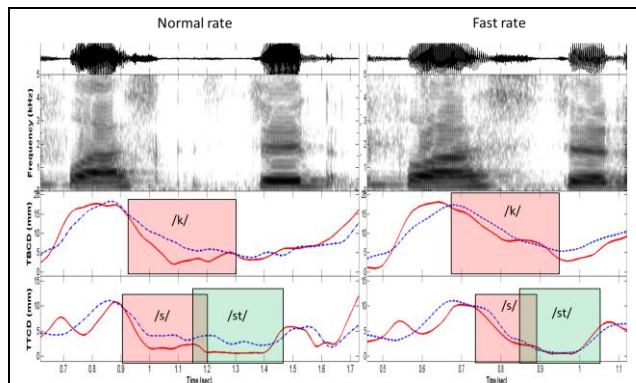

Figure 2: *Actual (red solid line) and estimated (blue dash dot line) TVs for "flask stood"*

However, in the fast-rate production of this utterance, the acoustics suggest that the /k/ in "flask" was not produced. Instead, it appears as if the /s/ in "flask" and the /s/ in stood are combined (the duration of this /s/ is about 30ms longer than

the ones in the normal-rate production) and this /s/ is then followed by a the /t/ in "stood". This appears to be a case where the fast-rate production resulted in no gesture being made for the /k/. Although there is lowering of the TBCD gesture during the /t/, this lowering appears to be due to the /t/ closure and can be seen in situations where a /s/ or /t/ is produced without an adjacent velar consonant. It is possible that this apparent deletion of the /k/ gesture is due to the complexity of these cluster sequences, which include four consecutive consonants.

5.2. Analysis of "workman"

Figure 3 shows spectrograms and the TVs for the normal-rate and fast-rate productions of sentence 2. The actual and estimated TVs are strongly correlated across the utterance. In particular, both show the /k/ constriction when it is produced as a stop in the normal-rate production and as a fricative in the fast-rate production. Note that the /k/ gesture in the fast-rate production of the utterance is weaker than it is in the normal-rate production of the same. This not surprising given the estimated TVs are derived from the acoustics. Finally, note that both sets of TVs show the closure of the lips for the /m/.
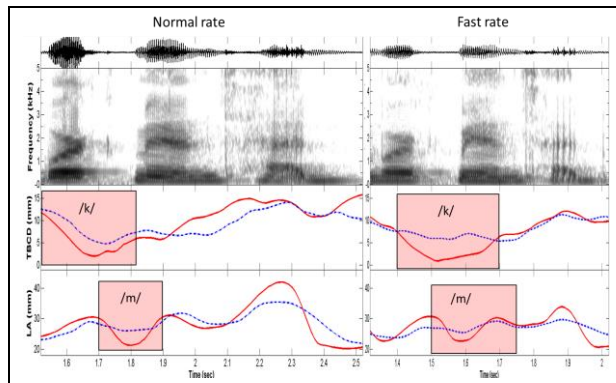

Figure 3: *Actual (red) and estimated (blue) TVs for "workman's head"*

5.3. Analysis of "perfect memory"

From Table 4, it can be seen that none of the TV estimators provide a reliable estimate of the 3 TVs that we are interested in analyzing. As a result, we trained speaker dependent TV estimators on all 46 speakers of the XRMB database. We then estimated TVs using each speaker's TV estimator and then selected the system that best correlated with the actual TVs. We observed that the TV estimator trained on speakers JW29 provided best correlations for normal rate and that trained on JW28 provided best correlations for fast rate. We used the estimated TVs from these models to analyze the "perfect memory" utterance.

Figure 4 shows spectrograms and the actual and estimated TVs for sentence 3. As can be seen in the normal-rate production, the acoustics show a release burst for /k/ but not /t/, followed by a period of silence and then the /m/ murmur at the beginning of "memory". Both sets of TVs show a tongue-body gesture for the /k/ that overlaps with the tongue-tip gesture for the /t/ and the lip gesture for the /m/. In contrast, there is no silence between the last vowel in "perfect" and the first vowel in "memory". Instead, this region appears acoustically as one sonorant consonant, i.e., the /m/. However, the articulatory data tell a different story. As in the normal

rate speech, we see gestures for the /k/ and /t/, but with considerably more overlap between the gestures. In particular, the /m/ gesture is fully overlapped with that of the other consonants. Thus, this fast-rate production of "perfect memory" contains what we refer to as "hidden gestures" for the /k/ and the /t/. Note that both of these gestures are apparent in the estimated TVs, although the closure for the lip gesture is weaker than the actual gesture.
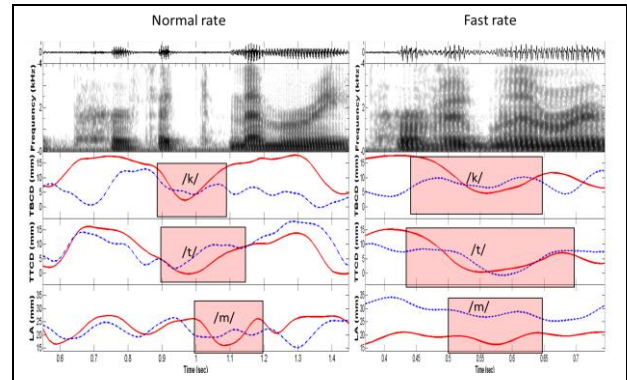

Figure 4: *Actual (red) and estimated (blue) TVs for "perfect memory"*

6. Conclusions

The results show that the speech inversion systems perform reasonably well on unseen data containing challenging coarticulatory phenomena. Working with naturally-spoken data can result in speech inversion systems that produce TVs that closely match TVs computed directly from articulatory data. However, the variability in the training data needs to be properly normalized or restricted. Thus, a future goal of this work is to develop methodologies for coping with variability and choosing which of several different speech inversion systems will work for any given speaker, especially if that speaker's data has not been used as part of the training data.

7. Acknowledgements

8. References

[1] C. P. Browman and L. Goldstein, "Articulatory Phonology : An Overview *," *Phonetica*, vol. 49, pp. 155–180, 1992.

[2] E. L. Saltzman and K. G. Munhall, "A Dynamical Approach to Gestural Patterning in Speech Production," *Ecol. Psychol.*, vol. 1, no. 4, pp. 333–382, Dec. 1989.

[3] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Retrieving Tract Variables From Acoustics: A Comparison of Different Machine Learning Strategies.," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 6, pp. 1027–1045, Sep. 2010.

[4] V. Mitra, "Articulatory Information For Robust Speech Recognition." Ph.D. dissertation, University of Maryland, College Park, 2010.

[5] E. Rothauser, W. Chapman, and N. Guttman, "IEEE Recommnded Pratice for Speech Quality Measurements,"

IEEE Trans. Audio Electroacoust., vol. 17, no. 3, pp. 225–246, Sep. 1969.

[6] J. R. Westbury, "Microbeam Speech Production Database User"'s Handbook," *IEEE Pers. Commun. - IEEE Pers. Commun.*, 1994.

[7] H. Nam, V. Mitra, M. Tiede, M. Hasegawa-Johnson, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "A procedure for estimating gestural scores from speech acoustics.," *J. Acoust. Soc. Am.*, vol. 132, no. 6, pp. 3980–9, Dec. 2012.

[8] M. K. Tiede, J. Perkell, M. Zandipour, and M. Matthies, "Gestural timing effects in the '"perfect memory"' sequence observed under three rates by electromagnetometry," *J. Acoust. Soc. Am.*, vol. 110, no. 5, p. 2657, Nov. 2001.