Information theoretic analysis of direct and estimated articulatory features for phonetic discrimination

Prasanta Kumar Ghosh and Shrikanth S. Narayanan
Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering,
University of Southern California, Los Angeles, CA 90089, USA.

It is well known that machine recognition of speech can be improved by including direct articulatory evidence in addition to the signal information derived from the acoustic speech. This has been shown through automatic phonetic recognition experiments [1] as well as by information theoretic analysis between phonetic classes and the acoustic/articulatory features [2]. However, access to such direct speech articulation data during the test phase of a recognizer is not feasible in practice. One way for estimating articulatory features is through acoustic-to-articulatory inversion; the question of interest then is how such derived articulatory information can contribute to machine speech recognition. First, we need a method for robust acoustic-articulatory inversion given any arbitrary talker. We have recently developed [3] a talker-independent acoustic-to-articulatory inversion which uses parallel acoustic and articulatory data from only one subject (we denote this subject as 'listener' in this work) and a generic acoustic model to estimate articulatory features from acoustic speech signal from any arbitrary talker. The goal of the present work is to investigate how much information these estimated articulatory features provide in addition to the acoustic features for phonetic discrimination. We do this by calculating the mutual information (MI) between the data (direct or derived articulatory data; acoustic data) and the phonetic classes of interest [6]. The effectiveness of the estimated articulatory features for phonetic discrimination can be inferred from the MI analysis using Fano's inequality [6] – higher MI indicates more discriminative power of the features.

For our experiments, we have used the MOCHA-TIMIT corpus [4], which provides parallel acoustic and electromagnetic articulography (EMA) data for one male and one female subject. We have used one of these subjects as listener and other as talker to perform our analysis; this means we have two different talker-listener pairs. We have used 14-dimensional mel frequency cepstral coefficients (MFCC) as the acoustic feature. We have designed articulatory features motivated by the concept of tract variables in the articulatory phonology [5]. According to articulatory phonology, speech is represented using an ensemble of gestures, which are defined as the dynamical control regimes for constriction actions in eight different constriction tract variables consisting of five constriction degree variables, lip aperture (LA), tongue body (TBCD), tongue tip(TTCD), velum (VEL), glottis (GLO), and three constriction location variables, lip protrusion (PRO), tongue tip (TTCL), tongue body (TBCL). Based on the available EMA data, we have estimated a subset of the above tract variables (TV) features, namely – LA, PRO, TTCD, TBCD, VEL, JAW_OPEN. JAW_OPEN is the vertical distance of the JAW sensor from the reference sensor on the nose. Thus our target articulatory feature is a 6-dimensional vector. In our experiment, we also examine the velocity (Δ) and acceleration ($\Delta\Delta$) of MFCC and TV features. The acoustic speech signal is passed through an automated force-alignment module (using Viterbi decoding) to obtain the phonetic labels at each analysis frame. An information theoretic analysis is performed on the acoustic and articulatory (direct and estimated by the listener) features of the talker.

We use MI [6] to quantify the amount of information that the acoustic or the joint acoustic-articulatory features provide about the phonetic classes. Following the procedure in [2], we have quantized the feature space into 128 centroids using K-means clustering (higher number of clusters did not change the result much). Note that the dimensionality of the feature vector changes depending on the type of the feature; however, the number of frames available for each subject is constant. Thus the estimate of the mutual

information may be affected by the sparsity of the data particularly for a high dimensional feature vector. Also there are small number of frames corresponding to some phonemes. Thus for reliability of the results and conclusions, we have used four broad phonetic classes – vowel, fricative, nasal, stops – and silence. The MI for different acoustic-articulatory feature combinations are shown in Fig. 1 for both the male and female subjects. It can be seen from Fig. 1 that both reference (direct) and estimated articulatory

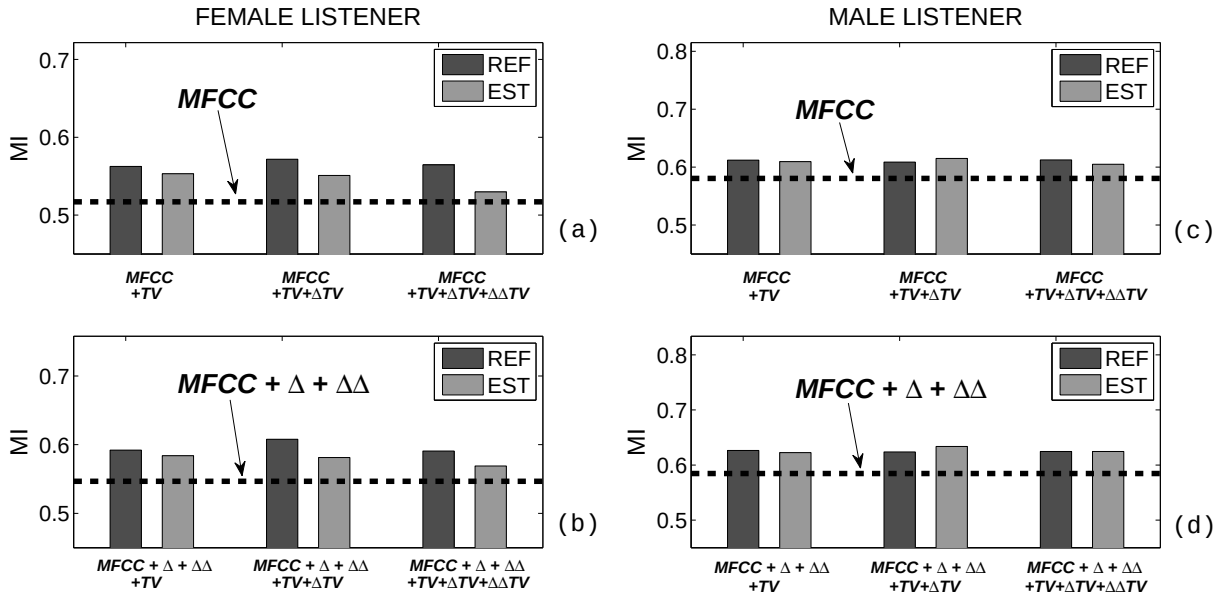

Figure 1: (a) and (b) – The MI obtained by adding direct (darker bar) and estimated (lighter bar) articulatory features to MFCC and MFCC+Δ + $\Delta\Delta$ features respectively (listener: female subject and talker: male subject). (c) and (d) – (a) and (b) repeated when male and female subjects act as listener and talker. Horizontal dashed lines indicate the MI obtained by the acoustic features in the respective cases.

features increases MI when added to the acoustic features. For the male talker (Fig. 1(a)-(b)), the MI is consistently higher for the direct TV features and its derivatives compared to the estimated ones. However, for female talker, the MI is more in the case of estimated TV+ΔTV compared to the reference ones. Overall, encouragingly, the MI values for the different cases considered suggest that the TV features, although estimated by using the knowledge of acoustic-to-articulatory map of a subject different from the target talker, provide similar information as that provided by the direct TV features. It is also worth noting that the use of estimated TV features results in improvement of MI when added to MFCC+Δ+$\Delta\Delta$ features although TV features were estimated only based on MFCC features. These observations indicate that the articulatory features estimated using a talker-independent acoustic-to-articulatory inversion can be potentially useful for phonetic discrimination task. Broad-class phonetic recognition experiments also support our findings based on MI analysis.

References

[1] J. Frankel and S. King, "ASR - articulatory speech recognition," Proc. Eurospeech, Scandinavia, pp. 599–602, 2001.

[2] J. Silva, V. Rangarajan, V. Rozgic, and S. S. Narayanan, "Information theoretic analysis of direct articulatory measurements for phonetic discrimination," Proc. ICASSP, pp. 457–460, 2007.

[3] Prasanta Kumar Ghosh and Shrikanth S. Narayanan, "A subject-independent acoustic-to-articulatory inversion", under review.

[4] 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, pp. 305–308, 2000.

[5] C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," Phonology, vol. 6, pp. 201–251, 1989.

[6] T. M. Cover and J. A. Thomas, Elements of Information Theory (Wiley Interscience, New York, 1991), pp. 12–49.