Improved subject-independent acoustic-to-articulatory inversion

Amber Afshan, Prasanta Kumar Ghosh

Abstract

In subject-independent acoustic-to-articulatory inversion, the articulatory kinematics of a test subject are estimated assuming that the training corpus does not include data from the test subject. The training corpus in subject-independent inversion (SII) is formed with acoustic and articulatory kinematics data and the acoustic mismatch between training and test subjects is then estimated by an acoustic normalization using acoustic data drawn from a large pool of speakers called generic acoustic space (GAS). In this work, we focus on improving the SII performance through better acoustic normalization and adaptation. We propose unsupervised and several supervised ways of clustering GAS for acoustic normalization. We perform an adaptation of acoustic models of GAS using the acoustic data of the training and test subjects in SII. It is found that SII performance significantly improves (~25% relative on average) over the subject-dependent inversion when the acoustic clusters in GAS correspond to phonetic units (or states of 3-state phonetic HMMs) and when the acoustic model built on GAS is adapted to training and test subjects while optimizing the inversion criterion.

Keywords: Acoustic-to-articulatory inversion; Subject-independence; Generic acoustic space; Adaptation

1. Introduction

Acoustic-to-articulatory inversion is the task of estimating speech articulatory representations from acoustic signal representations. While there are several spectro-temporal features or model parameters derivable from the acoustic speech signal, there are also a variety of articulatory representations obtainable including through (1) stylized models such as Maeda’s model (Maeda, 1979, 1990) or lossless tube model (Kelly and Lochbaum, 1962) of the vocal tract, (2) linguistic rule based models (Browman and Goldstein, 1989, 1986) or (3) direct physiological data-based representations of articulatory information (Wrench and William, 2000). In this work, we consider 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 instrumental means (e.g. electromagnetic articulograph (EMA)).

Acoustic-to-articulatory inversion can be subject-dependent or subject-independent depending on both the training and test subjects. In subject-dependent inversion (SDI), training data from the test subject are assumed to be available. On the other hand, in subject-independent inversion (SII), the test subjects could be, in general, different from the training subject pool; the acoustic mismatch between the training and test subjects makes SII more challenging than SDI. SII is useful for deriving articulatory features for supporting speech or speaker recognition (Ghosh and Narayanan, 2011a; Li et al., 2013) where articulatory data are not directly available from the test subject. Since direct articulatory evidence is scarce (limited amounts of data available for a small number of talkers),
SII plays a critical role for estimating articulatory information because it is robust to insufficient or non-availability of training data of the target test speaker.

Most existing techniques for acoustic-to-articulatory inversion belong to the SDI category; from a comprehensive summary (Toutios and Margaritis, 2003) of these techniques, it can be seen that the SDI techniques can be broadly divided into three categories: (1) Codebook lookup procedure (Ouni and Laprie, 2005), where the codebook is built from the training data, (2) analytical approaches involving an articulatory model such as Maeda’s (Laprie and Mathieu, 1998; Krstulovic, 2001) and (3) statistical (both parametric and non-parametric) modeling of the acoustic-articulatory map, including using a Gaussian mixture model (GMM) (Toda et al., 2004), mixture density network (MDN) (Richmond, 2006), a trajectory hidden-Markov model (HMM) (L. Zhang and Renals, 2008), generalized smoothness criterion (GSC) (Ghosh and Narayanan, 2010) and neural network-based modeling of the acoustic-to-articulatory mapping (Kirchhoff, 1999; King and Taylor, 2000). It has been shown across various studies that smoothing the articulatory trajectories improves the inversion performance. This is mainly because the articulatory trajectories have a lowpass spectrum where 99% of the spectral energy lies below 25 Hz for all articulators (Ghosh and Narayanan, 2010). Smoothing has been shown to be effective either as a post-processing step (Toda et al., 2004) or integrated directly within the objective function for inversion (Toda et al., 2004; Ghosh and Narayanan, 2010, 2013).

In contrast to SDI, there are fewer existing techniques for SII. One of the SII techniques (Ghosh and Narayanan, 2011b) works on the principle of representing an acoustic feature with respect to a generic acoustic space, obtained using speech data from a pool of talkers. When a test subject’s acoustic data is given for inversion, it is matched with the training subject’s acoustic data with respect to the generic acoustic space. Thus, the generic acoustic space is used to normalize the acoustic mismatch between the training and test subjects. It was shown that SII yields a better inversion performance than SDI techniques when the training and test subjects are different. However, the SII performance was found to be lower than the performance obtained under matched training and test conditions.

In this work, we focus on improving the SII performance by using better normalization techniques as well as various generic acoustic space designs. We experiment with different acoustic models and different numbers of speakers, as well as with different styles of talking – read and spontaneous. It is found that SII performance improves in the case of a generic acoustic space with more speakers and speaking styles matched to the subjects of the articulatory corpus. We have also investigated various means of normalization to compensate for the acoustic mismatch between the training and test subjects. In an earlier work (Ghosh and Narayanan, 2011b), clustering of the generic acoustic space was performed, and acoustic features of the training and test subjects were converted to probability features whose dimensions are equal to the number of clusters, where each element in the probability feature is the likelihood of an acoustic feature given the corresponding cluster. In addition to unsupervised clustering, in this work, we investigate supervised clustering methods based on the phonetic and sub-phonetic identities of the acoustic space obtained through a hidden Markov model (HMM). We find that supervised clustering leads to a better SII performance. We obtain an even better SII performance when the HMM parameters are adapted to the training and test subjects under consideration for inversion.

Speaker-independent articulatory HMMs were presented by Hiroya et al. for speaker adaptive inversion mapping (Hiroya and Mochida, 2006). Training articulatory HMMs requires EMA data from multiple subjects rather than only one in the present work. Given that EMA recording is expensive (Steiner et al., 2012), the requirement of minimal training data is one of the key advantages of the present formulation. The speaker adaptation (Hiroya and Mochida, 2006) uses a linear mapping between the source and target speakers’ articulatory features. However, it is well-known that vocal tract morphology varies across subjects; unlike cepstral features (Pitz and Ney, 2005), there is no evidence that the relationship should be linear. The proposed approach to SII in this work does not suffer from such limitations.

We begin by describing the articulatory datasets, and the different acoustic and articulatory features used in this work. In Section 3, we describe the SII criterion along with different inversion schemes and their solutions corresponding to different types of clustering and adaptation involved in modeling the generic acoustic space. In Section 4, we describe the experimental setup and discuss the results of acoustic-to-articulatory inversion using different articulatory features. In Section 5, we offer our conclusions.

2. Dataset and features

For the analysis and acoustic-to-articulatory inversion experiments of this paper, we use the Multichannel Articulatory (MOCHA) database (Wrench and William, 2000) that contains acoustic and corresponding electromagnetic articulography (EMA) data from one male and one female talker of British English. The articulatory position data have high frequency noise resulting from EMA measurement error. In addition, the mean position of the articulators changes from utterance to utterance (Richmond, 2002); hence, the position data needs pre-processing before it can be used for analysis. Following the available pre-processing steps (Richmond, 2002), we obtain parallel acoustic and articulatory data at a frame rate of 100 observations per second. For subject-independent inversion (SII) we use one subject’s data for training the inversion technique and the other subject’s data for testing.
We repeat this separately for both subjects. Of the 460 utterances available from the training subject, data from 368 utterances (80%) are used for training, 92 utterances (20%) as the development set (dev set); and all the utterances of the test subject as the test set for the SII.

We use 39-dimensional Mel frequency cepstral coefficients (MFCCs) with velocity and acceleration features as the acoustic feature. MFCCs are calculated using a 20 ms frame length and a 10 ms frame shift (Young, 1994). We experiment using two types of articulatory feature vectors similar to those used by Ghosh and Narayanan (2011b). First, we use 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)). We also use tract variable (TV) features motivated by articulatory phonology (Brownman and Goldstein, 1986, 1989). These features conceptualize speech as being produced as an ensemble of articulatory gestures. We use 6-dimensional TV features as outlined by Ghosh and Narayanan (2011b) – lip aperture (LA), lip protrusion (PRO), jaw opening (JAW_OPEN), tongue tip constriction degree (TTCD), tongue body constriction degree (TBCD), velum opening (VEL).

3. Subject-independent acoustic-to-articulatory inversion

We have proposed five different SII schemes depending on various types of acoustic normalization techniques between the training and test subjects. We describe the general framework for SII, followed by details of the proposed schemes along with the technique to solve the SII criterion.

3.1. Inversion criterion

Let the acoustic and articulatory feature vectors of the training subject in SII be denoted by $z$ and $x$ respectively.

$$x = [x^1 x^2 \ldots x^J],$$

where $J$ is the number of articulatory features and $x^j$ is the $j$th articulatory feature. Thus, the training data for SII are denoted by \{
\{(z_i, x_i); 1 \leq i \leq T\}\}, where $T$ is the total number of frames in the training corpus and $i$ denotes the frame index. Suppose the acoustic feature sequence for the test utterance is denoted by $u_n$, $1 \leq n \leq N$, where $N$ is the duration of the test utterance in number of frames. The articulatory feature vectors $y_n^*$, $1 \leq n \leq N$ for the test utterance are obtained following the principle of generalized smoothness criterion (GSC) (Ghosh and Narayanan, 2010) which imposes articulator specific smoothness in articulatory inversion. Using GSC, the $j$th articulatory trajectory $\{y_n^*; 1 \leq n \leq N\}$ is estimated as follows:

$$\{y_n^*; 1 \leq n \leq N\} = \arg \min_{\{y^j\}} J(y^1, \ldots, y^J$$

$$\triangleq \arg \min_{\{y^j\}} \left\{ \sum_{n=1}^{N} (y_n^j - h_n^j)^2 + \sum_{n=1}^{N} \sum_{l=1}^{L} (y_n^j - \eta_n^j)^2 \beta_n^j \right\},$$

where $J$ is the objective function in GSC-based optimization, which comprises two terms with $C$ as the trade-off parameter between the two. $*$ denotes the convolution operator (i.e. $y_n^j * h_n^j = \sum_{k=1}^{N} y_{n-k}^j h_{n-k}^j$), where $h_n^j$ is a high-pass filter (FIR or IIR) with cut-off frequency $f_c^j$ specific to $j$th articulator. The first term in $J$ represents the energy of the output of the high-pass filter; minimization of this energy ensures smoothness in the estimated articulatory trajectories. \{$\eta_n^j; 1 \leq i \leq L\} is the set of $L$ possible values of the $j$th articulatory feature at the $n$th frame. $\eta_n^j$ denotes the probability that $\eta_n^j$ is the value of the articulatory features at the $n$th frame given that $u_n$ is the test acoustic feature.

The second term in $J$ ensures that the estimated articulatory trajectory passes through a highly probable articulatory feature value. $\eta_n^j = [\eta_n^{j,1}, \eta_n^{j,2}, \ldots, \eta_n^{j,d}]$ and $p_n^j$ are determined using $u_n$, the training data \{(z_i, x_i); 1 \leq i \leq T\} and a generic acoustic space (GAS) as described below.

In SDI using the GSC (Ghosh and Narayanan, 2010), $\eta_n^j$ and $p_n^j$ are obtained by finding acoustic features $z$ from the training data which are close (in the Euclidean sense) to the test acoustic features. $p_n^j$ is calculated by computing the inverse of the Euclidean distance between the test acoustic feature and $z$; the corresponding articulatory features $x$ are used as $\eta_n^j$. However, for SII, the training and test acoustic features come from different subjects and thus the Euclidean distance between the training and the test features may not be appropriate as a closeness measure.

To circumvent the mismatch between training and test acoustics, we follow a normalization technique using a GAS as outlined by Ghosh and Narayanan (2011b). The GAS consists of a large pool of acoustic-only data from different talkers. Note that the acoustic data used for GAS is different from that from the training subject in SII. Let the acoustic feature vectors of the GAS be given by the set $A = \{a_r; 1 \leq r \leq R\}$. We consider $K$ many subsets of $A$ denoted by $A_k$ where $\bigcup_{k=1}^{K} A_k = A$. The subsets could be determined in a supervised or unsupervised way. The probability density function (PDF) of the acoustic feature vectors in each subset is modeled using an $M$-mixture Gaussian Mixture Model (GMM) as follows:

$$p(v|A_k) = \sum_{m=1}^{M} \pi_m \gamma (v; \mu_m^k, \Sigma_m^k),$$

where $v$ is the acoustic feature vector, $\mu_m^k$ and $\Sigma_m^k$ are the mean vector and the covariance matrix of the $m$th component of GMM in the $k$th subset, respectively. $\mu_m^k$ and $\Sigma_m^k$ are estimated using the expectation maximization algorithm (Dempster et al., 1977). $w_m$ is the weight for the $m$th component. Given an acoustic feature vector $v$, the posterior probability feature vector $^\Phi(v)$ is defined as follows:

$$^\Phi(v) = \arg \max_{k=1, \ldots, K} p(v|A_k)$$

We assume that the prior probabilities of all the subsets $A_k, k = 1, \ldots, K$ are equal.
\[ \Phi(v) \triangleq \frac{1}{Z} [p(v|A_1) \cdots p(v|A_K)]^T, \]

where \( Z = \sum_{k=1}^K p(v|A_k) \)

is a normalization term. ‘\( T \)' denotes the vector transpose operator. Thus, \( \Phi(v) \) is a \( K \) dimensional vector representing the likelihood of \( v \) given each of the \( K \) subsets in the acoustic space. \( \Phi(v) \) 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 \( v \).

The closeness between the test acoustic feature vector \( u \) and a training acoustic feature vector \( z \) is measured using \( \Phi(u) \) and \( \Phi(z) \) in SII. We hypothesize that the Euclidean distance between \( \Phi(u) \) and \( \Phi(z) \) would be a better estimate of the acoustic difference between \( u \) and \( z \), compared to measuring the Euclidean distance between \( u \) and \( z \) directly. If \( u \) and \( z \) are representations of the same acoustic unit, then the likelihood of both \( u \) and \( z \) will be higher given the GMM of the respective acoustic subset and lower for other acoustic subsets. This will cause the \( \Phi(u) \) and \( \Phi(z) \) vectors to have the highest valued element at the identical location and, hence, the Euclidean distance between the two would be small. Similarly, if \( u \) and \( z \) are representations of two different acoustic units, the \( \Phi(u) \) and \( \Phi(z) \) vectors will have the highest valued elements at two different locations resulting in a higher Euclidean distance between the two. By measuring the closeness between \( \Phi(u) \) and \( \Phi(z) \) for all training acoustic features, the \( L \) closest acoustic feature vectors from the training set are obtained, which are finally used to determine \( \{\eta_i^u, p_i^u; 1 \leq i \leq L \} \). However, finding the \( L \) closest acoustic feature vectors from the training set involves computing the Euclidean distances between \( \Phi(u) \) and \( \Phi(z) \), \( 1 \leq i \leq T \) at each test frame; this could be time-consuming and computationally expensive for a large \( T \) (i.e. large training corpus).\footnote{For example, the number of feature vectors \( T \) for each subset in the MOCHA corpus is \( \sim 10^6 \). Given a test utterance of duration 3 s (i.e. 300 frames), one needs to make \( 3 \times 10^7 \) Euclidean distance calculations.}

In particular, for each test frame, Euclidean distance computation requires \( KT \) multiplications and \( (K - 1)T \) additions.

We propose a computationally efficient way of determining \( \{\eta_i^u, p_i^u; 1 \leq i \leq L \} \) by creating subsets \( B_k \), \( 1 \leq k \leq K \) in the training corpus following the acoustic subsets \( A_k \) in the GAS and then estimating the subset \( B_k \) in the training corpus which best matches the given test acoustic feature vector \( u \). Let the training data in \( B_k \) be denoted by \( \{z, x\}; 1 \leq i \leq T_k \} \), where \( T_k \) is the number of frames in the \( k \)th subset of the training corpus. Note that \( \sum_{k=1}^K T_k = T \). Thus, \( \{\eta_i^u; 1 \leq i \leq L \} \) comprises all articulatory feature vectors in \( B_k \) with \( p_i^u \) being inversely proportional to the distance between \( \Phi(u) \) and \( \Phi(z^i) \) as follows:

\[ \eta_i^u = x_i^k, \ 1 \leq i \leq L (L = T_k) \]

\[ p_i^u = \frac{D^{-1}(\Phi(u), \Phi(z^i))}{\sum_{i=1}^L D^{-1}(\Phi(u), \Phi(z^i))} \]

where,

\[ D(\Phi(u), \Phi(z^i)) = \left| \max_{1 \leq i \leq K} p(u|A_i) - \max_{1 \leq i \leq K} p(z^i|A_i) \right| \]

Note that \( \max_{1 \leq i \leq K} p(u|A_i) \) is the highest valued element in the \( \Phi(u) \) vector; similarly, \( \max_{1 \leq i \leq K} p(z^i|A_i) \) in \( \Phi(z^i) \). Thus the distance between \( \Phi(u) \) and \( \Phi(z^i) \) is computed using the absolute difference between their highest values. Note that \( k \) is determined such that both \( \Phi(u) \) and \( \Phi(z^i) \) are likely to have their highest values at an identical location. Since the probability feature vector is typically sparse, we assume that the difference between the highest elements of two such vectors will be a good approximation to the Euclidean distance between the two.\footnote{In fact, we have found this approximated distance is correlated with the actual Euclidean distance. We have considered the probability features of the female talker in the MOCHA corpus and considered \( 9.9 \times 10^7 \) pairs of probability features belonging to the same subset. We found that the correlation coefficient between the approximate distance (as in (5)) and Euclidean distance is 0.83. We have also found (Section 4.2.6) that the inversion performance marginally drops by using the approximate Euclidean distance as opposed to using the actual Euclidean distance.}

Computing the difference between two scalar values is faster than computing the Euclidean distance between two \( K \) dimensional vectors. This accelerates the computation of \( \{\eta_i^u, p_i^u; 1 \leq i \leq L \} \). In particular, in each test frame, it requires only \( T_k \) additions; assuming acoustic subsets \( A_k \) of equal size, \( T_k \approx \frac{1}{K}T \) additions are required, which is significantly less than \( KT \) multiplications and \( (K - 1)T \) additions necessary for computing Euclidean distance. By computing the \( p_i^u \) as the inverse of the distance between the training and the test acoustic vector, we give higher probabilities to the possible articulatory features from the training corpus whose corresponding acoustic features are closer to the test acoustic feature vector.

The steps in SII are illustrated in Fig. 1. As shown in the figure, the acoustic feature vectors \( (z) \) in the parallel acoustic-articulatory features of the train subset are converted to the probability feature vector \( \Phi(z) \) using GAS (A). Similarly, the acoustic features \( u \) of the test subset are also converted to \( \Phi(u) \). \( \Phi(z) \), \( \Phi(u) \) and \( A \) are then used to compute \( \eta_i^u \) and \( p_i^u \) using (5). \( \eta_i^u \) and \( p_i^u \) are used finally to compute \( y^s \) using the GSC.

### 3.2. Different inversion schemes

The acoustic subsets \( A_k \), \( 1 \leq k \leq K \) (hence, subsets \( B_k \)) could be determined in a number of ways. Similarly, the estimation of \( B_k \) could also be done in several ways.
Depending on how \(A_k\), \(B_k\), and \(B^\_k\) are determined, one can have different inversion schemes.\(^4\) Below, we describe five different inversion schemes proposed in this work.

### 3.2.1. Inversion scheme – IS1

The acoustic subsets \(A_k\) in IS1 are determined by performing a K-means clustering of the acoustic feature vectors of GAS (i.e. the set \(A\)). The subset \(B_k = \{ (x_i^k, x_l^k); 1 \leq i \leq T_k \}\) in the training corpus is determined by the subset of \(x_i\) which yields the highest likelihood given the \(k\)th acoustic cluster \(A_k\) compared to all other clusters. In other words

\[
B_k = \{ (x_i^k, x_l^k); 1 \leq i \leq T_k \} = \arg \max_{1 \leq x_i \leq k} p(z_i | A_k) \quad (6)
\]

The estimation of \(B^\_k\) given a test acoustic feature vector \(u_n\) is done by finding \(k\) such that the likelihood of \(u_n\) is maximum given the \(A_k\) among all acoustic clusters as follows:

\[
\hat{k} = \arg \max_{1 \leq k \leq K} p(u_n | A_k) \quad (7)
\]

Thus, IS1 obtains acoustic clusters in GAS in an unsupervised way. The subsets \(B_k\) 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 \(u_n\), we first find which acoustic cluster \(k\) in GAS it is most likely to have come from as shown in \(7\) and then the corresponding subset \(B_k\) in the training corpus is used for computing \(\{\eta^u_{i}, p^u_{i}; 1 \leq l \leq L\}\). This is similar to the SII scheme proposed by Ghosh and Narayanan (2011b).

### 3.2.2. Inversion scheme – IS2H

Acoustic clusters of GAS in IS2H are determined by the phonemic identity of each feature vector. All acoustic feature vectors in GAS corresponding to one phoneme are taken as one acoustic cluster. Repeating this for \(K\) different phonemes yields \(K\) different acoustic clusters. The phonemic boundaries in each utterance of GAS are obtained by training a phonetic hidden Markov model (HMM) using the speech and the transcripts of all sentences in GAS and then performing a forced-alignment of the utterance with the phonetic transcription of the text. Since generating \(A_k\) requires transcription along with the speech acoustic signal, the acoustic subsets in GAS would be more representative of the actual phonetic clusters compared to those obtained by unsupervised clustering as in IS1. This could help in achieving better normalization in SII using GAS.

The subsets \(B_k = \{ (x_i^k, x_l^k); 1 \leq i \leq T_k \}\) are obtained by running a Viterbi decoding on each utterance in the training corpus using the HMM trained on the GAS and determining the phonemic boundaries. All acoustic and articulatory feature vectors belonging to the \(k\)th phoneme are used to construct \(B_k\). It is important to note that all the frames in an utterance are used to determine the phonemic category of an individual frame in that utterance; this is because the Viterbi algorithm incorporates the temporal context information while decoding.

\(B^\_k\) for a test acoustic feature vector \(u_n\) at the \(n\)th frame is estimated by performing a phone recognition on the utterance containing \(u_n\). The HMM trained on the GAS is used for this purpose. If the phone recognizer yields the \(k\)th phoneme as the label for the \(n\)th frame, \(B^\_k\) is used for computing \(\{\eta^u_{i}, p^u_{i}; 1 \leq l \leq L\}\). Thus, the estimation of \(\eta^u_{n}, p^u_{n}\) at the \(n\)th frame in IS2 requires all acoustic frames \(u_n, 1 \leq n \leq N\) of the test utterance unlike only \(u_n\) in the case of IS1. This is due to the temporal decoding nature of the Viterbi algorithm used for phone recognition.

### 3.2.3. Inversion scheme – IS3H

IS3H is similar to IS2H except for the identity of the acoustic clusters in GAS. In IS3H, a 3-state left-to-right HMM is trained using the speech and the corresponding transcripts of the GAS. All frames belonging to a

\(^4\) MATLAB implementations of different SII schemes can be found at [http://www.ee.iisc.ernet.in/new/people/faculty/prasantg/softwares.html#SII_IS2_IS3](http://www.ee.iisc.ernet.in/new/people/faculty/prasantg/softwares.html#SII_IS2_IS3).
particular state of a phonetic HMM are used to build acoustic clusters. Since the total number of states is three times the number of phonemes, the number \((K)\) of acoustic clusters in GAS for IS3H is three times the number of clusters for IS2H. Since, in a 3-state HMM, the states model the transient parts in the beginning and end of a phoneme and the stationary part in the middle of the phoneme, we expect to have finer acoustic clusters by using frames corresponding to the states of phonetic HMMs, unlike phonetic identity-based acoustic clusters. Finer acoustic clusters could potentially improve the normalization technique for resolving the acoustic mismatch between the train and test data in SII.

In IS3H, \(B_i\) in the training set and \(B_i\) for the test utterances are computed exactly following the steps outlined in IS2H except that the states of phonetic HMMs are used in place of phoneme labels.

### 3.2.4. Inversion scheme – IS2A

In IS2A, the acoustic clusters in GAS are created based on the phonetic identity of each frame as outlined in IS2H. However, in IS2H, generating \(B_i\) and \(B_i\) requires running a phone recognizer on the training and test subjects’ acoustic features using HMM trained on GAS. Since GAS does not contain acoustic features from the training or test corpora, the accuracy of recognition and hence, the generation of \(B_i\) and \(B_i\) could be improved by adapting the HMM parameters (prior to recognition) using the training and test acoustic features, respectively. In IS2A, the HMM parameters are adapted prior to computing \(B_i\) and \(B_i\). This could further improve the computation of \(\{\eta_i, p_i^l; 1 \leq l \leq L\}\) and, hence, the quality of SII.

### 3.2.5. Inversion scheme – IS3A

In IS3A, we explore the potential advantages of finer acoustic clusters (using states of phonetic HMMs as in IS3H) as well as HMM parameter adaptation (for computing \(B_i\) and \(B_i\) as in IS2A).

Note that in both IS2A and IS3A, computing \(B_i\) requires the availability of the entire training set. Similarly, \(B_i\) requires the availability of the entire test set. This is unlike IS2H and IS3H where access to only one utterance is sufficient as opposed to the entire training or test set. On the other hand, IS1 requires only one frame’s acoustic feature vector for the same task. Thus, different inversion schemes have different requirements for the amount of acoustic data for computing \(\{\eta_i^l, p_i^l; 1 \leq l \leq L\}\) – the least in IS1 and the most in IS2A or IS3A.

### 3.3. Solution to the inversion criterion

The various inversion schemes discussed above are different according to the way they determine \(\{\eta_i^l, p_i^l; 1 \leq l \leq L\}\) at the \(n\)th frame of the test utterance. These are used in the objective function in (1) to solve for the estimated articulatory feature vector \(\hat{y}_n^*\). \(1 \leq n \leq N\). Below, we describe the technique to solve the optimization problem in (1).

Ghosh and Narayanan (2013) showed that the solution of the optimization with an objective function as in (1) asymptotically matches a solution where \(\{\eta_i^l, p_i^l; 1 \leq l \leq L\}\) are at first used to compute the articulatory estimate at each frame separately. Then the smoothing (or low-pass filtering) is performed as a post-processing step on the estimated articulatory feature sequence. They showed that the asymptotic limit holds for sentences 1–2 s long, such as in the MOCHA corpus, which is, in fact, used in this work. Therefore, we use the two-step process of estimation and smoothing to solve the optimization problem in (1). First, the articulatory feature at the \(n\)th frame of the test utterance is estimated as follows:

\[
\hat{y}_n = \sum_{l=1}^{L} \eta_i^l p_i^l \]

where \(\hat{y}_n\) is a \(J\) dimensional articulatory feature vector – \(\hat{y}_n = [\hat{y}_1^l \hat{y}_2^l \cdots \hat{y}_J^l]^T\). The trajectory of the \(l\)th articulatory feature is then smoothed using a low-pass filter \(h_i^l\) with a cut-off frequency \(f_c^l\). To avoid any phase distortion due to the low-pass filtering on the articulatory trajectory, 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 \(f_c^l\) is optimized on the development set of the training subject.

### 4. Experiments and results

#### 4.1. Experimental setup

We conduct the SII experiments separately using three generic acoustic spaces, namely, the TIMIT corpus, the Boston University Radio Speech (BN) corpus and a combination of both the TIMIT and the BN corpora.

The TIMIT corpus contains read speech recordings of 630 speakers, each of them speaking ten sentences, recorded in a quiet environment. The corpus contains eight major dialects of American English. We have excluded the ‘sa1’, ‘sa2’ recordings as they were recorded for speaker calibration. Thus, we have used 5040 recordings of the corpus, spanning a total duration of \(~4.29\) h.

The BN corpus comprises recordings from radio broadcasts and also lab simulations of the broadcasts. In the lab news data, the announcers read the stories in both radio announcement style and non-radio style. The recordings are in American English, from seven FM radio announcers, four of them being male, and three female. The recordings are both in quiet and noisy environments. We have excluded the noisy recordings resulting in an audio data of \(~10.80\) h.

While the speakers in the TIMIT corpus outnumber that in the BN corpus, the TIMIT corpus is comprised only of read speech, unlike spontaneous speech in the BN corpus. We combine both corpora to improve the acoustic richness
in terms of speaker as well as speaking style variability in preparing GAS for SII. We refer to this GAS by TIMITBN.

The number of acoustic clusters $K$ is set to be 39 in the case of IS1: $K = 39$ is chosen to make the number of clusters identical to the number of phonemes used in IS2H and IS2A. $K$-means clustering with Euclidean distance is used for clustering the MFCCs of GAS in the case of IS1. The acoustic clusters for IS2H and IS2A correspond to 39 broad phonemes in the case of the TIMIT corpus. These broad phonemes are determined following the work by Lee and Won (1989). When GAS is chosen as BN and TIMITBN, we have added an extra phoneme to model breathing in the recording (i.e. a total of 40 units). This is because a significant portion of the recording in the BN corpus contains sounds due to breathing. The acoustic clusters in IS2H and IS2A are obtained by finding the phonetic boundaries for each sentence in the GAS using phonetic HMMs trained on the GAS and the corresponding transcriptions. Similarly the acoustic clusters in IS3H and IS3A correspond to the states of left-to-right phonetic HMMs with three emitting states. Thus, there are $117 (= 39 \times 3)$ acoustic clusters when TIMIT is chosen as the GAS and 120 clusters when BN and TIMITBN are chosen as the GAS (3 more states corresponding to the 3 states of the breathing HMM). The phone recognition accuracies on the TIMIT, BN, and TIMITBN with trained three-state HMMs are 87.11%, 79.76%, 74.44% respectively. It should be noted that no language model has been used for the recognition experiments. The drop in accuracy in the case of the corpora involving BN could be due to the fact that recording in the BN corpus is more noisy, unlike in the clean recordings of TIMIT made in a quiet environment; hence, discrimination among phonetic classes could be better in the case of TIMIT. For all inversion schemes, the number of mixtures for each acoustic cluster GMM is chosen to be $M = 256$ (as in (2)). For the HMM adaptation in IS2A and IS3A, we have used the MLLR adaptation method (Leggetter and Woodland, 1995). A supervised adaptation framework is used which involves a static two-pass adaptation approach. The entire set of available utterances of the particular speaker is used as the adaptation data.

The accuracy of SII depends on how accurately $B_k^j$ is estimated given a test acoustic feature vector $u_n$ at the $n$th frame. In IS2H, IS2A, IS3H, and IS3A, the estimation of $B_k^j$ is primarily determined by the recognition accuracy on the test utterance using HMMs trained on GAS. While IS2H and IS3H use a non-adapted HMM trained on GAS for recognition, IS2A and IS3A use HMMs adapted to the training and test subjects for SII. Table 1 shows the phone recognition accuracies for the MOCHA male and female subjects using both non-adapted and adapted HMM trained on different GASes. No language model has been used in phone recognition experiments. It is clear from Table 1 that for both the MOCHA subjects, adaptation improves recognition accuracy with relative improvement ranging from 50% to 100%. This suggests that, with adaptation, the estimation of $B_k^j$ and, hence, the quality of SII could improve.

Since the MOCHA corpus recording was done in a silent environment (similar to TIMIT) the recognition accuracy on the MOCHA subjects using HMM trained on TIMIT or TIMITBN is better than that using BN (Table 1). The degree of mismatch between the acoustic space of GAS and that of the MOCHA data determines the recognition accuracy and, hence, the normalization as well as the inversion quality. Thus, different choices of GAS would reveal the characteristic role of different GASes for inversion.

Table 1: Phone recognition accuracy on the MOCHA subjects using HMM trained on different GAS.

<table>
<thead>
<tr>
<th>Phone recognition accuracy (%) using HMM trained on GAS</th>
<th>No adaptation</th>
<th>With adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TIMIT</td>
<td>BN</td>
</tr>
<tr>
<td>MOCHA male</td>
<td>27.03</td>
<td>19.37</td>
</tr>
<tr>
<td>MOCHA female</td>
<td>34.53</td>
<td>23.07</td>
</tr>
</tbody>
</table>
A 10-fold cross validation setup on both subjects of the MOCHA corpus separately. Inversion is performed by a GMM based mapping between acoustic and articulatory spaces. One fold is used for testing, one fold for development, and the remaining 8 folds are used for training. This is repeated ten times in a round-robin fashion. Due to matched acoustics between the training and test sets, ISm1 would yield the best inversion accuracy for every test subject that could be achieved by any SII schemes. Thus, the performance of ISm1 can be considered to be an upper limit on the performance of the proposed SII schemes.

The development set for all inversion schemes – ISm1, IS0, IS1, IS2H, IS3H, IS2A, IS3A – is used to optimize the cut-off frequencies \( \{f_j^t, j = 1, 2, \ldots, J \} \) of the articulator-specific low-pass filter for smoothing estimated articulatory trajectories. This is done by a grid search method where the possible values of cut-off frequency are chosen from the following set: \{3, 4, 5, \ldots, 25\} Hz. Since the articulatory trajectories are low-pass in nature (Ghosh and Narayanan, 2010), the above set is assumed to cover possible ranges of required cut-off frequencies.

Let us denote the original articulatory trajectories corresponding to the test utterance by \( \textbf{w}_n = [w_n^1, \ldots, w_n^J]^T \), \( 1 \leq n \leq N \) corresponding to the acoustic features \( \textbf{u}_n \), \( 1 \leq n \leq N \). It is important to note that the estimated articulatory trajectories \( \textbf{y}_n^* \) correspond to the articulatory co-ordinates of the training subject; this is because \( \textbf{y}_n^* \) is derived using \( \eta_n^* \) which is, in fact, a subset of the training subject’s articulatory data \( \textbf{y}_n \). On the other hand, \( \textbf{w}_n \) corresponds to the articulatory co-ordinates of the test subject. Since \( \textbf{y}_n^* \) and \( \textbf{w}_n \) are mismatched due to the different shapes and sizes of the vocal tracts of the train and test subjects, we cannot directly use the root mean squared error (RMSE) as the performance measure, as typically done in the acoustic-to-articulatory inversion literature. Although the co-ordinates of \( \textbf{y}_n^* \) and \( \textbf{w}_n \) are different, the shape of the trajectories is expected to be similar for a successful inversion because both of them correspond to the same spoken utterance. The similarity in the shape of the trajectories is measured using Pearson’s correlation coefficients, \( \rho_j^t \) (defined in (9)) separately for individual articulators. Hence, the greater the value of \( \rho_j^t \), the better is the quality of inversion.

\[
\rho_j^t(\textbf{w}_n^t, \textbf{y}_n^*) = \frac{\sum_{n=1}^{N} (w_n^j - \overline{w}_j)(y_n^j - \overline{y}_j^*)}{\sqrt{\sum_{n=1}^{N} (w_n^j - \overline{w}_j)^2} \sqrt{\sum_{n=1}^{N} (y_n^j - \overline{y}_j^*)^2}} 
\]  

where \( \overline{w}_j = \frac{1}{N} \sum_{n=1}^{N} w_n^j \) and \( \overline{y}_j^* = \frac{1}{N} \sum_{n=1}^{N} y_n^j^* \).

4.2. Results and discussion

The inversion experiments have been conducted for both the EMA and the TV features separately and the results are described below.

4.2.1. Acoustic-to-articulatory inversion using EMA features

Error bar plots in Fig. 2(a) and (b) demonstrate the correlation coefficient (\( \rho_j^t \)) for different articulators, averaged
(± one std. dev.) over all test utterances, using different acoustic-to-articulatory inversion schemes; the TIMIT speech corpus is used for GAS. On average, ISm1 achieves the highest accuracy, as it is a subject dependent inversion scheme applied in the case when the train and test subjects are matched. It is clear from Fig. 2 that the performance using IS0 is inferior to ISm1 because, in IS0, the subject dependent inversion scheme is applied in the case when the train and test subjects are different.

All the SII schemes show a varying degree of performance with a trend of increasing averaged correlation coefficients in the following order – IS1, IS2H, IS3H, IS2A, IS3A. IS1 yields a similar performance with IS0 for the majority of the articulators except for TTx, TBx, TDx, Vx, ULy, TTy, TBx, TDy (for the male test subject) and for ULx, Vx, TBy (for the female test subject). This could be due to the poor clustering involved in obtaining acoustic clusters in GAS for the IS1 scheme. However, when the acoustic clusters are determined using phonetic identity (as in IS2H) or using states of a phonetic HMM (as in IS3H), the SII performance significantly improves and becomes better than IS1 for most of the articulators except for TTx, TDx, TBx, TDy (for male test subject) and for TBx and TDx (for female subject). Thus, there is a clear benefit obtained by using phonetic identity-based or phonetic HMM’s state-based acoustic clusters as opposed to K-means clustering-based acoustic clusters in SII. However, there is no significant difference between the performance from IS2H and IS3H; this suggests that more acoustic clusters (using states of phonetic HMMs) do not help improve the acoustic normalization compared to fewer acoustic clusters (using the phonetic identity of each frame). When the HMMs are adapted (as in IS2A and IS3A) to the training and test subjects in SII for finding the best subset \( B^k \), the SII performance further improves and becomes better than IS2H and IS3H for all articulators except for TDx (for the female test subject). This benefit is a direct consequence of improved HMM parameters due to adaptation which improves the recognition and, hence, the quality of \( B^k \). However, there is no significant difference between IS2A and IS3A suggesting that the SII performance using phonetic identity-based acoustic clusters is similar to that using finer acoustic clusters.

Fig. 3 shows the acoustic-to-articulatory performance on the EMA features when the BN corpus is used as GAS. The relative performances of different SII schemes remain similar to those using the TIMIT corpus for GAS (Fig. 2). However, on average, there is a drop in the correlation coefficient values when the BN corpus is used for GAS compared to when the TIMIT corpus is used. This could be due to the fact that the speaking styles in the BN corpus are different from those of the TIMIT ones leading to a better normalization due to greater acoustic similarities with the MOCHA-TIMIT. The drop in the correlation coefficient could also be due to there being fewer speakers in the BN corpus – TIMIT has 630 whereas BN has only 7. Thus, GAS of the BN corpus may not be rich with speaker variability unlike the GAS of the TIMIT corpus. Speaker variability in GAS could make the normalization more robust. It could also be due to the mismatch in

---

**Fig. 3.** Acoustic-to-articulatory inversion performances (in terms of average (± one SD) correlation coefficient) for 14 EMA features using various inversion schemes when the BN corpus is used as GAS: (a) Training and test data are from the MOCHA female and male speakers, respectively. (b) Training and test data are from the MOCHA male and female speakers, respectively. For ISm1, the training and test subjects are matched.
recording environments – TIMIT is in a quiet environment and BN is in both quiet and noisy environments, whereas the MOCHA corpus is recorded in a quiet environment.

Fig. 4 shows the acoustic-to-articulatory inversion performance when the GAS is formed by combining the TIMIT and the BN corpora. The relative performance of different SII schemes remains similar to those obtained using TIMIT as GAS. However, on average, the accuracy is slightly better than when BN is used as GAS and approximately the same when TIMIT is used as GAS. We have summarized the relative performances of different SII schemes for the three different GAS schemes – TIMIT, BN, and TIMITBN, in Table 2; the correlation coefficients of 14 different EMA features are averaged to report an average performance of each inversion scheme separately for all GASes. Each row in Table 2 corresponds to a combination of GAS, articulatory features and training and test subject choices; the highest performance among all schemes (except ISm1, which is underlined) in each row is marked in bold. Considering the rows corresponding to EMA features in Table 2, it is clear that IS3A (or IS2A) yields the highest inversion performance. Thus, when there is a mismatch between train and test subjects in inversion, the IS3A scheme performs better than IS0, which is typically used when the training and test subjects are identical.

4.2.2. Acoustic-to-articulatory inversion using TV features

Figs. 5–7 (similar to Figs. 2–4) show the acoustic-to-articulatory inversion performances on six TV features for both subjects from MOCHA using three GASes: TIMIT, BN, TIMITBN respectively. The performance of different SII schemes on TV features improves in the following order – IS1, IS2H, IS3H, IS2A and IS3A. This is identical to that for the EMA features. For most cases, the performances using IS2H and IS3H are identical; similarly, the performances using IS2A and IS3A are similar suggesting no significant benefit by increasing the number of acoustic clusters using states of phonetic HMMs, compared to acoustic clusters derived using just a phonetic identity. However, on average, the schemes with adapted HMMs (i.e. IS2A and IS3A) are better than schemes with no adaptation (i.e. IS2H, IS3H) indicating the benefit of adaptation in finding the subsets \( B_k \) for estimating TV features. ISm1 yields the highest correlation coefficient (oracle performance) and is considered to be the upper limit on the SII performance. It is interesting to observe that the performance using IS3A turns out to be closer to the oracle performance for some TV features; for example, the correlation coefficient in the case of LA using TIMIT as GAS on the male test subject (Fig. 5(a)) increases from 0.49 (IS1) to 0.71 (IS3A) whereas the oracle correlation coefficient is 0.86. For a few TV features (LA, VEL, TBCD) the performance using IS1 is worse than that using IS0. This could be due to the poor clustering of the GAS in IS1.

Table 2 shows the correlation coefficient averaged across six TV features for various acoustic-to-articulatory inversion schemes. In general, performance when using the TV features is better than that obtained with EMA features for all inversion schemes. This could mean that the TV features provide an articulatory representation which is more invariant for each acoustic unit compared to raw

---

**Table 2**

<table>
<thead>
<tr>
<th>Train: Female, Test: Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train: Male, Test: Female</td>
</tr>
</tbody>
</table>

Fig. 4. Acoustic-to-articulatory inversion performances (in terms of average (+one SD) correlation coefficient) for 14 EMA features using various inversion schemes when the TIMITBN corpus is used as GAS. (a) Training and test data are from the MOCHA female and male speakers, respectively. (b) Training and test data are from the MOCHA male and female speakers, respectively. For ISm1, the training and test subjects are matched.
kinematic) EMA features and hence, yield better inversion estimates. This is consistent with the observation made by Mitra et al. (2011). Similar to the EMA feature based results, the best performance among different SII schemes is obtained using IS3A (or IS2A in some cases, as indicated by the ‘bold’ font). A comparison of IS0 and IS3A for all rows in Table 2 reveals that the relative benefit of the improved SII over the typical SDI varies from 15.38% to 32.50% in terms of the average correlation coefficient. It is interesting to observe that this benefit is the smallest when BN is used as GAS, indicating that the acoustic normalization and, hence, the benefit in SII is greater when the GAS and the acoustics of the training and test subjects in SII are similar. Overall, IS3A turns out to be the best proposed SII inversion scheme for both the EMA and TV features. However, the oracle performance (ISm1) is still better than the IS3A performance, which indicates there is still room for improvement.

### 4.2.3. Role of acoustic clusters in SII

It is important to note that the acoustic subsets in IS1 are formed in an unsupervised manner while those in

<table>
<thead>
<tr>
<th>GAS</th>
<th>Feature</th>
<th>Train-test</th>
<th>ISm1</th>
<th>IS0</th>
<th>IS1</th>
<th>IS2H</th>
<th>IS3H</th>
<th>IS2A</th>
<th>IS3A</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>EMA</td>
<td>Female–male</td>
<td>0.66</td>
<td>0.42</td>
<td>0.34</td>
<td>0.47</td>
<td>0.48</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male–female</td>
<td>0.76</td>
<td>0.4</td>
<td>0.35</td>
<td>0.49</td>
<td>0.48</td>
<td>0.52</td>
<td>0.53</td>
</tr>
<tr>
<td>TV</td>
<td>Female–male</td>
<td>0.8</td>
<td>0.52</td>
<td>0.4</td>
<td>0.55</td>
<td>0.54</td>
<td>0.6</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Male–female</td>
<td>0.8</td>
<td>0.51</td>
<td>0.44</td>
<td>0.6</td>
<td>0.6</td>
<td>0.64</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>BN</td>
<td>EMA</td>
<td>Female–male</td>
<td>0.66</td>
<td>0.42</td>
<td>0.33</td>
<td>0.45</td>
<td>0.46</td>
<td>0.5</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male–female</td>
<td>0.76</td>
<td>0.4</td>
<td>0.36</td>
<td>0.47</td>
<td>0.47</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>TV</td>
<td>Female–male</td>
<td>0.8</td>
<td>0.52</td>
<td>0.42</td>
<td>0.53</td>
<td>0.54</td>
<td>0.59</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Male–female</td>
<td>0.8</td>
<td>0.51</td>
<td>0.47</td>
<td>0.58</td>
<td>0.59</td>
<td>0.63</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>TIMITBN</td>
<td>EMA</td>
<td>Female–male</td>
<td>0.66</td>
<td>0.42</td>
<td>0.33</td>
<td>0.47</td>
<td>0.48</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male–female</td>
<td>0.76</td>
<td>0.4</td>
<td>0.35</td>
<td>0.49</td>
<td>0.49</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>TV</td>
<td>Female–male</td>
<td>0.8</td>
<td>0.52</td>
<td>0.4</td>
<td>0.55</td>
<td>0.54</td>
<td>0.6</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Male–female</td>
<td>0.8</td>
<td>0.51</td>
<td>0.45</td>
<td>0.6</td>
<td>0.61</td>
<td>0.64</td>
<td>0.65</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 5. Acoustic-to-articulatory inversion performances (in terms of average (±one SD) correlation coefficient) for 6 TV features using various inversion schemes when the TIMIT corpus is used as GAS: (a) Training and test data are from the MOCHA female and male speakers, respectively. (b) Training and test data are from the MOCHA male and female speakers, respectively. For ISm1, the training and test subjects are matched.
IS2H (also IS2A, IS3H and IS3A) are formed in a supervised manner. There are also key differences in how the best subset $B_k$ in a test frame is identified – in IS1 the best subset is found by the index of the maximum valued element of the probability vector, while Viterbi decoding is used to find the best subset in IS2H. The better performance in IS2H over IS1 could be due to either better acoustic subsets or better estimations of $B_k$. To examine the role of acoustic
clusters in an SII performance, we perform an experiment in which we use acoustic subsets $A_k$ generated by the phonetic identity of each feature vector of GAS as in IS2H; however, $B_k$ and $B_l$ are determined using (6) and (7) as in IS1. We refer to this scheme by IS1-IS2Clusters. IS1-IS2Clusters yields a correlation coefficient of 0.45 for EMA features using the female as the train and the male as the test subject with the TIMIT corpus as the GAS; this is a significant improvement over that using IS1 (0.34) but less than that using IS2H (0.47). Similarly, IS1-IS2Clusters yields an inversion performance of 0.46 for the male as the train and the female as the test subject, which is between those using IS1 (0.35) and IS2H (0.49). Likewise, for the TV features, the inversion performances are 0.56 and 0.59 for the train-test combination as female–male and male–female respectively. When the BN corpus is used to construct the GAS, the inversion performances using IS1-IS2Clusters are 0.45, 0.45, 0.53 and 0.59 respectively for the above four cases. Similarly, when TIMITBN is used for the GAS, the performances are 0.45, 0.48, 0.55 and 0.6 respectively. In all these cases, the performance using IS1-IS2Clusters is significantly better than that using IS1 and very close to that using IS2H. This suggests that designing better acoustic subsets contributes more to improving the SII performance than better estimates of $B_k$.

### 4.2.4. Role of GAS in SII

In IS0, the acoustic-articulatory mapping of the training subject is modeled using a GMM, although the acoustics of the test subject is different from that of the training subject. In the proposed methods for SII, the acoustic-articulatory mapping is determined in a non-parametric manner using GAS. We would like to examine the SII performance using a technique in a non-parametric manner similar to the proposed methods, but without any GAS. This would provide another lower bound on the SII performance in addition to that using the parametric approach as in IS0. This is done using a KD-tree (denoted by IS0-KDtree) where the acoustic features of the training subject are used in building the KD-tree, followed by a $k$-nearest neighbor (KNN) search over the KD-tree to get the $L$ closest features $\left(\eta_n^l\right)$ from the test subject’s acoustic features. The inverse of the distances computed between these $L$ features and the test subject’s features are then used to compute $p_n^l$. The inversion performance of the IS0-KDtree is shown in Table 3. For comparison, the performance of the IS0 is also tabulated for different articulatory features and training-test subject combinations. It is clear that the performance of IS0-KDtree is slightly better than that of IS0 for male–female as the train-test combination for both EMA and TV features. However, the performance of IS0-KDtree is slightly worse than that of IS0 for female–male as a train-test combination. Overall, the performance of IS0-KDtree is not significantly different from that of IS0 suggesting that both parametric and non-parametric approaches perform similarly when there is an acoustic mismatch between the training and test subjects in inversion. Also, a comparison of the IS0-KDtree with the other proposed SII schemes suggests that acoustic normalization indeed helps in improving the inversion performance in the presence of acoustic mismatch.

#### 4.2.5. Efficiency in computing $\eta_n^l$ and $p_n^l$

$\eta_n^l$ and $p_n^l$ are calculated using (5), which reduces the computation by pruning the search space in the training data as well as using an approximation to the Euclidean distance. In this section, we examine the complexity and efficiency of computing $\eta_n^l$ and $p_n^l$ using (5) in comparison to a full search using Euclidean distance. A KD-tree is used for performing the full search. We use IS1 for this comparison. $\Phi(z_i)$ from IS1 are used in building a KD-tree, followed by a $k$-nearest neighbor (KNN) search over the KD-tree to get the $L$ closest training articulatory features $x_i$ from $\Phi(u_n)$. These are denoted by $\eta_n^l$, $1 \leq l \leq L$. The inverse of the Euclidean distances between the corresponding $\Phi(z_i)$ and $\Phi(u_n)$ is used to compute $p_n^l$, $1 \leq l \leq L$. We refer to this method as IS1-KDtree.

Table 4 shows the inversion performance using IS1-KDtree averaged over all articulators. For comparison,
the performance of IS1 is also shown in Table 4. It is clear from Table 4 that the correlation coefficient from IS1-KDtree is \( \sim 7.5\% \) (relative) higher than when using IS1. This demonstrates that a full search with Euclidean distance yields a better inversion performance than when using (5). However, IS1-KDtree is \( \sim 60 \) times slower compared to IS1. Average time to perform SII using IS1-KDtree for each audio file is approximately 8.15 s, while when using IS1 it is only 0.14 s; these are obtained by running the inversion experiments using MATLAB R2010b in a Desktop computer with an Intel Core i7 processor. Thus, there is a trade-off between the inversion performance and the speed of the inversion algorithm. This is also true for IS2H, IS2A, IS3H, IS3A because we have found a similar trade-off by performing a full search method (in a manner similar to IS1-KDtree) in those SII schemes. The average duration of each wavfile is 2.75 s and hence, for a real time inversion, IS1 could be a better choice than IS1-KDtree. It should also be noted that although IS1-KDtree performs better than IS1, it is still not the best performing inversion scheme. IS3A yields a better inversion performance than IS1-KDtree. Therefore, in the next section, we examine the performance of IS3A using Euclidean distance without any approximation.

### 4.2.6. Effect of approximate Euclidean distance on SII using IS3A

Once the best subset \( B_k^* \) is obtained in IS3A, the approximated Euclidean distance \( D(\Phi(u_n), \Phi(z_k^l)) \) (as defined in (5)) is used for computing \( p_{ik}^l \). In this subsection, we examine the performance of IS3A when the approximate Euclidean distance is replaced with Euclidean distance as follows:

\[
D(\Phi(u_n), \Phi(z_k^l)) = ||\Phi(u_n) - \Phi(z_k^l)||_2
\]  

We refer to this method as IS3A-Euclidean. The results obtained by IS3A-Euclidean are shown in Table 5. On average, the correlation coefficient obtained by using IS3A-Euclidean is 0.02 (absolute) over that of IS3A. This demonstrates that the proposed approximation to Euclidean distance does not lead to a considerable degradation of SII performance. Thus, the Euclidean distance could be used when there is no constraint on the computation time. But, in practice, the approximate Euclidean distance could be used without any significant drop in inversion performance.

### 4.2.7. Effect of recognition accuracy on SII performance

SII using IS2H, IS2A, IS3H and IS3A requires phone recognition of the test utterance to obtain the indices of the subsets \( B_k^* \) in each frame. A better estimate of the subset indices improves the SII performance. Thus, higher
recognition accuracy results in a better SII performance. In this section, we examine how the inversion performance varies with recognition accuracy. Recognition accuracy could vary depending on the amount of data used to train the recognizer or the degree of acoustic mismatch between the training data and the test acoustic. For example, the phone recognition accuracy on the MOCHA-TIMIT subjects is the highest when TIMIT is used for training a recognizer followed by that using TIMITBN and BN. This is true with and without adaptation (Table 1). We also observe a similar pattern in the SII performance for IS2H and IS3H (without adaptation) as well as for IS2A and IS3A (with adaptation) – the SII performance is the highest when TIMIT is used as GAS followed by TIMITBN and BN as GAS (Table 2). This suggests that better recognition accuracy would lead to a better SII performance.

Recognition accuracy could also vary depending on the amount of adaptation data. We experiment on the role of the amount of adaptation data on the SII performance by using 25%, 50% and 75% of the MOCHA-TIMIT subjects’ data for adaptation of the HMM parameters in IS3A. These inversion schemes are referred to as IS3A-25p, IS3A-50p and IS3A-75p respectively. Table 6 shows the SII performances obtained by IS3A-25p, IS3A-50p and IS3A-75p along with IS3H (no adaptation) and IS3A. Note that in IS3A the entire data of the MOCHA-TIMIT subjects are used for adaptation. It is clear from the table that, with 50% of the adaptation data, the SII performance becomes identical to that using the entire data for adaptation. With 25% of the adaptation data, the performance is very close to that using IS3A. This suggests that even 50% of adaptation data is sufficient to achieve an SII performance identical to that using the entire data for adaptation. It is also clear from Table 6 that the SII performance increases with an increase in adaptation data size.

5. Conclusions

In this work, we have found a better clustering of the generic acoustic space (GAS) for acoustic normalization in SII. Since the training and test subjects are different in SII, acoustic normalization between the two is a critical step for a successful SII. The proposed clustering of GAS yields a significant improvement in the SII performance compared to that using unsupervised K-means clustering. In particular, if the GAS is clustered into phonetic units or units corresponding to the states of a phonetic HMM, the SII performance improves by approximately 25% over the typical SDI applied in the subject-independent setting. From SII experiments with different GASes, we observed that the improvement in SII is greater when there is a similarity between the acoustics of GAS and the acoustics of the training and test subjects in SII. We also found that the estimates of the possible values of articulatory features along with their probabilities in SII become better by considering the acoustics of the test sentence or the entire test corpus compared to using the acoustic features of only the current test frame. The HMM parameters trained on the GAS, when adapted to the acoustics of the test subject, result in a better identification of the subset of the training corpus that each test frame might be similar to. This, in turn, results in a better estimate of possible articulatory features in SII and, hence, better SII performance. By applying the proposed normalization-based schemes in inversion with matched training-test condition (i.e. SDI), we observed that the inversion performance improves compared to that in SII case. But the performance is lower compared to the SDI performance obtained by GMM based inversion (i.e. Ism1) using acoustic features. Thus, for SDI, acoustic features without normalization would be preferred over those with normalization.

Although a significant improvement in the SII performance is obtained by using supervised clustering, model adaptation and selection of proper GASes, the performance of SII is still found to be lower than when training data from the target test subject are a priori available for learning the acoustic-to-articulatory map. Hence, the investigation of better acoustic normalization for further improving the SII performance is part of our ongoing research. Due to the limited availability of parallel acoustic-articulatory data, the impact of the acoustic mismatch between training and test subjects on the SII performance is not clear. It will be useful to conduct such a study and develop a system that could recommend the choice of the training subject from a pool of possible training subjects for a given test subject’s speech acoustic. Performing SII in a cross language setting could also reveal principles for designing a GAS that could potentially work across different languages, and with various dialects, accents and speaking styles.

Acknowledgement

The authors thank Prof. Shrikanth S Narayanan for useful discussions and constructive comments on the manuscript.

References


