Interlaced Derivation for HINDI phoneme- Viseme recognition from continuous speech

Arpana Mishra

G.L. Bajaj Institute of Technology, Greater Noida, India

Abstract: Visual phoneme (Viseme) is very important unit for lip reading. Statistics, pattern recognition and machine learning has been done with LDA to find the linear combinations of features that characterizes or separates two or more classes of objects or events. Feature extraction and reliable analysis of facial movement makeup an important part in many multimedia systems. viseme analysis is language dependent process. This paper included HINDI language. We proposed an image based approach for lip reading consists three steps

- Extracting the lip region
- Extracting interlaced derivative pattern(IDP) considering co-articulation effect
- Clusterization of viseme in HINDI language by mapping each into its subspace.

Keywords: LDA; linear discriminant analysis; Pseudo-Hue; color intensity; phonemes; viseme; MFC; Mel frequency Cepstrum, DCT; discrete cosine transform.

I. INTRODUCTION

Automatic speech recognition technique can be used in noise free environment but their performance accuracy decreases with noise interference. Audio and visual features of ASR system can be used to improve the performance in noisy environment. These audio visual features also can be used to develop the Hindi phoneme recognition systems. The DCT of all Mel-frequency cestrum on a nonlinear mel scale of frequency can provide better recognition over noisy channel. For speech reading and lip localization, Linear Discriminant Analysis (LDA) in pattern recognition has been used.

II. Automatic speech recognition: The sequence of events that makes any automatic speech recognition software, related to its of its sophistication, pickup and break down the words for analysis and response goes as follows:-

Speak to the software via an audio feed:

1) The device you are speaking to creates a wave file of your words (Transducer)
2) The wave file is cleaned by removing background noise and normalizing volume (signal processing)
3) The resulting filtered wave form is then broken down into are called phonemes,( phonemes are basic building block sounds of language and words. English has 44 of them ,consisting of sound blocks as “wh”, “th” “ka” etc
4) Each phoneme is like a chain link and by analyzing them in sequence, starting from the first phoneme, the ASR software uses statistical probability analysis to deduce whole words and then from there, complete sentences.
5) Now ASR understood your word, respond to you in meaning full way.

In computerized algorithm processing, an active contour model for inside mouth is considered to extract, which can be pursued in various utterances where one can trace points of interest in various utterances to detect lip variation. But it has neglected anterior movement of the lips in its clustering phase. besides, it has low accuracy and high execution time while the iteratin of this algorithm may cause local minimums. Viseme grouping has also been carried out in Swedish, Persian language using maximum likelihood classifier method aimed at sound phoneme articulation, and helping visual information, it has taken coarticulation effect into consideration and it is used video sequences recorded from one woman.

In [14] a hindi talking head application is developed, where English phonemes and visemes are used instead of HINDI ones.

HINDI visemes were neither identified nor classified yet.this cause such products not to be photorealistic.

Then a furthest neighbor of the weight value as a result of reconstruction is set as the criterion for comparing viseme dissimilarity, in order to indicate the robustness of the proposed database. Comparing the results of the clustering algorithm with that of the perceptual test given by an expert proves a reasonable evaluation of proposed algorithm. Viseme is the visual form of phoneme [24].in other words, visemes of some phonemes are alike, as/b/and /p/which are phonemically different, but the same in visual form, however, it should be noted that visemes of a single phonemes are not necessarily the same. in that a phoneme gets various shapes thanks to the influences exerted by its former and latter phonemes ,altogether termed coarticulation. See figure 1.
Figure 1. Different viseme for pronouncing /s/ in (a) /tas/, (b) /tes/, and (c) /tos/
Viseme classification should be done based on the visual information about lip, coarticulation effects, and applications. If the result is provided based on acoustic data, the result could be quite different, as/m/and/n/which are acoustically similar, but unlike in visual appearance.
In this study, considering coarticulation effect (CV and CVC combinations where C stands for Consonants and V for Vowel) which look similar are clustered together.

II. PROPOSED METHOD

The focus of this study is on the CV CVC combinations in HINDI. The proposed approach is based on both linguistic issues and algorithmic process. For linguistic issues, we considered phoneme position in each syllable and hindis viseme classification.

A. Linguistic Issues for Frame Selection

Considering linguistic issues as an algorithmic study distinguishes this study from other algorithmic studies. Phoneme position in speech pattern and co-articulation effects are two important factors in visemes appearance. Lip appearance in a speech pattern is relevant upon its place of articulation, whether it is beginning, in the middle or at end.

B. Lip Localization:

The first step in the proposed method is to localize lip images & face images. Since the speaker has some head movements in video sequence, and the number of need sequence is very large an automatic cropping is utilized to crop lip area.

C. Feature Extraction Using Interlaced Derivative Patterns:

In IDP technique an IDP image is produced from the original image. The IDP image is four-channel derivative image, four directional $n^{th}$ order derivative channels in $0^\circ, 45^\circ, 90^\circ$ and $135^\circ$ respectively.
For an $n^{th}$ order IDP operator the IDP image with four $(n-1)^{th}$ order derivative channels are produced.

These derivative channels present more detailed description of the image in all possible directions (see figure 2).

In IDP, the first order derivatives are calculated in four directions, and these derivatives are threshold with the center value of each directional channel to produce the final IDP. A support vector machine can be used as a classifier.

III. EXPERIMENTAL RESULTS

In order to demonstrate the performance of the proposed method, two sets of experiments are conducted. In the first set IDT was compared with the LDP, LBP, wavelet decomposition, and Kernel PCA with Gaussian and polynomial Kernels. In the second set, results of proposed algorithm were compared with a subjective test. These methods were examined on HINDI audio/visual data corpus [8] in order to achieve the maximum accuracy rate in the clustering and classification.

A. Data Set

Collecting a data corpus in the target language is the first step towards viseme extraction and analysis. We collected AVA, an audio-visual corpus [8] employed in this study. AVA data corpus comprises HINDI syllables, meets the requirements of our target application.
FIGURE: 3 shows six samples of AVA database speakers. (a) /bl/, (b) /vl/, and (c) /ls/

Table: Hindi consonants with phoneme form, letter form, and an example which consists of phonetic of example, the example in Persian script and its translation into English. Ph as phoneme, Le as letter, and Ex as Example.

<table>
<thead>
<tr>
<th>Ph</th>
<th>Le</th>
<th>Ex</th>
<th>Eng Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>/A/</td>
<td>अ, ए</td>
<td>patang (kite)</td>
<td>BUT</td>
</tr>
<tr>
<td>/AA/</td>
<td>आ, ए</td>
<td>aam (mango)</td>
<td>FAR</td>
</tr>
<tr>
<td>/U/</td>
<td>उ, ए</td>
<td>pita (father)</td>
<td>STILL</td>
</tr>
<tr>
<td>/EE/</td>
<td>ई, ए</td>
<td>peela (yellow)</td>
<td>FEE</td>
</tr>
<tr>
<td>/U/</td>
<td>उ, ए</td>
<td>pul (bridge)</td>
<td>BOOK</td>
</tr>
<tr>
<td>/OO/</td>
<td>ऊ, ए</td>
<td>fool (flower)</td>
<td>MOON</td>
</tr>
<tr>
<td>/AI/</td>
<td>एँ, ए</td>
<td>ped (tree)</td>
<td>HAIR</td>
</tr>
<tr>
<td>/AI/</td>
<td>ऐँ, ए</td>
<td>paisa (money)</td>
<td>HAIR</td>
</tr>
<tr>
<td>/I/</td>
<td>इ, ए</td>
<td>pita (father)</td>
<td>STILL</td>
</tr>
<tr>
<td>/EE/</td>
<td>ई, ए</td>
<td>peela (yellow)</td>
<td>FEE</td>
</tr>
<tr>
<td>/I/</td>
<td>इ, ए</td>
<td>pita (father)</td>
<td>STILL</td>
</tr>
<tr>
<td>/EE/</td>
<td>ई, ए</td>
<td>peela (yellow)</td>
<td>FEE</td>
</tr>
<tr>
<td>/I/</td>
<td>इ, ए</td>
<td>pita (father)</td>
<td>STILL</td>
</tr>
<tr>
<td>/EE/</td>
<td>ई, ए</td>
<td>peela (yellow)</td>
<td>FEE</td>
</tr>
<tr>
<td>/I/</td>
<td>इ, ए</td>
<td>pita (father)</td>
<td>STILL</td>
</tr>
<tr>
<td>/EE/</td>
<td>ई, ए</td>
<td>peela (yellow)</td>
<td>FEE</td>
</tr>
</tbody>
</table>

B. Evaluation of feature extraction method for clustering:

In order to evaluate the efficiency of IDP, the extracted features are compared the LDP, LBP, wavelet decomposition, Kernel PCA and Gaussian and polynomial. Resultant comparison has been shown in figure 4.
IV. CONCLUSION

This paper is concluded for Hindi language’s visemes. Visemes are classified with the proposed algorithm, having speech therapy applications and photo realistic speaking animation in target. Some phoneme coarticulation position in syllables are considered. Pictures are captured of two female’s responses; the first one who is aware of sound speech rules is used in viseme clustering. We rationally reduced the dimensions of images by applying IDP to each of the visemes. Then the weight criterion out of the construction of each viseme with the other is used for quantifying visemes’ dissimilarity through utilizing unweighted pair group method with arithmetic mean. Two evaluation procedures were considered for verifying our work. The algorithm was entirely applied to Hindi speakers so as to check the robustness of the feature extraction method as well as clustering and classifying approaches compare with the state of art methods. Comparing the results of the proposed algorithm with an expert speech therapist indicated the accuracy of this method.

References


