Better acoustic normalization in subject independent acoustic-to-articulatory inversion: benefit to recognition

Amber Afshan[1]*, Prasanta Kumar Ghosh*[2]

[1]Department of Electrical Engineering, University of California, Los Angeles, USA,
[2]Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India.
amberafshan0107@gmail.com, prasantg@ee.iisc.ernet.in

Abstract

In subject independent acoustic-to-articulatory inversion (SII), the training and test subjects are in general different, whereas subject dependent inversion (SDI) uses the same training and test subjects. Thus, acoustic normalization is used to compensate for the mismatch between the training and the test subjects in SII. We show that a better acoustic normalization not only results in better articulatory estimates using SII, but also improves the broad class phonetic recognition accuracy, when the articulatory features estimated from SII are used for recognition. Recognition experiments using male and female subjects from the MOCHA-TIMIT corpus also show that there is no significant difference between the recognition accuracy using the articulatory features obtained by the best acoustic normalization in SII and that obtained using SDI as well as directly measured articulatory features.

Index Terms: broad class phonetic recognition, acoustic-to-articulatory inversion, subject independent inversion, acoustic normalization

1. Introduction

Kinematics of speech articulators (e.g., lips, jaw, tongue, velum) recorded during speech production are known to provide cues for automatic speech recognition (ASR) [1, 2]. These articulatory features are also known to provide information complementary to acoustic features obtained from the speech signal [3]. Recording articulatory kinematics is not convenient in practice unlike recording of speech signal. This hinders the use of directly measured articulatory data for ASR. In the absence of directly measured articulatory features, estimating them from the speech signal becomes a plausible option. The task of estimating articulatory features from acoustic representation is known as acoustic-to-articulatory inversion (AAI) [4]. AAI can be of two types: 1) subject-dependent inversion (SDI), where acoustic-articulatory data from the test subject is available for training AAI algorithm [5, 6, 7, 8, 9, 10, 11, 12, 13, 14], 2) subject-independent inversion (SII) [15, 16, 17], where the test subject can in general be different from the training subject. SII is more challenging compared to SDI due to the mismatch between the training and test subjects. At the same time, SII is more appropriate compared to SDI when the estimated articulatory features are to be used for ASR on any arbitrary test subject, because the acoustic or articulatory data from the test subject may not be available a-priori. Thus, in this work we conduct speech recognition using articulatory features estimated using SII.

Acoustic normalization is used to compensate for the mismatch between the training and test subjects' acoustics in SII [16, 17]. This is done by constructing a probability feature vector by transforming the acoustic features of train and test subjects on a generic acoustic space (GAS) consisting of a large pool of acoustic features from multiple speakers [16]. GAS may not include the acoustic data of the training and test subjects of SII. The probability feature vectors of two subjects are comparable unlike their acoustic feature vectors. It has been shown that the acoustic normalization in SII can be improved by appropriately choosing the acoustic units in GAS [17]. For example, the phonetic units are found to be more effective for normalization compared to acoustic units obtained by unsupervised clustering. Similarly, when the states of a phonetic hidden Markov model (HMM) are used as the acoustic units, the acoustic normalization is even better compared to that using the phonetic units [17]. Better acoustic normalization, in turn, results in better estimates of the articulatory features.

Although the effect of different acoustic normalizations in SII has been studied on the quality of the estimated articulatory features [17], it is not clear how the ASR performance would change when the estimated articulators using different acoustic normalizations in SII are used for recognition. In this work, we study the effect of different acoustic normalizations on broad class phonetic recognition accuracy, the recognition being done based on the estimated articulatory features. The goal is to compare the amount of phonetic cues present in the articulatory features obtained using different acoustic normalization techniques. We also compare the recognition accuracies obtained by the articulatory features estimated from SII with those estimated from SDI as well as the directly measured articulatory features.

Broad class phonetic recognition experiment reveals that a better acoustic normalization leads to a better recognition accuracy. This suggests that when the estimated articulatory features match the original ones, they also provide more discrimination among broad phonetic classes. It is also found that, on an average, the recognition accuracy obtained by the articulatory features estimated using the best acoustic normalization technique in SII is better than that using SDI. Interestingly, the recognition accuracy using articulatory features from SII is found to be similar to that using the directly measured articulatory features. All these findings indicate the potential of the articulatory features estimated using SII for phonetic recognition.

We begin with the description of the dataset and the acoustic and articulatory features. In section 3, we briefly describe different acoustic normalization techniques used for comparison in this work. The recognition experiments and results are discussed in section 4. Conclusions and future works are summarized in section 5.

2. Dataset and features

For the recognition experiments and AAI in our work, we have used the Multichannel Articulatory (MOCHA) database [18]. This

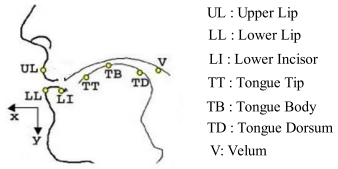

Figure 1: Illustration of EMA sensors' placement

database has one male and one female talker of British English. The dataset comprises of the parallel acoustic and articulatory kinematics recordings using electromagnetic articulography (EMA) corresponding to 460 utterances spoken by each of the subjects. We use 39-dimensional Mel frequency cepstral coefficients (MFCCs) along with their velocity and acceleration coefficients as the acoustic features. These MFCCs are calculated using a 20 ms frame length and a 10 ms frame shift. The articulatory features which we use are 14-dimensional raw EMA features (i.e., X and Y co-ordinates of the upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue body (TB), tongue dorsum (TD), and velum (V)). Along with the articulatory positions, we also use their velocity and acceleration values, hence 42-dimensional articulatory feature vector is used. For illustration, EMA sensors' placement on these articulators is shown in Fig. 1. As the mean position of articulators changes in every utterance, we perform a pre-processing on the EMA data following the steps outlined in [19].

For the SII, we use TIMIT database [20] as GAS. This is because the best SII performance was reported using TIMIT as the GAS [17]. The TIMIT corpus comprises of recordings in a quite environment for ten sentences each spoken by 630 speakers. The corpus contains eight major dialects of American English. As the 'sa1', 'sa2' recordings are for speaker calibration in TIMIT corpus, we have excluded them resulting in 5040 recordings, spanning a total duration of ≈4.29h.

3. Acoustic normalization in subject independent inversion

Let us denote the training data for inversion by $\{(\mathbf{z}_i, \mathbf{x}_i); 1 \leq i \leq T\}$, where i is the frame index, T denotes the duration of the test utterance in number of frames and $\mathbf{z}$ and $\mathbf{x}$ denote the acoustic and articulatory feature vectors respectively. Let the acoustic feature vectors for the test utterance be $\mathbf{u}_n,\ 1 \leq n \leq N$, where N is the total number of frames of the utterance. In SII, following the principle of generalized smoothness criterion (GSC) [21], the articulatory feature vectors $\mathbf{y}_n^\star,\ 1 \leq n \leq N$ for the test utterance are estimated. GSC imposes articulator specific smoothness in inversion.

In GSC the j-th articulatory trajectory $\left\{y_n^{j\star}; 1 \leq n \leq N\right\}$ is estimated using the acoustic features $\mathbf{z}_i$ which are close to $\mathbf{u}_n$ in the Euclidean sense. But in case of the subject-independent (SI) setup, there is a mismatch between the acoustics of the train and test subjects. To overcome this, several normalization techniques are proposed by Afshan et al [17], which uses GAS as outlined by Ghosh et al. [16]. Let the acoustic feature vectors of the GAS be given by the set $\mathcal{A} = \{\mathbf{c}_r; 1 \leq r \leq R\}$. We use K subsets of $\mathcal{A}$, determined either in supervised or unsupervised manner denoted by $\mathcal{A}_k$. Each acoustic subset is then represented by its probability density function (PDF) of the acoustic feature vectors using an M-component Gaussian mixture model (GMM).

The posterior probability feature vector[1] $\Phi(\mathbf{v})$ for an acoustic feature vector $\mathbf{v}$, is defined as follows:

$$\Phi(\mathbf{v}) \triangleq \frac{1}{Z}\left[p(\mathbf{v}|\mathcal{A}_1)\cdots p(\mathbf{v}|\mathcal{A}_K)\right]^{\mathrm{T}}, \quad (1)$$

$$\text{where } Z = \sum_{k=1}^{K} p(\mathbf{v}|\mathcal{A}_k)$$

is a normalization term. 'T' denotes the vector transpose operator. Thus, $\Phi(\mathbf{v})$ is a K dimensional vector representing the likelihood of $\mathbf{v}$ given each of the K subsets in the acoustic space. $\Phi(\mathbf{v})$ is typically a sparse vector with the highest value at the element corresponding to the acoustic subset which gives the maximum likelihood of the acoustic feature $\mathbf{v}$. In SII, the closeness between the test acoustic feature vector $\mathbf{u}_n$ and a training acoustic feature vector $\mathbf{z}_i$ is measured using $\Phi(\mathbf{u}_n)$ and $\Phi(\mathbf{z}_i)$ [17]. By measuring the closeness between $\Phi(\mathbf{u}_n)$ and $\Phi(\mathbf{z}_i)$ for all training acoustic features, the L closest acoustic feature vectors from the training set are obtained and the corresponding articulatory feature vectors are used in GSC for inversion.

To make the process of estimation computationally efficient, subsets $\mathcal{B}_k,\ 1 \leq k \leq K$ in the training corpus are created following the acoustic subsets $\mathcal{A}_k$ in the GAS. Given a test acoustic feature vector $\mathbf{u}_n$, we select the best matching subset $\mathcal{B}_{\hat{k}}$ $\left\{(\mathbf{z}_i^{\hat{k}}, \mathbf{x}_i^{\hat{k}}); 1 \leq i \leq T_{\hat{k}}\right\}$, where $T_{\hat{k}}$ is the number of frames in the $\hat{k}$-th subset of the training corpus. The articulatory feature vectors in $\mathcal{B}_{\hat{k}}$ are used in GSC for inversion (i.e., $L = T_{\hat{k}}$).

Several SII schemes were proposed by Afshan et al.[17]. These schemes typically vary depending on how acoustic subsets $\mathcal{A}_k$ as well as $\mathcal{B}_{\hat{k}}$ are formed. We use three among these schemes for comparison in this work. They are briefly described in the following subsections.

3.1. Inversion scheme - IS1

IS1 is similar to the SII scheme proposed by Ghosh et al. [16]. In IS1, a K-means clustering of the acoustic feature vectors of GAS (i.e., the set $\mathcal{A}$) is performed to obtain the acoustic subsets $\mathcal{A}_k$. The subset $\mathcal{B}_k = \left\{(\mathbf{z}_i^k, \mathbf{x}_i^k); 1 \leq i \leq T_k\right\}$ in the training corpus is determined by the subset of $\mathbf{z}_i$ which yields the highest likelihood given the k-th acoustic cluster $\mathcal{A}_k$ compared to all other clusters.

The estimation of $\mathcal{B}_{\hat{k}}$ given a test acoustic feature vector $\mathbf{u}_n$ is done by finding $\hat{k}$ such that the likelihood of $\mathbf{u}_n$ is maximum given the $\mathcal{A}_{\hat{k}}$ among all acoustic clusters as follows:

$$\hat{k} = \arg \max_{1 \leq k \leq K} p(\mathbf{u}_n|\mathcal{A}_k) \quad (2)$$

Thus, the acoustic clusters in GAS are obtained in an unsupervised way in IS1. The subsets $\mathcal{B}_k$ in the training corpus are formed such that each subset is acoustically similar to one acoustic cluster in GAS. Given the test acoustic feature vector $\mathbf{u}_n$ we first find which acoustic cluster $\hat{k}$ in GAS it is most likely to have come from as shown in (2) and then the corresponding subset $\mathcal{B}_{\hat{k}}$ in the training corpus is used for computing possible articulatory feature values and their probabilities $\left\{\eta_n^l, p_n^l; 1 \leq l \leq L\right\}$ required for GSC .

3.2. Inversion scheme - IS3H

In IS3H, 3-state left-to-right phonetic HMMs are trained using the speech and the corresponding transcripts of the GAS and then a forced-alignment of the utterance is performed with the phonetic transcription of the text. Feature vectors in GAS with identical phonetic state label are used to form acoustic clusters. Since generating $\mathcal{A}_k$ requires transcription along with the speech acoustic

[1] We assume that the prior probabilities of all the subsets $\mathcal{A}_k$, $k = 1, \ldots K$ are equal

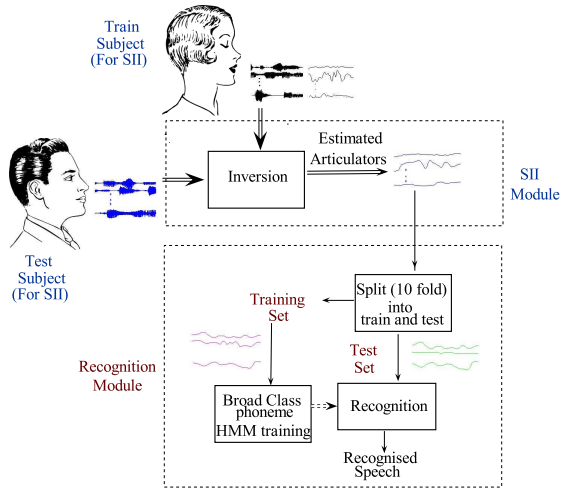

Figure 2: Experimental setup of the speech recognition using articulatory features estimated from inversion

signal, the acoustic subsets in GAS are found to be more representative of the actual sub-phonetic clusters compared to those obtained by unsupervised clustering as in IS1 [17]. This helps in achieving better normalization in SII using GAS. The subset $\mathcal{B}_k$ is obtained by running a Viterbi decoding on each training utterance using the HMMs trained on the GAS and determining the frames corresponding to states of the phonetic HMMs. Similarly, $\mathcal{B}_{\hat{k}}$ is estimated by running a Viterbi decoding on the test utterance using HMMs trained on GAS.

3.3. Inversion scheme - IS3A

IS3A is similar to IS3H except that before computing $\mathcal{B}_k$ and $\mathcal{B}_{\hat{k}}$ the parameters of HMMs trained on GAS are adapted using the training and test acoustics. Due to adaptation, it was found that acoustic normalization in IS3A is better than that in IS3H leading to better SII performance. In fact, IS3A was reported to be the best performing SII scheme by Afshan et al. [17].

4. Experiments and results

4.1. Experimental setup

Fig. 2 shows the block diagram of the experimental setup for the inversion and recognition using articulatory features estimated from inversion. The train subject's acoustic and articulatory data are used to train the inversion system, which is used to estimate articulatory features from test subject's speech signal. In order to perform recognition using the estimated articulatory features, the set of estimated articulatory features are split into training and test sets. Features from the training set are used to build three state left-to-right phonetic HMMs. These trained HMMs are used to perform the phonetic recognition on the estimated articulatory features from the test set.

Note that the inversion block could be based on either SII or SDI schemes. Also the training and the test subjects could be same or different resulting in a subject dependent (SD) or SI setup. When the SDI scheme is used in an SD setup, it is referred to as ISm1. Similarly, when the SDI scheme is used in an SI setup, it is referred to as IS0. The SII schemes used in the SI setup are IS1, IS3H, and IS3A as explained in Section 3. The acoustic clusters in IS1 are obtained with K=39. In IS3H and IS3A, we use 117 (39×3) acoustic clusters. We use 256 mixture components in each acoustic subset GMM for all the SII schemes. In IS3A, the adaptation is done following a supervised adaptation frame-work involving a static two-pass adaptation approach based on MLLR adaptation [22]. We use the entire available test subject's acoustics for the adaptation.

For ISm1 and IS0, we use GMM based AAI [23], where the articulatory features are estimated using the minimum mean squared error (MMSE) criterion. Since ISm1 uses an SDI scheme in SD setup, the inversion performance obtained from ISm1 may indicate an upper bound on the performance from any SII scheme. On the other hand, IS0 represents a baseline – it reflects the performance of the traditional SDI scheme when the training and test subjects are not matched.

Since there are only 460 utterance recordings for each subject, the number of frames for several fine class phonemes turns out to be small, which is insufficient for building a good statistical model for each phoneme required for a fine class phonetic recognition. Hence, following the work by Sainath et al [24], we combine the fine class phonemes into four broad classes, namely VOWELS, STOPS, FRICATIVES, NASALS as shown in Table 1. To obtain the ground truth broad class phonetic boundaries, we perform a fine class phonetic forced-alignment (using 39 phoneme set [25]) using HMM Tool Kit (HTK) [26] and the available transcriptions of the utterances spoken by the MOCHA-TIMIT subjects. The fine class phonetic boundaries thus obtained are manually checked and corrected if required. Thereafter, the broad class phonetic boundaries are obtained from the fine class phonetic boundaries and used for recognition.

FINE CLASS PHONEMES	BROAD PHONEMES CLASS
AA,AE,AH,AO,AW,AY, EH,ER,EY,IH,IY,JH,L, OW,OY,R,UH,UW,W,Y	VOWELS
B,CH,D,G,K,P,T	STOPS
DH,F,HH,S,SH,TH,V,Z,ZH	FRICATIVES
M,N,NG	NASALS

Table 1: Mapping of the fine class to broad class phonemes

The recognition using the estimated articulators is also performed using HTK. A three-state left-to-right HMM with state emission PDF as GMMs with 256 mixtures is used for the recognition. Note that no language model is used in the recognition. We also perform recognition using acoustic features i.e., MFCCs as well as directly measured articulatory features. These are referred to as MFCC and EMA respectively. Experiment using EMA is done to understand the performance in recognition using estimated articulators in relation to that obtained using the original articulatory kinematics.

Both inversion as well as recognition are done in a 10-fold cross validation setup. In the SD setup, 8 folds are used for training the SDI scheme, 1 fold is used as the development set and the remaining fold as the test set and this is repeated 10 times. This is repeated separately for two subjects in the MOCHA-TIMIT database. In the SI setup, the articulators for all the utterances of a test subject are estimated using 8-folds and 2-folds of the other subject as the training and development set. Once the articulatory features are estimated for all folds, they are used for recognition, where 9-folds are used for training the recognizer and the remaining 1-fold is used as the test set in a round-robin fashion.

4.2. Results

In this work, we report the inversion performance in addition to the recognition performance. The inversion performance is re-

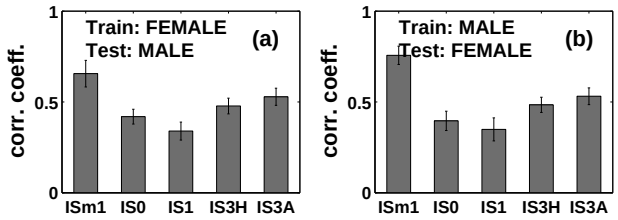

Figure 3: AAI performances (in terms of average ($\pm$ one SD) correlation coefficient) averaged over different articulatory features for various inversion schemes: (a) Training and test being the MOCHA female and male speakers, respectively, (b) Training and test data being the MOCHA male and female speakers, respectively. For ISm1, the training subject is chosen identical to the test subject.

ported in terms of the Pearson's correlation coefficient [27] between the original articulatory trajectory and the estimated one. Thus a higher correlation coefficient would indicate more similarity between the original and the estimated articulatory trajectories and hence better inversion performance. Fig 3 shows the inversion performance for various schemes considered. Top three inversion schemes in the increasing order of their inversion performance are IS3H, IS3A and ISm1. Due to the matched training and test subjects, ISm1 achieves the highest performance which is followed by IS3A, the best among all the SII schemes. Acoustic normalization using GAS as well as adaptation of HMM parameters is the key to the high performance of IS3A. However, it should be noted that the correlation coefficient obtained using IS3A is lower (by 0.13 and 0.23 absolute) than that using ISm1 for the male and the female test subjects respectively. Thus there is still scope for improvement in the SII performance.

Table 2 shows the recognition accuracies averaged across all folds using different acoustic and articulatory features for both male and female speakers. It is clear that the acoustic features result in a better recognition accuracy compared to that using articulatory (original and estimated) features. Among the articulatory features, the four highest average recognition accuracies for both male and female subjects are obtained using EMA, IS3A, IS3H and ISm1. It should be noted that the statistical test (t-test) reveals that there is no significant difference ($p \geq 0.05$) between the recognition accuracies obtained using EMA and IS3A for both subjects. However, the recognition accuracy using IS3A is significantly ($p < 0.01$) better than that using ISm1 in case of the male subject, while that is not true for the female subject. The recognition accuracy using IS3A is significantly ($p < 0.01$) better than that using IS3H for both subjects. Thus IS3A yields not only the best inversion performance but also the best recognition performance among different SII schemes. Moreover, the recognition accuracy using IS3A is similar to that obtained by the original articulatory features as well as the articulatory features estimated from the SDI scheme. It is interesting to note that although ISm1 has a higher inversion performance compared to the best performing SII scheme, i.e., IS3A, there is no significant difference in their recognition accuracies.

Detailed investigation of the confusion matrices of phonetic recognition using different acoustic and articulatory features show that the FRICATIVES recognition accuracy using IS3A is higher than that using ISm1 and is comparable to that using EMA. This is not the case for STOPS and VOWELS. In the case of NASALS, the recognition accuracy using IS3A is higher than that using EMA and ISm1 for the female subject, while they are similar for the male subject. However, for each of the broad class phoneme,

Subject	Recognition accuracy (in percent) using						
	MFCC	EMA	ISm1	IS0	IS1	IS3H	IS3A
Female	86.08 (1.31)	74.24 (3.01)	72.24 (1.33)	65.76 (1.86)	67.14 (2.34)	69.83 (1.61)	73.17 (2.09)
Male	84.48 (1.38)	70.97 (1.2)	66.47 (1.58)	63.5 (1.33)	62.05 (2.66)	66.76 (1.31)	69.68 (1.62)

Table 2: Average recognition accuracy in the MOCHA-TIMIT corpus (with standard deviation shown in braces).

the recognition accuracy using the acoustic features is higher than that using articulatory features.

4.3. Discussion

In IS3H and IS3A, the best matching subset $\mathcal{B}_{\hat{k}}$ for a frame of a given sentence is computed using the acoustic features of all the remaining frames in the sentence leading to a better estimate of the subset. In addition, $\mathcal{B}_k$ corresponds to finer phonetic segments and the estimated articulatory feature in a frame is computed as a weighted combination of the articulatory features from $\mathcal{B}_{\hat{k}}$ which in turn could encode the phonetic information in the estimated articulatory features resulting in the best SII performance. In the SDI using GMM based mapping (ISm1), no such phonetic information is directly embedded in the estimated articulatory features. This could result in a better average recognition accuracy in the case of IS3A compared to that of ISm1, although IS3A based estimates are worse compared to that of ISm1 based estimated articulatory features when compared against the measured articulatory features (EMA). Similar recognition accuracies using IS3A and EMA suggest that statistically both of them provide equal discrimination among different phoneme classes although the estimated and original articulatory trajectories are dissimilar. This could also be a direct consequence of selecting the relevant articulatory features in each test frame based on finer phonetic class identity. It should be noted that the recognition accuracy using IS2A is found to be worse than that of IS3A indicating that finer phonetic subsets in GAS is beneficial for both SII as well as recognition performance compared to typical phonetic subsets. Thus, finer the clusters in GAS, better is the recognition accuracy using corresponding estimated articulatory features.

5. Conclusions

In this work, we have found that a better acoustic normalization and adaptation in SII not only improve inversion performance but also improve recognition accuracy when the articulatory features from SII are used for recognition. We also find that the recognition performance of the best SII scheme is similar to that using the SDI scheme as well as the original articulatory features. However, the inversion performance of SDI is better than that using the best SII scheme. This indicates that a worse inversion performance may not necessarily result in a worse performance in recognition using the articulatory features obtained from inversion. While it is well-known that the original articulatory features provide complementary information to the acoustic features [3], investigation is required to find out if the estimated articulatory features are also complementary to the acoustics for recognition. This benefit due to articulatory features could be more evident when a better modeling technique such as the deep neural network (DNN) is used. It would also be interesting to examine the variations in the recognition benefits for different fine class phonemes by using a corpus larger than the MOCHA database used in this work. These are parts of our future work.

6. References

[1] E. McDermott and A. Nakamura, "Production-oriented models for speech recognition," *IEICE Trans. Inf. & Syst.*, vol. E89-D, no. 3, pp. 1006–1014, 2006.

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

[3] A. A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," *Proc. ICSLP, Beijing, China*, pp. 145–148, 2000.

[4] A. Toutios and K. Margaritis, "A rough guide to the acoustic-to-articulatory inversion of speech," *Proceedings of HERCMA*, pp. 746–753, September 2003.

[5] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *J. Acoust. Soc. Am*, vol. 63, pp. 1535–1555, May 1978.

[6] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 444–460, 2005.

[7] Y. Laprie and B. Mathieu, "A variational approach for estimating vocal tract shapes from the speech signal," *Proc. ICASSP*, pp. 929–932, 1998.

[8] K. Kirchhoff, *Robust Speech Recognition using Articulatory Information*. PhD. Thesis, University of Bielefeld, 1999.

[9] S. Krstulovic, *Speech analysis with production constraints*. PhD Thesis, Ecole Polytechnique Federale de Lausanne, 2001.

[10] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamic and phonological constraint," *the 5th Speech Production Seminar, Munich, Germany*, pp. 237–240, 2000.

[11] A. Lammert, D. P. W. Ellis, and P. Divenyi, "Data-driven articulatory inversion incorporating articulator priors," *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition, SAPA, Brisbane, Australia*, 21 September 2008.

[12] T. Toda, A. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with Gaussian mixture model," *Proc. ICSLP*, pp. 1129–1132, October 4-8 2004.

[13] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," *Proc. ICSLP*, pp. 577–580, September 2006.

[14] L. Zhang and S. Renals, "Acoustic-articulatory modeling with the trajectory HMM," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[15] S. Hiroya and T. Mochida, "Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMMs," *Speech Communication*, vol. 48, no. 12, pp. 1677–1690, 2006.

[16] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," *ICASSP*, pp. 4624–4627, 2011.

[17] A. Afshan and P. K. Ghosh, "Improved subject-independent acoustic-to-articulatory inversion," *Speech Communication*, vol. 66, pp. 1–16, 2015.

[18] 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.

[19] K. Richmond, *Estimating articulatory parameters from the acoustic speech signal*. Ph.D. Thesis, The Centre for Speech Technology Research, Edinburgh University, 2002.

[20] DARPA-TIMIT, *Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1*, 1990.

[21] 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.

[22] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[23] T. Toda, A. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, pp. 215–217, 2008.

[24] T. Sainath, D. Kanevsky, and B. Ramabhadran, "Broad phonetic class recognition in a hidden Markov model framework using extended baum-welch transformations," in *ASRU*. IEEE, 2007, pp. 306–311.

[25] K. F. Lee and H. W. Won, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[26] S. J. Young, "The HTK hidden Markov model toolkit: Design and philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[27] K. Pearson, "Notes on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, no. 347-352, pp. 240–242, 1895.