Information theoretic acoustic feature selection for acoustic-to-articulatory inversion

Prasanta Kumar Ghosh[1], *Shrikanth S Narayanan*[2]

[1]Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India
[2]Department of Electrical Engineering, University of Southern California, Los Angeles, U.S.A.
prasantg@ee.iisc.ernet.in, shri@sipi.usc.edu

Abstract

We use mutual information as the criterion to rank the Mel frequency cepstral coefficients (MFCCs) and their derivatives according to the information they provide about different articulatory features in acoustic-to-articulatory (AtoA) inversion. It is found that just a small subset of the coefficients encodes maximal information about articulatory features and interestingly, this subset is articulatory feature specific. We use these subsets of MFCCs(+derivatives) in AtoA inversion using Gaussian mixture model (GMM) mapping. Inversion experiments with articulatory data support the information theoretic finding that the subsets of MFCCs(+derivatives) as selected by feature ranking method are sufficient to achieve an inversion performance similar to that obtained by a conventional full set of MFCCs(+derivatives). This drastically reduces the modeling complexity of the acoustic-articulatory map using GMM without degrading inversion performance significantly.
Index Terms: Acoustic-to-articulatory inversion, mutual information, Gaussian mixture model.

1. Introduction

Estimation of articulatory representation from speech acoustics is known as acoustic-to-articulatory (AtoA) inversion. The articulatory space can be represented in a variety of ways including through 1) stylized models such as Maeda's model[1, 2] or the lossless tube model[3] of the vocal tract, 2) linguistic rule based models[4, 5, 6] or 3) direct physiological data-based representations of articulatory information[7]. In this work, we consider the physiological data based representation of the articulatory space, where articulatory data (e.g. position of the lips, jaw, tongue, velum etc.) during speech production are obtained directly from the talkers by means of a specialized instrument such as an electromagnetic articulograph (EMA).

The speech acoustic space could be represented using different types of features including Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Coefficients (LPC), Line Spectral Pairs (LSP), and Log Area Ratios (LAR). Experimental comparison among different commonly used acoustic features for AtoA inversion was reported in [8, 9]. It was found that MFCCs are the best acoustic representation in terms of providing maximum information about the articulatory movements and hence the best choice for AtoA inversion. Based on this experimental analysis we have chosen MFCCs along with its first and second derivatives as the acoustic representation for our present investigation.

While MFCCs were shown to be the best acoustic features for AtoA inversion, the relative role of different coefficients in the MFCC feature vector in estimating different articulatory features has not been well explored. The premise of this work is that the information about the movements of different articulators could be encoded along different degrees in various cepstral coefficients. The goal in this paper is to rank different cepstral coefficients and their derivatives according to the amount of information provided by them to estimate various articulatory features in AtoA inversion. We use mutual information (MI) as the ranking criterion for this purpose. Since articulatory dynamics describe the speech production process, the information theoretic analysis done in this work could provide a production-oriented interpretation of various MFCCs.

We then extend our analysis of MI based acoustic feature selection by conducting AtoA inversion experiments. This is done to examine the efficacy of the acoustic feature selection method directly in terms of the AtoA inversion performance. AtoA inversion is performed using a subset of cepstral coefficients selected in an information theoretic manner. There are several parametric and non-parametric approaches to statistically model the mapping between acoustic and articulatory representations; this mapping function is used to invert acoustic features to the respective articulatory representations. A good summary of different approaches for AtoA inversion can be found in [10]. Gaussian mixture model (GMM) is one of the well-known parametric models used for AtoA mapping [11]. In this work, we use GMM mapping for AtoA inversion task. By conducting GMM based AtoA inversion we quantify the inversion performance by replacing the full acoustic feature vector with its subset as selected by the proposed MI criterion.

2. Dataset and pre-processing

For the analysis and experiments of this paper, we use the Multichannel Articulatory (MOCHA) database [12] that contains speech and corresponding ElectroMagnetic Articulography (EMA) data from one male and one female talker of British English. The EMA data consist of dynamic positions of the EMA sensors in the midsagittal plane of the talker. Seven sensors are placed on upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue body (TB), tongue dorsum (TD), and velum (VEL)). Thus, we use 14 dimensional raw EMA features for representing articulatory space (i.e., X and Y co-ordinates of seven EMA sensors), namely ULx, LLx, LIx, TTx, TBx, TDx, VELx, ULy, LLy, LIy, TTy, TBy, TDy, VELy. The articulatory position data have high frequency noise resulting from the EMA measurement error. Also the mean position of the articulators changes from utterance to utterance; hence, the position data needs pre-processing before it can be used for analysis. Following the preprocessing steps outlined in [8], we obtain parallel acoustic and articulatory data at a frame rate of

100 observations per second. Acoustic feature MFCCs are computed [13] using 20 msec frame length with 10 msec shift. Each MFCC feature vector is 13 dimensional, where 13-th coefficient represent the log energy of the short-time frame. The first and second derivatives of MFCCs are computed and appended to the MFCC feature vector constructing a 39 dimensional acoustic feature vector.

3. Acoustic feature selection

Given a 39 dimensional acoustic feature vector $\mathbf{x}$ (i.e., MFCCs and their derivatives), we aim to select the best K ($1 \leq K \leq 39$) acoustic features denoted by $\mathbf{x}^K$ which will provide least uncertainty about a given articulatory feature z (i.e., one of the 14 EMA sensors). In other words we would like to find a subset of acoustic features of cardinality K such that the mutual information (MI) between $\mathbf{x}^K$ and z, denoted by $I(\mathbf{x}^K, z)$, is maximum. MI quantifies the statistical dependency between $\mathbf{x}^K$ and z and is negatively proportional to the conditional uncertainty $H\left(z|\mathbf{x}^K\right)$ [14]. Thus higher the MI, better will be the estimate of z when $\mathbf{x}^K$ will be used as the acoustic representation in AtoA inversion. Note that $\mathbf{x}^{39}$=$\mathbf{x}$, i.e., K=39 corresponds to the full acoustic feature vector without any feature selection.

$\mathbf{x}^K$ and z are continuous random variables and hence the joint probability density function (PDF) of $\mathbf{x}^K$ and z is necessary to compute MI. However, we only have realizations of $\mathbf{x}^K$ and z in the MOCHA database and hence their joint PDF is not known; therefore, we consider MI estimation by quantization of the space of $\mathbf{x}^K$ and z and applying a definition of MI for discrete random variables as was done in [8]. K-means clustering is used to perform this quantization (for details of MI computation see [8]). 64 quantization bins are used to compute MI. Higher values of quantization bins yield similar result and hence we report results with only 64 bins.

Selection of K best coefficients among 39 MFCCs is a combinatorial problem and hence computationally expensive. Therefore we use a forward feature selection method [15]. In the i-th iteration of the forward selection method, the i-th best feature is selected from the remaining 40-i coefficients (i.e., 39 coefficients except the already selected i-1 coefficients). The steps for this MI based feature selection are described in Algorithm 1. In each iteration, the best feature is determined by picking the one which, when added to already selected features, yields maximum MI with the articulatory feature. The indices of the best K coefficients and the corresponding MI values are returned in ζ and γ respectively. The l-th element of γ, i.e., γ_l is the MI between the articulatory feature and the subset of acoustic features with cardinality K=l, whose indices are stored in $\zeta_1, \cdots, \zeta_l$.

We compute γ and ζ for all 14 articulatory features separately for both the male and female subject in the MOCHA corpus. Fig. 1 and 2 show plots of γ_l vs l=1,...,39 with respect to LLx, VELx, ULy, TTy for the male and female subject respectively. These four articulatory features are chosen to illustrate the increasing trend of the MI with increasing values of K. It is clear from the plots in Fig. 1 and 2 that only 5-10 selected cepstral coefficients yield an MI similar to that obtained by all the coefficients. A similar trend is observed with other articulatory features too. Such a steep increase of MI (γ_l) with l suggests that the information about the individual articulators is essentially encoded in a few cepstral coefficients. For example, with only the top 10 selected acoustic features, we obtain an MI which is at least greater than 97% (and 94%) of the corresponding maximum MI for any articulatory feature for the male (and female) subject. It should be noted (from Fig. 1 and 2) that γ_l does not always increase monotonically with l. This would have been the case if we could have access to the true MI values. γ being an estimate of the original MI, such non-monotonicity could happen due to the estimation error.

Algorithm 1 MI based acoustic feature selection - inputs: $\mathbf{x}$, z and K

1: $\mathbf{x}_r = \mathbf{x}$, (note that the cardinality of $\mathbf{x}$ ($|\mathbf{x}|$) = 39). γ, ζ are initialized as empty vectors. $\mathbf{x}_\zeta$ denotes a $|\zeta|$-dimensional vector obtained from $\mathbf{x}$ keeping only elements with indices in ζ.
2: **for** l=1 to K **do**
3: **for** i=1 to $|\mathbf{x}_r|$ **do**
4: $\eta_i \leftarrow I(\mathbf{x}_{r,i}, z)$, where $\mathbf{x}_{r,i}$ is the i-th element of $\mathbf{x}_r$
5: **end for**
6: $\gamma_l \leftarrow \max_i \eta_i$
7: $\zeta_l \leftarrow \arg\max_i \eta_i$
8: $\mathbf{x}_r \leftarrow \mathbf{x} \setminus \mathbf{x}_\zeta$
9: **end for**
10: Return γ and ζ

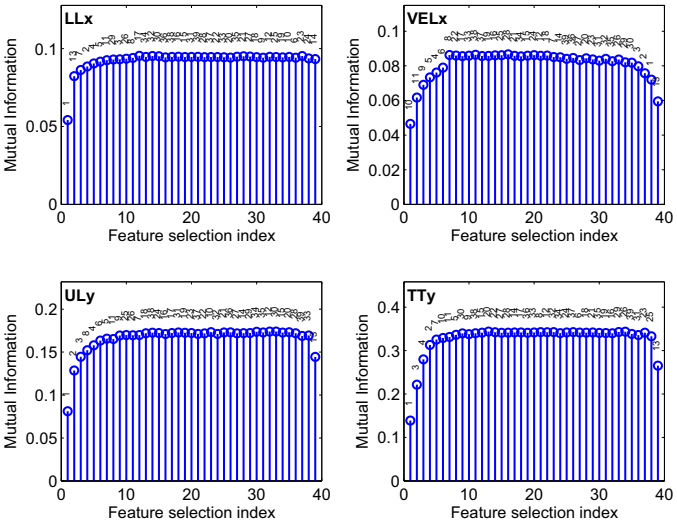

Figure 1: *MI based feature selection for MOCHA-male subject. γ_l vs $l = 1, ..., 39$ is plotted for four different articulators. The integers printed on the stem plot of γ_l are the corresponding ζ_l, the l-th best acoustic feature index selected using MI criterion.*

Fig. 1 and 2 also show the indices ζ_l of the selected MFCCs above the γ_l plots for each l. Considering K=10, it is interesting to see that, for most of the articulators, the dynamic features (i.e., first and second derivatives of MFCCs) appear to be more information bearing than some of the static features. For example, top 10 (K=10) selected acoustic features have at least one first or second derivative feature in the case of 11 (13) articulatory features for the male (female) subject. It is also interesting to note that the top 10 selected features are articulatory feature specific. This suggests that the subset of acoustic features that encodes most of the information about articulatory features changes depending on the articulator. However, if different articulators are correlated to each other then the corresponding selected acoustic features also appear to be similar. For example, the tongue sensors' movements are expected to be correlated

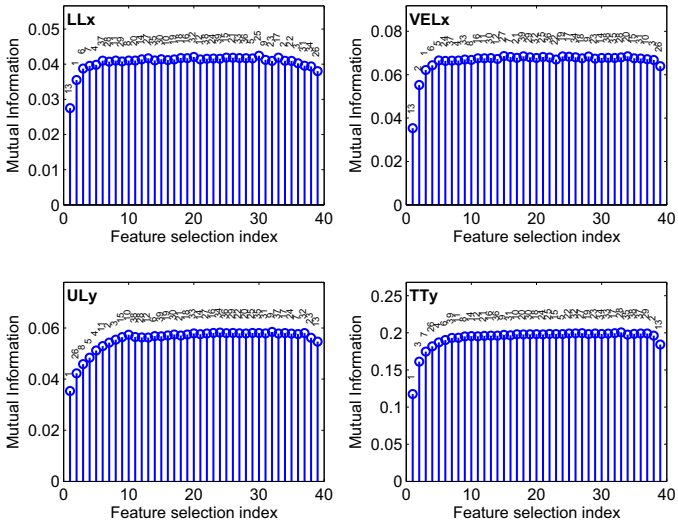

Figure 2: *MI based feature selection for MOCHA-female subject. The plots are similar to that of Fig. 1.*

since the tongue sensors are attached to different parts of one articulator namely tongue. We find that, in the case of the female subject, the indices of the top 5 selected acoustic features for TTx, TBx, TDx are {1 3 2 5 7}, {1 3 2 8 5} and {4 3 2 1 5} respectively indicating a similarity among the selected acoustic feature subsets for these correlated articulatory features. Articulatory feature specific nature of the subsets selected by MI criterion also suggests that different cepstral features provides varying degrees of information about the different articulators.

The selected subset of acoustic features also changes depending on the MOCHA subject considered even for the same articulator. This is probably due to the fact that the details of the acoustic-articulatory map are different for different subjects and hence information about one articulatory feature may be encoded in different cepstral features depending on the subject.

We also observe that unlike any other coefficients the 13-th MFCC (i.e. the log energy coefficient) occurs most of the time as the last selected acoustic features for both the subjects suggesting the energy coefficient to be least information bearing for most of the articulators. For example in the case of male subject 13-th MFCC appears as the last selected acoustic features for all articulators except LLx, LLy suggesting short-time energy to be least information bearing for most of the articulators. For LLx, LLy 13-th MFCC appear in the top 2 selected features. Similarly for the female subject 13-th MFCC appears as the last selected acoustic features for all articulators except LLx, LIx, VELx, LLy, LIy, VELy, for which 13-th MFCC appears in the top 3 selected features. Appearance of the energy coefficients in the top few selected features for lower lip movement for both subjects indicates a strong correlation of the log energy trajectory with the lower lip sensor trajectory.

4. Acoustic-to-articulatory inversion experiments

The key finding of the MI based acoustic feature selection experiment is the compact encoding of articulatory information in few cepstral coefficients. This could be potentially useful to reduce the complexity of the modeling of the acoustic-articulatory map in AtoA inversion as we will see in this section. We conduct AtoA inversion experiment on both subjects of the MOCHA database to evaluate the potential of MI based acoustic feature selection for AtoA inversion directly in terms of inversion performance. Since only a few acoustic features carry information about each articulatory feature, it is expected that use of a few selected acoustic features should be sufficient to achieve an inversion performance similar to that using all acoustic features. Below we describe the AtoA inversion experimental setup and results.

4.1. Experimental setup

AtoA inversion is performed separately on the male and female subject of the MOCHA corpus. Each subject's acoustic and articulatory data is used in a 5-fold cross-validation setup for this purpose. AtoA inversion is performed in an articulatory feature specific manner using the corresponding top K selected acoustic features. We report the inversion performance as an average performance over all sentences of all folds. Root mean squared error (RMSE) between the original and estimate articulatory trajectories of each sentence is used as a performance measure. The RMSE reflects the average closeness between the original and estimated articulatory features. However, a minimum RMSE does not always mean the trajectories are similar; for example, the estimated one can be very jagged although it might be close to the actual one. Thus, as an additional performance measure, we compute Pearson correlation coefficient (PCC) [17] between the original and estimated feature trajectories of each sentence. Average RMSE and PCC over all sentences are used to measure the quality of inversion.

GMMs are used to model the acoustic-articulatory map of the training data separately for each articulator in each fold. Toda et. al. [11] showed that a GMM with 64 mixture components results in the minimum RMSE on the MOCHA corpus [11]. Therefore we use 64 mixture component GMMs with full co-variance matrices for our experiments in this paper. The Expectation Maximization algorithm is used to estimate the parameters of GMM [16]. In each fold we separately consider every articulatory feature along with its corresponding top K selected acoustic features. A GMM is trained on K+1 dimensional space using the training data and a conditional GMM of the articulatory feature given the K dimensional acoustic feature space is used to obtain the estimate of the articulatory feature from the acoustic features (for details on GMM based inversion see [11]).

It is well-known that the estimated articulatory trajectory obtained using GMM is rough and jagged. Smoothing the GMM based estimate reduces the RMSE of the AtoA inversion [11]. We also follow a similar procedure and smooth the estimated articulatory trajectory by low-pass filtering. To avoid any phase distortion due to the low pass filtering on the estimated trajectories, the filtering process is performed twice ("zero-phase filtering") - the trajectory is initially filtered and then reversed and filtered again and reversed once more finally. The cut-off frequency of the low-pass filter is varied from 3Hz to 25Hz with a step-size of 0.5Hz for each articulator. The cut-off frequency which led to minimum RMSE is used for smoothing and the respective average RMSE and PCC are reported as the inversion performance.

4.2. Results and discussions

We compare the AtoA inversion performance with and without acoustic feature selection for each articulator of both subjects in MOCHA database. K in acoustic feature selection is chosen

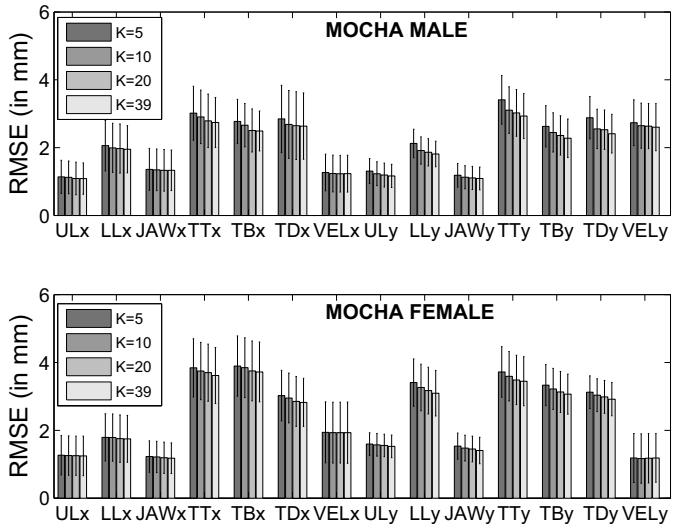

Figure 3: *RMSE of AtoA inversion for different articulatory features in case of the male and female subject of MOCHA database. Average RMSE values (± one SD) are indicated by error bars for different number of selected acoustic features (K) – 5, 10, and 20. RMSE is also shown for K=39, which is equivalent to acoustic feature without any feature selection.*

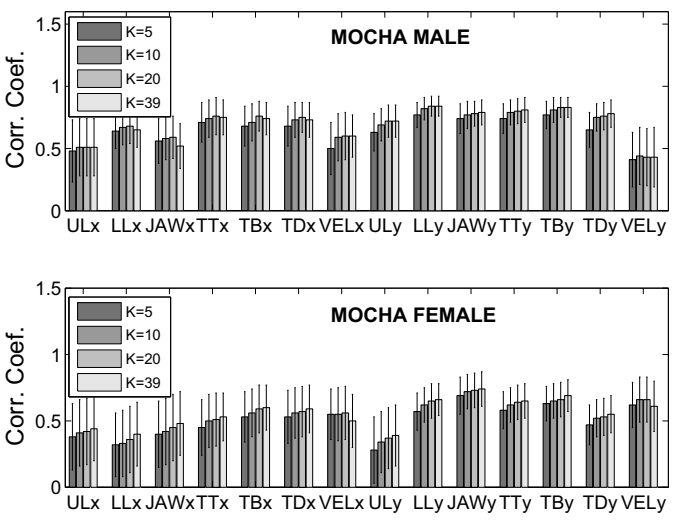

Figure 4: *Pearson Correlation Coefficient of AtoA inversion for different articulatory features in case of the male and female subject of MOCHA database (Similar to the RMSE measure of Fig. 3).*

empirically as 5, 10 and 20 in our experiment; whereas K=39 corresponds to the case when no acoustic feature selection is done. Average RMSE (± one standard deviation (SD)) for each articulator with K=5, 10, 20 and 39 are shown in Fig. 3 for both the male and female subject. Similar graphs of inversion performance using PCC are shown in Fig. 4. It should be noted that usually in GMM-based AtoA inversion [11], the articulatory features are jointly estimated by modeling the acoustic and all articulatory features together using GMM. However, we perform GMM-based AtoA inversion for each articulator separately even for K=39. We have found that there is no significant difference in performance between inversion done on each articulator separately and inversion on all articulators jointly.

It is clear from Fig. 3 that with only the top five selected acoustic features (i.e., K=5) the average RMSE is very close to that obtained without any feature selection; in fact the difference between the two is statistically insignificant which is also clear from the overlapping errorbars of K=5 and K=39 cases. This is true for all articulators of both MOCHA subjects. The average RMSE decreases with increasing K; however, the difference with K=39 is statistically insignificant for all choices of K. A similar result is found when PCC is used as the performance measure for inversion in Fig. 4. We observe that the average PCC increases with K although there is no statistically significant difference between with (K=5, 10, 20) and without (K=39) feature selection cases. For some articulators the average PCC for K=39 is lower than that for K=10. This happens, for example, in the case of MOCHA female subject for VELx, VELy and in the case of MOCHA male subject for LLx, JAWx. This could be due to the poor quality of GMM parameter estimation particularly in the case of high acoustic feature dimension (K=39); for K=39 the number GMM parameters is large and in the order of the number of training data points unlike that for K=10 and hence the quality of parameter estimate could suffer for K=39 causing a lower average PCC compared to that for K=10.

The acoustic feature after feature selection has less dimension (K) than that of the full acoustic feature; this in turn reduces the complexity of modeling the acoustic-to-articulatory map using GMMs. With the full acoustic feature set of 39 dimensions, the number of GMM parameters for each articulator is $64\times(1+40+40^2)$ = 105024 (GMM is trained on 39+1=40 dimension for each articulator and parameters consist of 64 mixture weights, 40-dim mean vector and 40×40-dim full covariance matrix for each mixture). On the other hand for acoustic feature with feature selection K=5, the number of GMM parameters for each articulator is $64\times(1+6+6^2)$ = 2752. This is 97.38% reduction in the number of GMM parameters required to be trained for AtoA inversion. Such a large percentage of reduction in the number of parameters improves both the quality of GMM parameter estimates as well as the complexity of the GMM based modeling of AtoA mapping without losing the inversion performance significantly.

5. Conclusions

We found that a subset of acoustic features MFCC (+derivatives) can be used to estimate the articulatory features in AtoA inversion with an accuracy similar to that provided by a conventional full set of acoustic features. We have also shown that these subsets of acoustic features can be determined a-priori using a forward feature selection method using an information theoretic criterion. Due to the reduction in acoustic feature space required for AtoA inversion, the modeling complexity of the acoustic-to-articulatory map goes down appreciably. The feature selection also provides insight about the amount of information that different acoustic features provide about various articulatory features. This analysis also offers a production-oriented interpretation of some of the MFCCs, which are computed based on speech perception studies.

6. Acknowledgements

Work supported by Department of Science and Technology (DST), Govt. of India.

7. References

[1] S. Maeda, "Un modele articulatoire de la langue avec des composantes lineaires (an articulatory model of the tongue with linear components)," Actes 10emes Journees dEtude sur la Parole (Grenoble, France), 152–162, 1979.

[2] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," Speech production and speech modelling, edited by W. Hardcastle and A. Marchal (Kluwer Academic Publishers, Dordrecht, The Netherlands), 131–149, 1990.

[3] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," Proc. Fourth Int. Congr. Acoust., Copenhagen, 1–4, 1962.

[4] C. P. Browman and L. Goldstein, "Towards an articulatory phonology," Phonology Yearbook 3, 219–252, 1986.

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

[6] C. P. Browman and L. Goldstein, "Gestural specification using dynamically-defined articulatory structures," Journal of Phonetics 18, 299–320, 1990.

[7] 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, 305–308, 2000.

[8] P. K. Ghosh and S. S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," J. Acoust. Soc. Am., vol. 128, no. 4, pp. 2162–2172, 2010.

[9] C. Qin and M. A. Carreira-Perpinan, "A comparison of acoustic features for articulatory inversion," Proc. Interspeech, 2469–2472, 2007.

[10] A. Toutios and K. Margaritis, "Acoustic-to-articulatory inversion of speech: a review," Proceedings of the International 12th TAINN, 2003.

[11] T. Toda, A. W. Black, and K. Tokuda. "Acoustic-to-articulatory inversion mapping with Gaussian mixture model", Proc. INTERSPEECH, pp. 1129–1132, Jeju, Korea, Oct. 2004.

[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, 305–308, 2000.

[13] Hidden Markov Model Toolkit (HTK), Online: http://htk.eng.cam.ac.uk, accessed on 16 Mar 2013.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, (Wiley Interscience, New York, 1991).

[15] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (New York:Wiley, 1983).

[16] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Series B, 39(1):1–38, 1977.

[17] J. Benesty, J. Chen, Y. Huang, and I. Cohen. "Pearson Correlation Coefficient." Noise reduction in speech processing, Springer, 1–4, 2009.