Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion

Prasanta Kumar Ghosh[a] **and Shrikanth Narayanan**
Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089
prasantg@usc.edu, shri@sipi.usc.edu

Abstract: An automatic speech recognition approach is presented which uses articulatory features estimated by a subject-independent acoustic-to-articulatory inversion. The inversion allows estimation of articulatory features from any talker's speech acoustics using only an exemplary subject's articulatory-to-acoustic map. Results are reported on a broad class phonetic classification experiment on speech from English talkers using data from three distinct English talkers as exemplars for inversion. Results indicate that the inclusion of the articulatory information improves classification accuracy but the improvement is more significant when the speaking style of the exemplar and the talker are matched compared to when they are mismatched.

PACS number(s): 43.72.Ar [DO]
Date Received: June 28, 2011 **Date Accepted:** August 17, 2011

1. Introduction

The role of articulatory features in automatic speech recognition has been investigated for several decades. One straightforward approach to improve recognition using articulatory features would be to access direct articulatory measurements from the talker and use them in addition to the acoustic speech features. For example, Frankel *et al.*[1] showed improvement in speech recognition accuracy by combining acoustic and articulatory features from a talker. However, it is not practical to assume the availability of direct articulatory measurements from a talker in real-world speech recognition scenarios. To address this challenge, a number of techniques have been proposed[2–4] where instead of relying on features from direct articulatory measurements, abstracted articulatory knowledge is incorporated in designing models [e.g., dynamic Bayesian network (DBN), hidden Markov model (HMM)] which can be gainfully used for automatic speech recognition. A summary of such techniques can be found in McDermott and Nakamura (2006).[5] Multi-steam architectures[6] have been also proposed as an alternative approach where linguistically derived articulatory (or more generally, phonetic) features are estimated from the acoustic speech signal, typically using artificial neural networks (ANN), and then used to either replace or augment acoustic observations in an existing HMM based speech recognition system.

In the context of articulatory data-driven approaches for speech recognition, acoustic-to- articulatory inversion offers a promising venue.[7–9] The goal of acoustic-to-articulatory inversion is to estimate the vocal tract shape or articulatory trajectories from a talker's speech; the estimated articulatory information can in turn be used for improving speech recognition. However, estimating articulatory trajectories for an arbitrary talker is quite challenging without having access to parallel articulatory-acoustic training data from that talker. This is because the shape and size of the vocal tract and articulators vary across subjects and so do their speaking styles. The requirement of

[a]Author to whom correspondence should be addressed.

talker-specific training data for inversion in fact has been a major impediment in developing automatic speech recognizers that can exploit such estimated articulatory features. Recently we proposed a subject-independent approach to inversion,[10] where parallel articulatory-acoustic training data from one exemplary subject (we refer here as "exemplar") can be used to estimate articulatory features from any arbitrary talker's speech. It was shown that the resulting estimated trajectories are significantly correlated to the measured articulatory trajectories from the talker. Thus, the subject-independent inversion offers us a potential way to develop an articulatory-data based approach for speech recognition. It should be noted that when the talker and exemplar are different, acoustic adaption techniques[11] can be used to normalize the talker's and exemplar's acoustic differences before performing acoustic-to-articulatory inversion. However, adaptation may not be a feasible option when a single utterance from the talker is available for recognition because one utterance may not provide sufficient acoustic data for adapting exemplar's acoustic model.

The goal in this paper is to experimentally study the effectiveness of using articulatory features estimated through subject-independent inversion for speech recognition. The experiments are performed using parallel acoustic and articulatory data from three native speakers of English from two distinct databases. Automatic speech recognition experiments using both acoustic-only speech features and joint acoustic-articulatory features are performed for each subject (talker) separately. To experimentally explore the effect of using estimates derived from different articulatory-acoustic maps (i.e., exemplars), we cross-test each exemplar-based model against the data of the others. Thus for each subject in our study, we have three different estimates of the articulatory features (using two other subjects and the talker itself as exemplars) as well as the original articulatory features—overall, four different versions of the articulatory features for each subject. We investigate the nature of acoustic-articulatory recognition accuracy compared to acoustic-only recognition accuracy for the different versions of the articulatory features. The availability of direct articulatory data allows us to investigate the extent and nature of the recognition benefit we can obtain when we replace the original articulatory features by the estimated ones. We next describe the articulatory datasets used in this work.

2. Datasets and features

The present study uses articulatory-acoustic data drawn from two different sources. The first one is from the multichannel articulatory (MOCHA) database[12] that contains electromagnetic articulography (EMA) data for 460 utterances (~20 min) read by a male and a female talker of British English. We refer to these subjects as EN_RD_MALE and EN_RD_FEMALE, respectively. The EMA data consist of trajectories of sensors placed in the midsagittal plane of the subject on upper lip (UL), lower lip (LL), jaw (LI), tongue tip (TT), tongue body (TB), tongue dorsum (TD), and velum (V).

The second source of parallel articulatory-acoustic data comes from the EMA data collected at the University of Southern California (USC) from a male talker of American English (EN_SP_MALE) as a part of the Multi-University Research Initiative (MURI) project.[13] In contrast to the read speech in the MOCHA database, the articulatory data in the MURI database were collected when the subject was engaged in a spontaneous conversation (~50 min) with an interlocutor. Unlike MOCHA, the second corpus has articulatory position data only for UL, LL, LI, TT, TB, and TD. The articulatory data from the MURI corpus are preprocessed, in a manner similar to that used for the MOCHA database, to obtain a frame rate of 100 Hz.

To specify articulatory features, we have used the tract variable (TV) definition[14] motivated by the basic role of constriction formation during speech production in articulatory phonology.[15] The data from the three subjects we have considered in this study do not correspond to identical set of articulators; thus, for consistency, we have chosen five TV features for each subject, namely, lip aperture (LA), lip protrusion (PRO), jaw opening (JAW_OPEN), tongue tip constriction degree (TTCD), and tongue

body constriction degree (TBCD). These TV features are illustrated in Fig. 1 and are computed from the raw position values of the sensors using the definitions given by Ghosh *et al.*[10] We use 13-dimensional mel frequency cepstral co-efficients (MFCCs) as speech acoustic features at a frame rate (100 Hz) identical to the rate of the articulatory features.

The implementation of subject-independent inversion[10] requires a generic acoustic model, the design of which requires a large acoustic speech corpus. For this purpose, we have considered the speech data from the TIMIT[16] corpus. Because TIMIT is a phonetically balanced database of English and our experiments are limited to English talkers, we assume the TIMIT training corpus adequately represents all variabilities in acoustic space required for subject-independent inversion.

3. Subject-independent inversion

In subject-independent acoustic-to-articulatory inversion,[10] the articulatory-to-acoustic map of an "exemplar" is used to estimate the articulatory trajectory corresponding to any arbitrary talker's speech. The inversion scheme itself is based on the generalized smoothness criterion based approach recently proposed.[17] Because the acoustics of the "exemplar" and the talker can be, in general, different, the basic idea of enabling subject-independent inversion[10] is to normalize this inter-subject acoustic variability by computing the likelihood of the acoustic features for both the exemplar and the target talker using a general acoustic model and predict the articulatory position values based on the closeness between likelihood scores. Because the articulatory configuration of the "exemplar" is in general different from that of the talker, it was shown in Ref. 10 that the range and values of the estimated articulatory trajectories correspond to those of the exemplar's articulatory trajectories as if the exemplar spoke the target utterance spoken by the talker.

To examine the correlation values between the original (x) and estimated TV features ($\hat{x}$), we report the average correlation coefficients (ρ) in Table 1 computed over all utterances calculated by considering in turn each of the three subjects in our data set as an exemplar and the others two as the talker. For each exemplar and talker combination, we also performed a linear regression analysis $\hat{x} = ax + b + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2)$ and a hypothesis test on the slope a (H_0: $a = 0$, H_a: $a \neq 0$) for each TV feature. We found that the estimated feature values have a significant correlation (P value $= 10^{-5}$) with the original ones (i.e., there is sufficient evidence to reject the null hypothesis H_0). To investigate

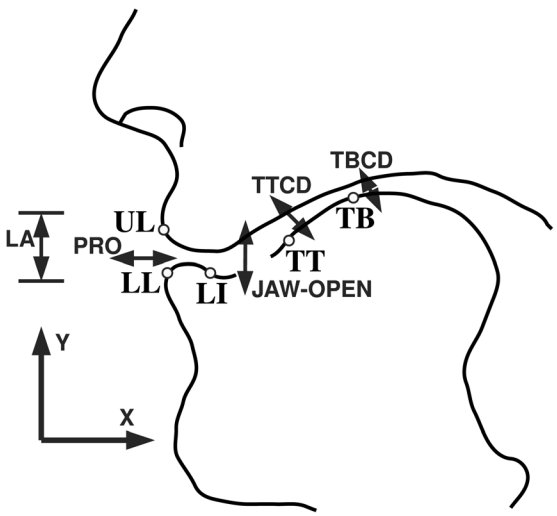

Fig. 1. Illustration of the TV features in the midsagittal plane.

the potential benefit of these estimated articulatory feature to automatic speech recognition, we performed experiments using the estimated TV features and compared the results with those obtained by using the measured (original) TV features.

4. Automatic speech recognition experiments

Our experiment focuses on a frame-based broad-class phonetic classification using GMM classifiers using data from each of the three subjects. Because we do not have sufficient data to train a full fledged HMM-based automatic speech recognition system, we assume the frame- based phonetic classification accuracy to be indicative of the recognition performance. The phone boundaries were obtained using the forced-alignment module in the SONIC recognizer[18] using an acoustic model set of 51 phones. Because the total number of frames corresponding to some phones was too few in our data to build individual GMMs for them, we grouped the data into five broad phone classes, namely, vowel, fricative, stops, nasal, and silence. Ninety percent of each subject's data (acoustic as well as articulatory) was used for training, and the remaining 10% was used as test set in a ten-fold cross validation setup. Note that in addition to original articulatory features, we also have three different estimates of the articulatory features for each subject from subject-independent acoustic-to-articulatory inversion by using the remaining two subjects as well as the talker himself as exemplars. When the talker is used as exemplar, the parallel acoustic-articulatory data in the training set for each fold is used as the training set for acoustic-to-articulatory inversion and the articulatory features for the utterances in the test set are estimated.

The feature space corresponding to each broad phone class is modeled using GMMs with four mixtures, and each mixture is modeled with multivariate Gaussian distribution with full co-variance matrices. The GMM parameters were estimated using the expectation maximization (EM) algorithm.[19] Each test frame is classified as one of the five broad phonetic categories using a max-*a posteriori* (MAP) rule. Table 2 shows the acoustic and acoustic-articulatory feature (using both original and estimated articulatory features) based phonetic classification accuracies for each English talker separately. Average classification accuracy using just acoustic features (MFCC) over 10-fold cross validation is reported in Table 2.[20] In addition, average accuracies for each talker are reported when estimated articulatory features are used to augment acoustic feature vectors. This allows us to compare the classification performance for the different exemplar choices. Table 2 also shows the average accuracy when the acoustic feature vector is augmented with the directly measured (original) articulatory features. We perform the Wilcoxon test[21] between acoustic-only and acoustic-articulatory feature based classification to investigate whether there is any significant improvement for including articulatory features. The P value resulting from the Wilcoxon test is reported next to the average accuracy. A lower P value indicates more significant improvement in the classification accuracy. In Table 2, we mark the accuracies by "bold" when the average classification accuracy using acoustic-articulatory feature is significantly better at 95%

Table 1. Average correlation coefficient (ρ) between original and estimated TV features for different talker and exemplar combinations.

		ρ for different TV features				
Test subject	Exemplar	LA	PRO	JAW_OPEN	TTCD	TBCD
EN_RD_MALE	EN_RD_FEMALE	0.45	0.20	0.58	0.60	0.63
	EN_SP_MALE	0.50	0.29	0.59	0.59	0.62
EN_RD_FEMALE	EN_RD_MALE	0.45	0.16	0.63	0.69	0.65
	EN_SP_MALE	0.60	0.15	0.66	0.73	0.59
EN_SP_MALE	EN_RD_FEMALE	0.55	0.28	0.52	0.70	0.62
	EN_RD_MALE	0.36	0.29	0.44	0.65	0.64

Table 2. Average phonetic classification accuracy using acoustic and acoustic-articulatory (both measured and estimated) features separately for each English subject. *P* values indicate the significance in the change of classification accuracy from the acoustic to acoustic-articulatory feature based phonetic classification.

		Acoustic-articulatory accuracy (*P* value)			
			Using exemplar (%)		
Talker	Acoustic only accuracy (%)	Using original TV features (%)	EN_RD_MALE	EN_RD_FEMALE	EN_SP_MALE
EN_RD_MALE	76.79	***79.05*** (0.002)	**77.37** (0.002)	**77.99** (0.002)	**77.23** (0.020)
EN_RD_FEMALE	79.10	***81.28*** (0.002)	**80.25** (0.002)	**80.16** (0.002)	**79.75** (0.010)
EN_SP_MALE	74.84	***76.29*** (0.002)	74.87 (0.084)	74.97 (0.131)	**75.17** (0.002)

significance level than just with the acoustic feature. We also mark the highest acoustic-articulatory phonetic classification accuracy for individual subject by "underline."

5. Discussion of the experimental results

The results of Table 2 show that the frame-based phonetic classification accuracy for all the English talkers significantly improves when the measured (original) TV features are used to augment the acoustic features. When the original TV features are replaced with the estimated TV features, the nature of improvement obtained depends on the characteristics of the "exemplar" used to estimate the TV features. We also observe that the benefit due to original articulatory features is more compared to that due to estimated articulatory features for each talker considered in this experiment. This observation suggests that the better estimates of the articulatory features could lead to better phonetic classification accuracy. The selection of "exemplar" also plays a crucial role in determining the quality of the articulatory estimates and, hence, the recognition benefit. For example, when we consider EN_RD_MALE or EN_RD_FEMALE as talker, the improvement in classification due to EN_RD_FEMALE and EN_RD_MALE as exemplars is more than that due to EN_SP_MALE as exemplar. This could be due to the fact that EN_RD_MALE and EN_RD_FEMALE are British English talkers and EN_SP_MALE an American English talker. Furthermore the MOCHA TIMIT represents read speech while the USC MURI database, spontaneous speech, and this could contribute to poor estimates due to increased talker-exemplar speaking style mismatch. It is also interesting to observe that the EN_SP_MALE exemplar and EN_RD_MALE talker are of the same gender, yet an exemplar of a different gender (EN_RD_FEMALE) provided more phonetic classification accuracy for EN_RD_MALE talker. When we consider the American English talker (EN_SP_MALE), we find the classification improvement obtained using British English subjects as exemplars is not statistically significant. This means that the general acoustic space in subject-independent inversion could account for gender differences more effectively compared to speaking dialectal and style differences between the talker and the exemplar.

Finally, frame-based phonetic classification experiments with articulatory features estimated using identical exemplar-talker combination were performed to examine the extent of improvement in classification when the talker's articulatory-to-acoustic map itself is used for subject-independent inversion. For every training-test set combination of individual talkers, the parallel articulatory and acoustic data of the training set are used to estimate the articulatory features for the test sentences. It appears that the use of identical talker and exemplar does not always guarantee the maximum improvement in phonetic classification among different exemplars. This may reflect the data limitations in deriving articulatory–acoustic maps that can cover the range of expected test feature variability. For example, when EN_RD_MALE is considered as talker, the exemplar EN_RD_FEMALE of similar speaking style resulted in

more benefit compared to that using identical talker-exemplar scenario. This also holds for EN_RD_FEMALE talker and EN_RD_MALE exemplar. However, for EN_SP_-MALE talker, the identical talker-exemplar combination provides the highest improvement in phonetic classification among other exemplars.

Thus, the choice of "exemplar" plays a critical role to improve the recognition for a given talker. Our results suggest that when the "exemplar" is chosen to have same speaking style as that of the talker, there is a significant benefit in using estimated articulatory features in addition to the acoustic features to improve speech recognition.

6. Conclusions

We investigated the potential of using articulatory features estimated through acoustic-articulatory inversion in automatic speech recognition. We conducted subject-specific broad- class phonetic classification experiments using data from three different native English speaking subjects. We find that the selection of "exemplar" for the subject-independent acoustic-to- articulatory inversion has a critical impact on the quality of the articulatory feature estimates and, hence, the final phonetic classification accuracy. In particular, our experimental results suggest that when the talker and the "exemplar" characteristics are matched in their speaking styles, the improvement in classification due to estimated articulatory features is significant.

Acknowledgments

This work was supported in part by NIH and ONR-MURI.

References and links

[1]J. Frankel and S. King, "ASR—articulatory speech recognition," in Proceedings of Eurospeech, Scandinavia, (2001), pp. 599–602.

[2]L. Deng, G. Ramsay, and D. Sun,"Production models as a structural basis for automatic speech recognition," Speech Commun. **22**(2), 93–112 (1997).

[3]H. Attias, L. Lee, and L. Deng, "Variational inference and learning for segmental switching state space models of hidden speech dynamics," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, Vol. 1 (2003), pp. I-872 – I-875.

[4]J. Ma and L. Deng,"Target-directed mixture dynamic models for spontaneous speech recognition," IEEE Trans. Speech Audio Process. **12**(1), 47–58 (2004).

[5]E. McDermott and A. Nakamura, "Production-oriented models for speech recognition," IEICE Trans. Inf. Syst. **E89-D**(3), 1006–1014 (2006).

[6]F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," International Conference on Spoken Language Processing, Denver, CO, USA (2002), pp. 2133–2136.

[7]J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman,"Accurate recovery of articulator positions from acoustics: New conclusions based on human data," J. Acoust. Soc. Am. **100**(3), 1819–1834 (1996).

[8]H. Yehia, "A study on the speech acoustic-to-articulatory mapping using morphological constraints," Ph.D. thesis, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan, 2002.

[9]T. Toda, A. Black, K. Tokuda,"Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," Speech Commun. **50**, 215–217 (2008).

[10]P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Prague, Czech Republic (2011), pp. 4624–4627.

[11]D. Yu, L. Deng, and A. Acero,"Speaker-adaptive learning of resonance targets in a hidden trajectory model of speech coarticulation," Comput. Speech Lang. **27**, 72–87 (2007).

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

[13]J. Silva, V. Rangarajan, V. Rozgic, and S. S. Narayanan, "Information theoretic analysis of direct articulatory measurements for phonetic discrimination," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Honolulu, HI, USA (2007), pp. 457–460.

[14]For articulatory representation, one can also use raw X, Y values of the sensor positions of common articulators across subjects. TV features represent a "low-dimensional" (5×1) control regime for constriction actions in speech production and are considered more invariant in a linguistic sense.

[15]C. P. Browman and L. Goldstein, "Towards an articulatory phonology," Phonol. Yearbook **3**, 219–252 (1986).
[16]John S. Garofolo *et al.*, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, Philadelphia, PA (1993).
[17]P. K. Ghosh and S. S. Narayanan, "A generalized smoothness criterion for acoustic-to- articulatory inversion," J. Acoust. Soc. Am. **128**(4), 2162–2172 (2010).
[18]B. Pellom and K. Hacioglu, "Sonic: The University of Colorado continuous speech recognizer," Technical Report No. TR-CSLR-2001-01 (2005).
[19]A. P. Dempster, N. M. Laird, and D. B. Rubin,"Maximum likelihood from incomplete data via the EM algorithm," J. R. Stat. Soc. Ser. B (Methodol.) **39**(1), 1–38 (1977).
[20]We did not explore delta and delta-delta MFCC as acoustic features due to the increase in feature dimension, which in turn requires more data for reliable estimates of GMM parameters; this is not afforded by the corpus limitations of the present study.
[21]M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods* (Wiley, New Jersey, 1999).