Toward Clustering Persian Vowel Viseme: A New Clustering Approach based on HMM

Maedeh Moghadam  
Faculty of Electrical, Computer and IT Islamic Azad University, Qazvin Branch, Qazvin, Iran  
maedeh.moghadam@gmail.com

Azam Bastanfard  
Faculty of Engineering, Islamic Azad University, Karaj Branch, Karaj, Iran  
bastanfard@kiau.ac.ir

Mohammad Mahdi Dehshibi  
Faculty of Electrical, Computer and IT Islamic Azad University, Qazvin Branch, Qazvin, Iran  
mohammad.dehshibi@yahoo.com

Abstract- This paper sorts out the problem of Persian Vowel viseme clustering. Clustering audio-visual data has been discussed for a decade or so. However, it is an open problem due to shortcoming of appropriate data and its dependency to target language. Here, we propose a speaker-independent and robust method for Persian viseme class identification as our main contribution. The overall process of the proposed method consists of three main steps including (I) Mouth region segmentation, (II) Feature extraction, and (IV) Hierarchical clustering. After segmenting the mouth region in all frames, the feature vectors are extracted based on a new look at Hidden Markov Model. This is another contribution to this work, which utilizes HMM as a probabilistic model-based feature detector. Finally, a hierarchical clustering approach is utilized to cluster Persian Vowel viseme. The main advantage of this work over others is producing a single clustering output for all subjects, which can simplify the research process in other applications. In order to prove the efficiency of the proposed method a set of experiments is conducted on AVAII.

Keywords- Viseme, Visual Speech, HMM

I. INTRODUCTION

A viseme [1] is the visual equivalent of a phoneme in spoken language, and it is the basic visible distinguishable unit of visual speech. This linguistic information has been shown its impact in researches such as speech recognition [2], speech processing [3], computer facial animation [4], and lip synchronization [5].

Preliminary researches in the area of viseme classification is carried out by Binnie et al. [6] and Owens et al. [7]. In these studies, classes are defined by conducting a subjective test. Same subjective test is carried out in Persian language [8] which consonant visemes are classified. Caldognetto et al. [9] were researchers who investigate machine vision techniques for viseme set class identification. They calculate lips height, lips width, and anterior/posterior movements of lips by utilizing a movement analyzer. These features are used to classify a single subject data into three viseme classes. Melenchón et al. [10] use Eigen space for visual speech clustering, and they extract three viseme classes for three speakers. Other researchers consider this theme and apply different methods to classify Chinese [11] and Persian viseme [12]. However, previous methods are subjected to three shortcomings in which the size of database is merely small, different viseme classes are introduced for different subjects, and only some user-selected frames are analyzed instead. The last limitation disturbs temporal nature of visual speech and therefore, the stretch of speech is not considered.

In this paper, we propose a Persian Vowel clustering method in which only one cluster for all subjects. The proposed method considers all lips change during speech and utilizes HMM as a probabilistic model-based feature detector. HMM was applied in one dimension data clustering [13] but applying this model for visual speech clustering was done in this paper for first time.

Vowels are basic unit of language and therefore Persian vowel viseme set class was identified in this paper. Persian vowel consist of: {/a/, /æ/, /e/, /i/, /o/, /u/}

Other parts of this paper are organized as follows: Section II focus on the importance of data corpus. Section III describes mouth segmentation algorithm. Section IV explains how output of HMM is used as extracted features. Hierarchical clustering is seen in section V. Section VI explains experimental result. Finally, the paper is concluded in section VII.

II. DATA CORPUS IMPORTANCE

Audio-visual data corpus has been discussed for a decade. However, it is an open problem due to shortcoming of appropriate data. AVAII [14] data corpus was selected for viseme class identification. Each speaker’s film was analyzed to syllable and saved in different file. Analysis was done using silence between each syllable. The analysis method is based on two simple audio features, namely the signal energy and the spectral centroid. As long as the audio feature sequences are extracted, a simple thresholding criterion is applied in order to determine the silence and
speech interval. Considering this interval, film was cut and saved.

III. MOUTH SEGMENTATION

Since the mouth area includes the most important visual information in speech, mouth segmentation is an important step in viseme class identification. To enhance the presence of the lip in the image, the pseudo-hue component is calculated from the RGB representation for each frame in the video sequence [15]. Pseudo-hue is calculated according to Equation (1). In Equation (1), G(x, y), R(x, y) respectively show green and red in (x, y) coordinates and H(x, y) shows pseudo-hue in that coordinates.

\[
H(x,y) = \frac{R(x,y)}{G(x,y) + R(x,y)}
\]  

Equation (1)

The region around the lips is extracted by applying a histogram-thresholding scheme. The threshold value is adaptively selected as the local minima between the first and the second peak of the pseudo-hue histogram. By applying threshold technique according to Equation (2), the pseudo-hue image would be binaries. In this formula, f(i, j) and g(i, j) is respectively input image and output image and Th is image threshold. Finally, with regard to nose, chin and lip corners information and applying morphological technique, mouth area is extracted. To neutralize the effect of lighting changes, the mouth area image are normalized. The images resulting from the lip segmentation procedure are as shown in Fig. 1.

\[
g(i,j) = \begin{cases} 1 & \text{if } f(i,j) \leq T_h \\ 0 & \text{otherwise} \end{cases}
\]  

Equation (2)

(a) Original RGB Image (b) ROI extracted from the original image (f) Normalized image

IV. FEATURE EXTRACTION

After segmenting the mouth region in all frames, the feature vectors are extracted based on a new look at Hidden Markov. First the dimensionality of the data resulting from the lip segmentation process is reduced and then HMM is utilized as a probabilistic model-based feature detector

A. Dimension reduction

In order to reduce the dimensionality of the data compression techniques are applied to extract the lip-features from each frame in the video sequence. To achieve this goal, Linear Discriminate Analysis (LDA) is applied to obtain a compact representation for all images resulting from the lip segmentation procedure. For dimension reduction using LDA, within-class scatter and between-class scatter for each vowel class is calculated and then separation criteria is got. Within-class scatter is calculated according to Equation (3).

\[
b_w = \sum_{j} (m_j - m)^T S_{jw} (m_j - m)
\]  

Equation (3)

In Equation 3, \( m \) is within class scatter for class \( j \). The between-class scatter is calculated according to Equation (4).

\[
b_b = \sum_{j} N_j \left( m_j - \mu \right)^T \left( m_j - \mu \right)
\]  

Equation (4)

\( m \) is mean of entire dataset’s. \( N_j \) is number of samples in class \( j \). For best separation between classes, LDA should maxim criteria in Equation (5).

\[
C_{riteria} = S_{w}^{-1} \times S_{w}
\]  

Equation (5)

\( S_w \) and \( S_b \) are calculated for all segmented mouth in all frame. By maximum criteria, Eigen values, and Eigen vectors is obtained which called LDA subspace. The \( m \) Eigen vector was selected. Each mouth image is multiplied to these Eigen vector and reduced dimension feature is obtain. These reduced dimension features have maximum similarity to within-class data and minimum similarity to other classes.

B. HMM as feature detector

HMMs can be viewed as stochastic generalizations of finite-state automata, when both transitions between states and generation of output symbols are governed by probability distributions [16]. Fig. 2 shows the structure of a typical left–right HMM. For convenience, we denote an HMM as a triplet (6).

\[ \lambda = \left( A, B, \Pi \right) \]

Equation (6)

\( \lambda \) is state transition probability distribution, \( B \) is observation symbol probability distribution and \( \Pi \) is initial state probability distribution. For extracting features, we follow this algorithm.

- Define a HMM for each six vowels in Persian. These six HMM are identified by \( \lambda_i \), \( 1 \leq i \leq 6 \)
- Train each model \( \lambda_i \) by \( s_i \), \( 1 \leq i \leq 13 \). \( s_i \) is sample data. For example, HMM related to vowel \( /e/ \) is trained by all 13 samples of vowel \( /e/ \). 13th sample is maintained for test.

Figure 2. A left-right HMM with 3 states
Test each model by \( t_i \) \((1 \leq i \leq 6)\). \( t_i \) is test data. For example HMM related to /e/ vowel is tested by /æ/, /æ/, /æ/, /æ/ test data and its probability to generate the test data is evaluated.

By following above algorithm, a six-dimension feature vector is obtained for each vowel.

V. HIERARCHICAL CLUSTERING

In order to identify Persian vowel viseme set class, a hierarchical clustering approach is utilized. In previous step, 6-dimension feature vector was obtained for each vowel. In this step, each feature vector of vowels is compared and the most similar ones are placed in a viseme class.

For applying hierarchical clustering, Euclidean distance between each pairs of feature vector is calculated. Then each pair of feature vector that are in close proximity were linked and form a cluster. Euclidean distance calculates again between new cluster and feature vector. This work is done until all feature vectors will be placed in a single cluster. By pruning branches off the bottom of the hierarchical tree, and assign all the objects below each cut to a single cluster, Persian viseme set class was obtained.

VI. EXPERIMENTAL RESULT

First step in viseme clustering is preparing data corpus. At first, we selected AVAII data corpus. Each speaker pronounced Persian syllables. There was 5-seconds silence between each syllable. Each speaker’s film analyzed into syllable and saved in different files. Vowel’s file was selected from 14 speakers. Result of first step is 14 samples for each vowel. Each vowel’s visual speech has 10 frames in average.

In second step, mouth area is segmented in all 14 samples of six vowels and saved in different file. Segmented mouth is a 60*70-pixel image. Considering that there is 14 samples from six vowels and each sample have 10 frames in average, result of mouth segmentation step is included 840 image in average which each image size is 60*70 pixels.

In feature extraction step, at first each mouth image dimension was reduced by applying LDA technique. Three principle features was extracted. Then six HMM for each vowel was defined. Each model was trained by related sample. For example, HMM related to vowel /e/ is trained by all 13 samples of /e/. Each sample consists of 10 images in average. 14th sample is maintained for test.

After training all models, each model was tested by 14th sample. All vowels of 14th sample were used. For example, HMM related to vowel /e/ is tested by /æ/, /æ/, /æ/, /æ/, /æ/, /æ/ test data and its probability to generate the test data is evaluated. The HMM’s test result is a six-dimension feature vector. Each dimension is HMM likelihood for each /æ/, /æ/, /æ/, /æ/, /æ/, /æ/. More similarity between test data and model will be resulted to a better probability. For example for HMM related to vowel /æ/, the best likelihood is related to /æ/ test data and then the vowel /æ/. The worst one is related to the vowel /æ/. These result means that vowel /æ/ have most similarity to itself at first and then to vowel /æ/. It has the least similarity to vowel /æ/. All models were tested. Result of feature extraction is six feature vectors for each vowel that each feature vector dimension is six.

In last step, Euclidean distance between each pair of feature vector is calculated and most similar feature vector place in a cluster. Clustering result can be seen in Fig 3. Vowels are in horizontal axis and distance between them has specified in vertical axis. By pruning branches off the bottom of the hierarchical tree, and assign all the objects below each cut to a single cluster, Persian viseme class was obtained. Persian vowel viseme can be seen in Table 1. First column of this table show viseme class and second column show vowels, which belong to class. For example viseme class 3 consist of vowel /æ/ and /æ/. This means that these two vowels are so similar in visual dimension and distinguishing them from each other is difficult. Therefore, these two vowels form a unit in visual dimension.

Table1. Viseme classes by applying HMM method on 14 speakers

<table>
<thead>
<tr>
<th>Viseme class</th>
<th>vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viseme1</td>
<td>/æ/</td>
</tr>
<tr>
<td>Viseme2</td>
<td>/æ/</td>
</tr>
<tr>
<td>Viseme3</td>
<td>/æ/, /æ/</td>
</tr>
<tr>
<td>Viseme4</td>
<td>/æ/, /æ/</td>
</tr>
</tbody>
</table>

Proposed method in [10] was implemented for evaluating. Melenchón et al. method used PCA and only 13 frames instead of all frame. They did not explain how these frame is selected. Their method implemented on extended AVAII.

By using only first 13 frames, we obtained 14 different results for each speaker. Melenchón et al. method not only used selected frames but also proposed different results for different speakers. Result of implementing Melenchón et al.
method over speaker 3 and 9 could be seen in Fig3. and Fig4. Then Table2 and Table 3 illustrate Persian viseme classes by applying Melenchón et al. method on speaker 3 and 9.

Table 2. Viseme classes by applying Melenchón et al. method on speaker 3

<table>
<thead>
<tr>
<th>Viseme class</th>
<th>Vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viseme1</td>
<td>i, e, æ</td>
</tr>
<tr>
<td>Viseme2</td>
<td>æ</td>
</tr>
<tr>
<td>Viseme3</td>
<td>o, u</td>
</tr>
</tbody>
</table>

Table 3. Viseme classes by applying Melenchón et al. method on speaker 9

<table>
<thead>
<tr>
<th>Viseme class</th>
<th>Vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viseme1</td>
<td>a</td>
</tr>
<tr>
<td>Viseme2</td>
<td>æ</td>
</tr>
<tr>
<td>Viseme3</td>
<td>o</td>
</tr>
<tr>
<td>Viseme4</td>
<td>e, i, u</td>
</tr>
</tbody>
</table>

Therefore, each group of above vowels separately is a single unit in visual dimension and form a viseme class. /æ/ and /a/ are not similar to any other vowel in visual dimension. Therefore, each one is a unit in visual domain and form a viseme class. For the future work we would like extend our algorithm.

REFERENCES


VII. CONCLUSION

In this paper, we proposed a Persian vowel clustering method in which only one cluster with four classes is generated for all 14 subjects. First class includes vowel /æ/, second class includes /æ/, third class includes vowels (/æ, /æ/) and fourth class includes vowels (/æ, /æ/). Vowels (/æ, /æ/) and vowels (/æ, /æ/) are too similar in visual dimension;