A Comparative Study of Articulatory Features From Facial Video and Acoustic-To-Articulatory Inversion for Phonetic Discrimination

Abhishek Narwekar
Dept. of Electrical Engg.
Indian Institute of Technology Madras
Chennai, 600036, India
Email: ee11b129@ee.iitm.ac.in

Prasanta Kumar Ghosh
Dept. of Electrical Engg.
Indian Institute of Science
Bangalore, 560012, India
Email: prasanta@ee.iisc.ernet.in

Abstract—Several studies in the past have shown that the features based on the kinematics of speech articulators improve the phonetic recognition accuracy when combined with the acoustic features. It is also known that the audio-visual speech recognition performance is better than that of the audio-only recognition, which, in turn, indicates that the information from the visible articulators is complementary to that provided by the acoustic features. Typically, visible articulators can be extracted directly from a facial video. On the other hand, the speech articulators are recorded using electromagnetic articulography (EMA), which requires highly specialized equipment. Thus, the latter is not directly available in practice and hence usually estimated from speech using acoustic-to-articulatory inversion. In this work, we compare the information provided by the visible and the estimated articulators about different phonetic classes when used with and without acoustic features. The information provided by different visible, articulatory, acoustic and combined features is quantified by the mutual information (MI). For this study, we have created a large phonetically rich audio-visual (PRA V) dataset comprising of 9000 TIMIT sentences spoken by four subjects. Experiments using PRA V corpus reveal that the articulatory features estimated by inversion are more informative than the visible features but less informative than the acoustic features. This suggests that the advantage of visible articulatory features in recognition could be achieved by recovering them from the acoustic signal itself.

Index Terms - audio-visual corpus, speech articulators, visual articulators, mutual information.

I. INTRODUCTION

Kinematics of speech articulators (such as lips, jaw, tongue, velum) involved in speech articulatory movement also encodes phonetic information about various phonetic units. The acoustic speech signal obtained as a result of articulatory movement also encodes phonetic information. Several studies [1, 2, 3] indicate that the nature of information encoding in the articulatory movement and the speech signal are different and complementary. Phonetic recognition accuracy is found to improve when the articulatory features are used along with the acoustic features. Evidence from the audio-visual speech recognition also reveals similar findings. For example, recognition experiments with lip movement tracked from facial video shows that the visual features carry phonetic information but lower than that given by acoustic features [4]. When the visual features are used along with acoustic features, the phonetic recognition accuracy is found to increase compared to the individual ones [5, 6] indicating that the information encoded in the lip movement is also complementary to that in the acoustic signal.

Influence of the visual information from seeing speaker’s mouth in speech perception is well-known as the McGurk effect [7]. It has been shown that the visual information improves the speech intelligibility [8] particularly in noisy conditions [9, 10]. Interestingly, it is also shown that speech perception is improved even if a synthetic facial movement in sync with the audio is presented to the listener [11, 12].

In this work, we compare different articulatory features with and without acoustic features on their role in phonetic discrimination. The main motivation behind this study is to compare the information present in external visual articulatory features and the articulatory features estimated using acoustic-to-articulatory inversion (AAI). The study also quantifies how complementary these features are to the acoustic features. It is important to note that articulatory features from AAI would differ from the actual articulatory kinematics. However a comparison with the estimated features would reveal if the information from articulatory kinematics embedded in the speech signal is comparable to the directly available visual articulatory features. This, in turn, would answer if the visual modality could be replaced with AAI for improving the phonetic discriminability.

We begin with a brief description of a large phonetically rich audio-visual (PRA V) corpus that was created for this study.

II. DATABASE

While there are several audio-visual corpora in the literature, none of them is large and phonetically rich enough to conduct the comparative study done in this work. Hence, we created a large phonetically rich audio-visual dataset (PRA V) with 2368 TIMIT [13] sentences each spoken by four subjects (2 male - M1, M2 - and 2 female - F1, F2) aged 20-22 years. The subjects were fluent in reading, writing and speaking English.
The recording was done in a well illuminated sound-proof chamber. The video was recorded at a frame rate of 25 frames/second using a Sony Handycam – model HDR-CX280E. The distance between the camera and the subject was maintained constant by fixing the position of the seat for the subject, and varying only the height of the camera so as to capture the face of the subject completely. A light backdrop was used for the video recording.

The audio was recorded and stored at a sampling frequency of 16 KHz using Praat [14]. Each subject was explained about the functioning of the setup and allowed to practise before the actual recordings in order to familiarize them with the protocol. The sentences were displayed one per slide on a laptop computer placed between the subject and below the stand of the camera in such way that it did not block the camera view. In order to remove the incorrect utterances due to stuttering and mispronunciation, we annotated all the correct utterances in the database by delineating their start and end points using Audacity [15]. From the video frames of the PRAV corpus, we extract the visual articulatory features.

III. ARTICULATORY FEATURES FROM FACIAL VIDEO

We use the Active Appearance Model [16] to extract features from each frame of the video. These features are constructed using 22 points along the contour of the subject’s lips as shown in Fig. 1(a). Such features have been previously been used in work by Neti et al. [17] and Bregler and Konig [18]. We use the implementation for IC AAM [19] as a template, and modify it. We had to overcome certain shortcomings of the IC AAM in order to extract lip features effectively in PRAV corpus.

To generate training data for AAM, we manually select 19 frames from the recorded videos where the lip makes various shapes in different phonetic contexts. We observe that IC AAM often diverges to regions outside the lip due to the presence of high gradient areas like the nose and chin which attract the AAM contour. If this happens in more than 3 consecutive frames, we discard those frames. Otherwise, the points in these frames are interpolated using neighboring frames.

We denote the articulatory feature from a facial video frame as a 44-dimensional vector: an ordered sequence of the X- and Y-coordinates of 22 points, \( F_{VA} = \{p_{1,x}, p_{2,x}, \ldots, p_{22,x}, p_{1,y}, p_{2,y}, \ldots, p_{22,y}\} \), where \( p_{i,x} \) and \( p_{i,y} \) refer to the \( x \) and \( y \) coordinates of the \( i^{th} \) point on the contour.

IV. ACOUSTIC-TO-ARTICULATORY INVERSION

In order to estimate articulatory kinematics from speech signal, AAI is used. AAI is a challenging problem owing to the non-linear and one-to-many mapping between the acoustic and articulatory spaces. AAI is of two types: 1) subject-dependent inversion [20], where the test subject’s data is also available for training and 2) subject-independent inversion (SII) [21, 22], where the test subject’s data is not available for training. In our experiments, we implement SII since directly measured articulatory data was not available from any of the subjects in PRAV corpus.

We have used the Multichannel Articulatory (MOCHA) database [23] as the training data for the SII in this work. MOCHA has two British English speakers - 1 male and 1 female. It comprises of the parallel recordings of both the acoustic and articulatory kinematics using electromagnetic articulography (EMA) for 460 utterances spoken by each of the four subjects. The articulatory features given by MOCHA contain the X- and Y-coordinates of 7 points, corresponding to the upper and lower lips, lower incisor, tongue tip, tongue body, tongue dorsum and velum, as shown in Figure 1(c). Similar to the case of the AAM coordinates, we denote the

![Fig. 1. 1(a): 22 manually marked points long the lip-contour. 1(b): A sample AAM output. 1(c): Placements of the EMA sensors in the MOCHA database. UL-Upper lip, LL-Lower lip, LI-Lower Incisor, TT-Tongue Tip, TB-Tongue Body, TD-Tongue Dorsum, V-Velum.](image)

The key step in SII is the compensation for the acoustic mismatch between the train and the test subjects. To do so, a probability feature vector is constructed using a transformation of the acoustic features onto a generic acoustic space (GAS) that consists of acoustic features from a large number of speakers. This normalization helps in the comparison of the probability feature vectors from training and test subjects. Acoustic normalization has been shown to improve by choosing the sound units used in GAS appropriately [22].

The general procedure in an SII algorithm is as follows: the acoustic feature vectors in the GAS are clustered into \( K \) components. For any acoustic feature vector, a \( K \)-dimensional probability feature vector are obtained, where the \( i^{th} \) element \( (1 \leq i \leq K) \) is the posterior probability of that vector given the \( i^{th} \) cluster. Following the normalized representation for the training and test data, we find the \( L \) closest feature vectors in the training data to the test vector and use the corresponding articulatory feature vectors to estimate the articulatory feature of the test vector.

We perform the SII experiments separately using 3 different GAS: the TIMIT corpus, the Boston University Radio Speech (BN) corpus and a combination of both the TIMIT and the BN corpora (TIMITBN). The TIMIT corpus comprises of audio recordings of 630 speakers speaking 10 sentences each, in a quiet environment. The BN corpus contains quiet and noisy recordings from radio broadcasts and lab simulations of the broadcasts. In the lab simulations, the announcers read the stories in both manner of a radio announcement and a non-radio style announcement. The TIMIT corpus has a far greater number of speakers than the BN corpus. However, the TIMIT corpus contains only read speech, while the BN corpus has spontaneous speech.

We have used the Multichannel Articulatory (MOCHA) database [23] as the training data for the SII in this work. MOCHA has two British English speakers - 1 male and 1 female. It comprises of the parallel recordings of both the acoustic and articulatory kinematics using electromagnetic articulography (EMA) for 460 utterances spoken by each of the four subjects. The articulatory features given by MOCHA contain the X- and Y-coordinates of 7 points, corresponding to the upper and lower lips, lower incisor, tongue tip, tongue body, tongue dorsum and velum, as shown in Figure 1(c). Similar to the case of the AAM coordinates, we denote the
articulatory feature from the EMA as a 14-dimensional vector: an ordered sequence of the X- and Y-coordinates of 7 points, \( F_{AA} = \{ q_{1,x}, q_{2, x} \ldots q_{7,x}, q_{1,y}, q_{2,y} \ldots q_{7,y} \} \). The points \( q_1 \) to \( q_7 \) correspond to upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue body (TB), tongue dorsum (TD) and velum (V) respectively.

V. Mutual Information Based Analysis

We use mutual information (MI) [24] for quantifying the information provided by different feature sets about phonetic categories. MI indicates the statistical dependency between two random variables. Let \( z \) be a random vector representing the visual features and \( X \in \{ 1, 2 \cdots B \} \) be a discrete random variable representing the \( B \) phonetic categories. MI \( (I(z, X)) \) between \( z \) and \( X \) is calculated using features in all available frames with their phonetic labels. We know:

\[
I(z, X) = H(X) - H(X|z),
\]

where \( H(X) \) is the entropy of \( X \) [24]. Thus, higher the MI, lower the uncertainty of the phonetic categories \((X)\) given the feature vector \((z)\), i.e., \( H(X|z) \). Note that \( z \) is a continuous random variable. Thus, to compute MI, the probability density function (PDF) of \( z \) and the probability mass functions (PMF) of \( X \) need to be known. Since the PDF of \( z \) is unknown, we quantize the space of \( z \) (denoted by \( Q(z) \)) using the feature vectors with a finite number \((K)\) of quantization bins using the K-means vector quantization [24, 25]. Then, we compute the MI by estimating the joint distribution of \( z \) and \( X \) in the finite alphabet space \((\mathbb{R}^{K \times B})\) using standard maximum likelihood criterion – frequency counts [25] and finally applying the definition of MI for discrete random variables as follows:

\[
I(Q(z), X) = \sum_{x=1}^{B} \sum_{z=1}^{K} P(Q(z) = z, X = x) \times \log \frac{P(Q(z) = z, X = x)}{P(Q(z) = z)P(X = x)}
\]

\( I(Q(z), X) \) is a lower bound of the MI between \( z \) and \( X \). \( I(Q(z), X) \) converges to the actual MI with more quantization bins \((K)\). We have chosen \( K=256 \) since increasing \( K \) further does not change the results significantly.

VI. Experiments and Results

A. Experimental Setup

The sampling frequency for the EMA features is 100 Hz. However, the frame rate of the video recording is 25 Hz. In order to match the sampling frequency of the visual features with the estimated articulatory features, we upsample the visual feature vector sequence by 4. The number of frames, thus, retained for M1, M2, F1 and F2 are 839846, 768663, 1006107 and 871392 respectively.

The Mel Frequency Cepstral Coefficients (MFCC) are known to be effective for automatic speech recognition (ASR) systems. In our computation of 39-dimensional MFCC features, we windowed the speech signal using a 20 ms window with a window overlap of 10 ms to match with the rate of articulatory feature sequence. For computing MI between phonemes and different acoustic and articulatory features, we need to determine phonetic boundaries in the speech recordings, which are obtained by a phonetic forced alignment (using 39 phoneme set [26]) using MFCC features with HMM Tool Kit (HTK) [27] and the available transcription of the utterances spoken by the subjects in the PRAV corpus.

We compute the MI between the visual and estimated articulatory and the MFCC feature sets with 39 phonemes. For estimating features from SII, we train the model with articulatory data from the male and female speakers of the MOCHA corpus separately and report their respective results. Moreover, we experiment with various GAS, viz. TIMIT, BN and TIMITBN, as described in Section IV. We also explore if the visual and estimated articulatory features are complementary to acoustic features. Using a combination of visual and estimated articulatory features separately with the MFCC features, we obtain two combined feature sets. We then compute the MI with these combined feature sets and the phonemes. Since acoustic and articulatory features have different ranges of values, we have normalized both sets of features to zero mean and unit standard deviation.

The estimated articulatory features from SII have only two points on the lips, but there are 22 points on the lips in the visual features. For a fair comparison, we reduce both feature sets. We denote the reduced estimated articulatory feature vector by \( F'_{AA} = \{ q_{1,x}, q_{2,x}, q_{1,y}, q_{2,y} \} \). For the visual features, we perform a similar reduction by extracting the features that correspond to the X- and Y-coordinates of the upper and lower lips. We define the location the upper lip as the mean of the coordinates of the inner and outer points of the upper lip as shown in Figure 1(a). This corresponds to points 17 and 5. Similarly, for the lower lip, the points of interest are 21 and 12. Thus the reduced feature vector is denoted by \( F'_{VA} = \{ P_{1,x} + P_{1', x}, P_{12,x} + P_{21,x}, P_{5,y} + P_{17,y}, P_{12,y} + P_{21,y} \} \).

B. Phonetic information by visual articulatory features

We compute MI between phonemes and different visual features in combination with MFCC – \( F_{VA}, F_{VA}+MFCC, F'_{VA} \) and \( F'_{VA}+MFCC \). MI is computed in batches of 100,000 frames, for each subject and each feature type. We report the mean and standard deviation (SD) of MI across all batches in Table I.

As seen in Table I, the average MI for the visual features alone lies in the range of 0.121 - 0.197 for all subjects. As expected, when the reduced visual features are considered, the MI drops by \( \sim 0.02 \) for all subjects. The MFCC features, on the other hand, have a significantly higher MI (\( p < 0.001 \) for all subjects), which lies in the range of 1.206 - 1.392. Concatenation of visual features with MFCC features does not always improve the MI. However, merging the MFCC features with reduced visual increases the average MI of the MFCC features in all 4 cases, although these improvements are not significant.
C. Phonetic information by estimated articulatory features

In Table II, we compare the MI using 14-dim articulatory features $F_{AA}$ estimated using SII using three different GAS (TIMITBN, TIMIT, BN) with and without the MFCC features. For training SII, the male and female subject from MOCHA were used. MI in these two cases are computed and reported in the columns of Table II as indicated by M and F respectively before the GAS.

The choices of GAS in the decreasing order of MI using $F_{AA}$ are TIMITBN, BN and TIMIT though the differences in the three quantities are not significant. It is important to note that the MI for $F_{AA}$ is always more than 0.4 - significantly higher ($p < 0.001$) than MI for the visual features, which is consistently below 0.2. This is expected since visual features capture only lip movement while the estimated articulatory features capture tongue movement as well.

When the estimated articulators are used along with MFCC, the MI significantly ($p < 0.05$) improves compared to that using only MFCC when the SII training is done with the male speaker’s features and the GAS is TIMITBN. This is unlike visual features, suggesting complementarity of the estimated articulatory movements with MFCC. Comparing Table I and Table II, it is also observed that the MI from estimated articulatory features combined with acoustic features is significantly more for all subjects compared to that from visual features combined with MFCC for the (M,TIMITBN) scheme ($p < 0.05$).

Similar to Table II, Table III shows the MI obtained using reduced articulatory features estimated from SII. It is clear that due to the reduction in features, MI has reduced by ~0.1. This drop is more compared to that obtained by reducing $F_{VA}$. This suggests that the articulatory features other than the lips, such as the tongue, velum and lower incisor, are more informative than the points on the lip that were discarded while reducing the visual features.

When the reduced estimated articulatory features are considered in combination with the MFCC features ($F_{VA}+MFCC$), the MI marginally exceeds that of MFCC features in each case. When MI of $F_{VA}+MFCC$ is compared to the MI of $F_{AA}+MFCC$, we do not observe a significant difference in MI values.

Comparing the last but one column of Table I with ‘w/o. MFCC’ column under ‘M,TIMITBN’ in Table III, it is clear that MI obtained using $F_{AA}$ is significantly more ($p < 0.001$) than that using $F_{VA}$ although both represent the movement of

### Table I

<table>
<thead>
<tr>
<th>Subjects</th>
<th>MI</th>
<th>$F_{VA}$ w/o. MFCC</th>
<th>$F_{VA}$ w. MFCC</th>
<th>$F_{VA}$ w/o. MFCC</th>
<th>$F_{VA}$ w. MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>1.206 0.021</td>
<td>0.164 0.002</td>
<td>1.238 0.042</td>
<td>0.139 0.004</td>
<td>1.239 0.019</td>
</tr>
<tr>
<td>M2</td>
<td>1.392 0.028</td>
<td>0.168 0.029</td>
<td>1.380 0.045</td>
<td>0.147 0.024</td>
<td>1.481 0.070</td>
</tr>
<tr>
<td>F1</td>
<td>1.281 0.013</td>
<td>0.121 0.019</td>
<td>1.258 0.049</td>
<td>0.106 0.013</td>
<td>1.281 0.036</td>
</tr>
<tr>
<td>F2</td>
<td>1.326 0.062</td>
<td>0.197 0.008</td>
<td>1.391 0.063</td>
<td>0.160 0.011</td>
<td>1.370 0.055</td>
</tr>
</tbody>
</table>

### Table II

<table>
<thead>
<tr>
<th>Subjects</th>
<th>M2</th>
<th>M1</th>
<th>M</th>
<th>F</th>
<th>M2</th>
<th>M1</th>
<th>M</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMITBN</td>
<td>0.456 0.197</td>
<td>0.447 0.014</td>
<td>0.440 0.010</td>
<td>0.439 0.015</td>
<td>0.421 0.011</td>
<td>0.414 0.013</td>
<td>0.395 0.026</td>
<td></td>
</tr>
<tr>
<td>TIMIT</td>
<td>1.315 0.022</td>
<td>1.306 0.022</td>
<td>1.296 0.024</td>
<td>1.299 0.018</td>
<td>1.295 0.018</td>
<td>1.277 0.013</td>
<td>1.272 0.026</td>
<td></td>
</tr>
<tr>
<td>BN</td>
<td>0.353 0.019</td>
<td>0.355 0.007</td>
<td>0.352 0.022</td>
<td>0.356 0.011</td>
<td>0.351 0.005</td>
<td>0.344 0.005</td>
<td>0.350 0.021</td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>1.401 0.017</td>
<td>1.395 0.060</td>
<td>1.389 0.072</td>
<td>1.394 0.064</td>
<td>1.387 0.070</td>
<td>1.384 0.009</td>
<td>1.380 0.061</td>
<td></td>
</tr>
</tbody>
</table>

### Table III

<table>
<thead>
<tr>
<th>Subjects</th>
<th>M2</th>
<th>M1</th>
<th>M</th>
<th>F</th>
<th>M2</th>
<th>M1</th>
<th>M</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMITBN</td>
<td>0.456 0.197</td>
<td>0.447 0.014</td>
<td>0.440 0.010</td>
<td>0.439 0.015</td>
<td>0.421 0.011</td>
<td>0.414 0.013</td>
<td>0.395 0.026</td>
<td></td>
</tr>
<tr>
<td>TIMIT</td>
<td>1.315 0.022</td>
<td>1.306 0.022</td>
<td>1.296 0.024</td>
<td>1.299 0.018</td>
<td>1.295 0.018</td>
<td>1.277 0.013</td>
<td>1.272 0.026</td>
<td></td>
</tr>
<tr>
<td>BN</td>
<td>0.353 0.019</td>
<td>0.355 0.007</td>
<td>0.352 0.022</td>
<td>0.356 0.011</td>
<td>0.351 0.005</td>
<td>0.344 0.005</td>
<td>0.350 0.021</td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>1.401 0.017</td>
<td>1.395 0.060</td>
<td>1.389 0.072</td>
<td>1.394 0.064</td>
<td>1.387 0.070</td>
<td>1.384 0.009</td>
<td>1.380 0.061</td>
<td></td>
</tr>
</tbody>
</table>
upper and lower lips points. This could be because $F_{AA}^*$ is estimated by regressing against the MFCC features, which carry more information than $F_{VA}^*$. It can also be due to the estimation error in the visual features introduced by AAM. This in turn causes estimated articulatory features to provide more MI compared to the ones obtained from visual input.

VII. Conclusion

In this paper, we use mutual information to perform a comparative study for the information provided by the articulators obtained from the facial video (using AAM) with that of the articulators estimated from speech signal using SII. In order to facilitate the study, we create the PRAV corpus, a phonetically rich audio-visual dataset. We also compare two feature sets when used along with the acoustic features (MFCC) to investigate their complementary property. Based on a mutual information-based quantification of the information content in these feature sets, we find that the estimated articulators extracted from speech are more informative than the visual articulators extracted using from facial video. This finding would be useful in developing better speech recognition framework by extracting articulatory features without any need for additional modality such as facial video. It would be interesting to extend the study to more subjects using recordings from realistic environments. Performing the MI based analysis for various classes of phonemes such as vowels, consonants, nasals and fricatives could provide important insights too. These are parts of our future work.

Acknowledgment

This work was supported by the Department of Science and Technology (DST), Govt. of India.

References