Arpana Mishra et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 4, Issue 3, Sept 2017, pp. 172-176

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 persued in various utternances where one can trace points of interest in various utterances to detect lip variation. But it has neglected anteriour movements 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 mininums. 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 exterted 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 hindi viseme classification.

A.Linguistic Issues for Frame selection

Considering linguistic issues as an algorithm prerequisite distinguishes this study from other algorithmic study .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:

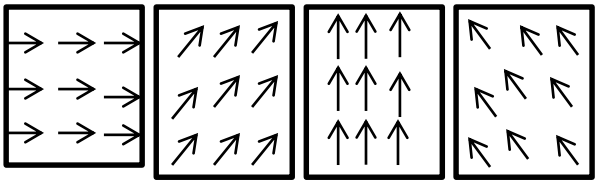
In IDP technique an IDP image is produced from the original image.

The IDP image is four-channel derivative image,four directional n^{th} order derivative channels in 0^o,45^o,90^o and 135^o,respectively.

For an nth 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).

Z_1	Z_2	Z_3
Z_8	Z_0	Z_4
Z_7	Z_6	Z_5



Channel 0^0 Channel 45^0 Channel 90^0 channel135^0

Figure2. (a)A 3X3neighborhood around a pixel.(b)four directional derivative channel in the IDP image**.**

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.

(a)

(b)

(c)

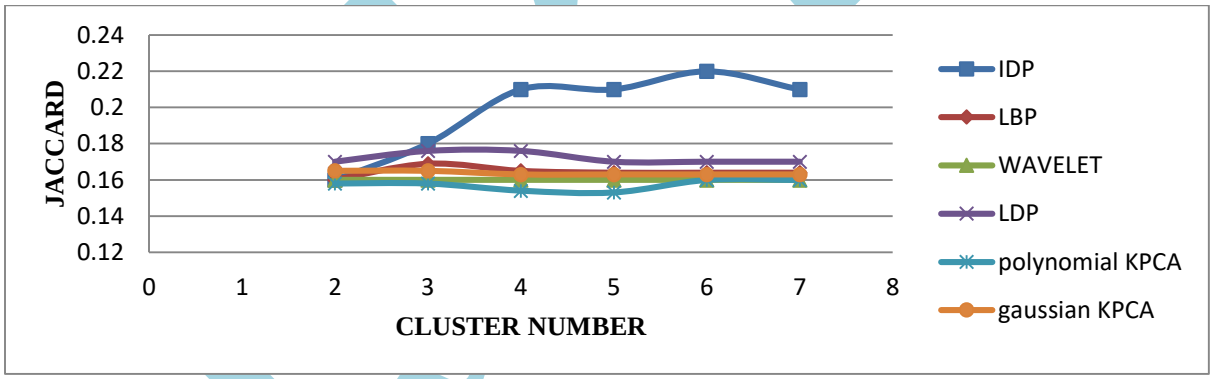
FIGURE: 3 shows six samples of AVA database. speakers.
(a) /b/, (b) /v/, and (c) /s/

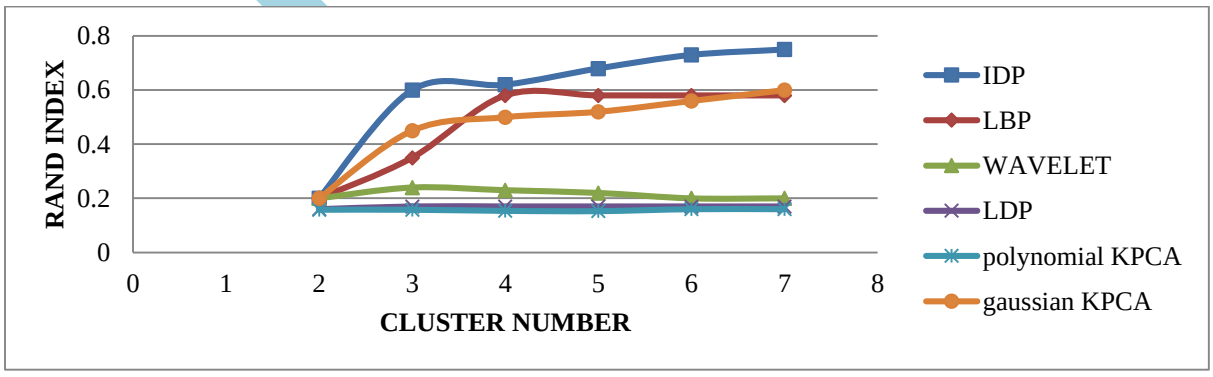
Table:Hindi consonants with phoneme form ,letter form and an example which consists of phonetic of example,the example in Persian script and its translation into English. Ph as phoneme,Le as letter,and Ex as Example

Ph	Le	Ex	Eng Equivalent
/A/	अ प	patang(kite)	BUT
/AA/	आ□□	aam(mango)	FAR
/I/	इ, □□	pita(father)	STILL
/EE/	ई, □□	peela(yellow)	FEE
/U/	उ, □□	pul(bridge)	BOOK
/OO/	ऊ□□	fool(flower)	MOON
/E/,/A/	ए, □□	ped(tree)	HAIR
/AI/	ऐ, □□	paisa(money)	HAIR
/O/	ओ□□	POSCHA(wiper)	OR
/AU/	औ□□	aurat(woman)	OR
/ANG/	□	angoor(grapes)	NASAL VOWEL FAUN
/AN/	□	chanda(moon)	JUNGLE
/AH/	□		AHEAD

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, Kernal PCA and Gaussian and polynomial. Resultant comparison has been shown in figure 4.





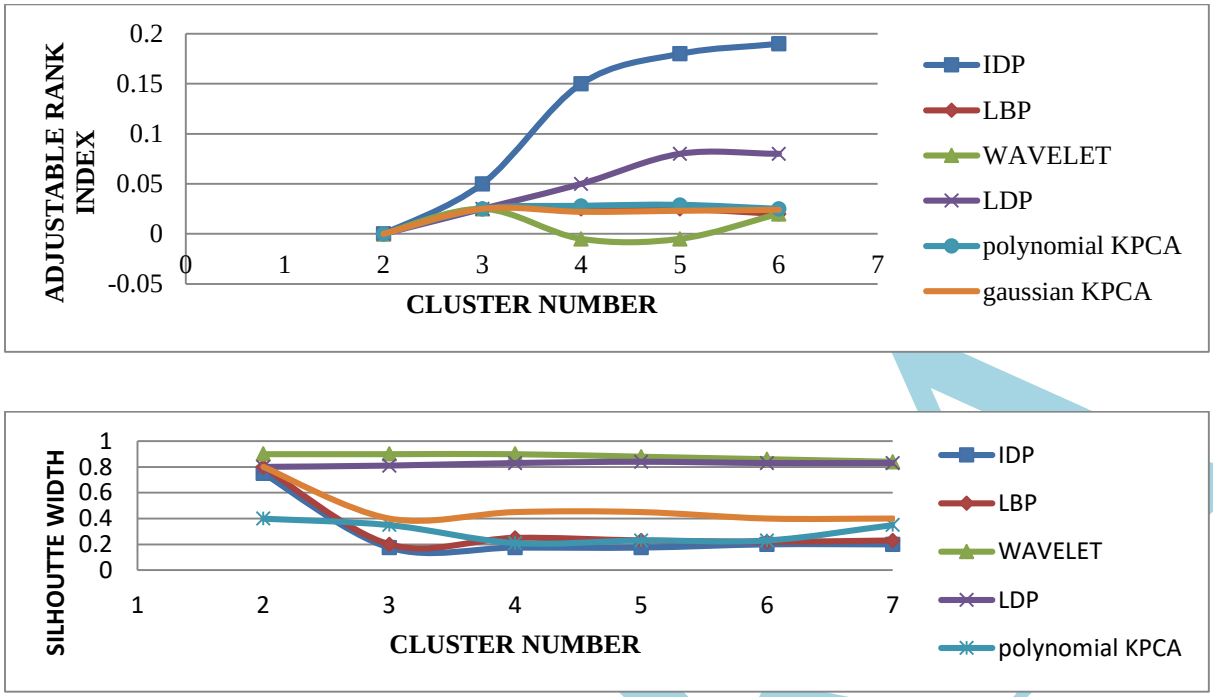


Figure 4.Comparing IDP with different feature extraction method. Four results shows the clustering property of IDP.

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 ofthe visemes. Then the weight criterion out of there construction of each viseme with the other is use 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 the art methods. Comparing the results of the proposed algorithm with an expert speech therapist indicated the accuracy of this method.

References

[1]. "A wavelet-based framework for face recognition," C. Garcia, et al., Proc.Int. Workshop on Advances in Facial Image Anal. RecognitionTechnology, 5th European Conf. Computer Vision, Citeseer, 1998.

[2]. "Facial family similarity recognition using Local Gabor Binary Pattern Histogram Sequence," M.M. Dehshibi, et al., Proc. Hybrid Intelligent Systems (HIS), 2012 12th International Conference on, IEEE, 2012, pp.219-224

[3]. "Portability: A New Challenge onDesigning Family Image Database," M.M. Dehshibi and A. Bastanfard, Proc. IPCV, 2010, pp. 270-276.

[4]. "A new algorithm for age recognition from facial images," . M.M. Dehshibi and A. Bastanfard, Signal Processing, vol. 90, no. 8, 2010, pp. 2431-2444.

[5]. "Generic Visual Recognition on Non-Uniform Distributions Based on AdaBoost Codebooks," M.M. Dehshibi and S.M. Alavi, Proc.International Conference on Image Processing, Computer Vision, and Pattern Recognition, 2011, pp. 1046-1051.

[6]. "Unsupervised Feature Based Facial Family Similarity Recognition," M.M. Dehshibi and A. Bastanfard, Proc. International Converence on Image and Video Processing and Computer Vision (IVPCV-10), ISRST, 2010, pp. 132-138.

[7]. "LPT: Eye Features Localizer in an NDimensional Image Space," M.M. Dehshibi, et al., Proc. IPCV, 2010, pp. 347-352.

[8]. "AUT-Talk: a farsi talking head," R. Safabakhsh and F. Mirzazadeh, Proc. Information and Communication Technologies, 2006. ICTTA'06. 2nd IEEE, 2006, pp. 2994-2998.

[9]. "Graphical speech training system for hearing impaired," K. Resmi, et al.Proc. Image Information Processing (ICIIP), 2011 International Conference on, IEEE, 2011, pp. 1-6.

[10]. "Linear principal transformation: toward locating features in N-dimensional image space," M.M. Dehshibi, et al., Multimedia Tools and Applications, vol. 72, no. 3, 2014, pp. 2249-2273.

[11]. "Linear principal transformation: toward locating features in N-dimensional image space," M.M. Dehshibi, et al., Multimedia Tools and Applications, 2013, pp. 1-25.

[12]. "Lip localization and viseme classification for visual speech recognition," S. Werda, et al., arXiv preprint arXiv: 1301.4558, 2013.

[13]. "Viseme & Frame rate analysis for multimedia applications to assist speech reading," . J.J. Williams, et al., Journal of VLSI signal processing.

[14]. "Support-vector networks," C. Cortes and V. Vapnik, Machine learning,vol. 20, no. 3, 1995, pp. 273-297.

[15]. "Audio-visual phoneme classification for pronunciation training applications," H. Kjellström, et al., Proc. INTERSPEECH, 2007, pp. 702-705.

[16]. "Viseme analysis for speech-driven facial animation for Czech audio-visual speech synthesis," Z. Krňoul, et al., SPECOM 2005 proceedings, 2005.

[17]. "Confusions among visually perceived consonants," . C.G. Fisher, Journal of Speech, Language, and Hearing Research, vol. 11, no. 4, 1968, pp.796-804.

[18]. "Gender classification using interlaced derivative patterns," A. Shobeirinejad and Y. Gao, Proc. Pattern Recognition (ICPR), 2010 20th International Conference on, IEEE, 2010, pp. 1509-1512.

[19]. "Face description with local binary patterns: Application to face recognition," T. Ahonen, et al., Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, no. 12, 2006, pp. 2037-2041.

[20]. "Objective viseme extraction and audiovisual uncertainty: estimation limits between auditory and visual modes," J. Melenchón, et al., Proc. AVSP, 2007, pp. 13.

[21]. "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vsall schemes," M. Galar, et al., Pattern Recognition, vol. 44, no. 8, 2011, pp. 1761-1776.

[22]. "Nonlinear component analysis as a kernel eigenvalue problem," B. Schölkopf, et al., Neural computation, vol. 10, no. 5, 1998, pp. 1299-1319.

[23]. "Audio-visual speech recognition in vehicular noise using a multi-classifier approach," H. Karabalkan and H. Erdoğan, 2007 systems for signal, image and video technology, vol. 20, no. 1-2, 1998,pp. 7-23

[24]. "EigenCoin: sassanid coins classification based on Bhattacharyya distance," Proc. International Conference on Information Technology, AWERProcedia Information Technology & Computer Science, 2012, pp. 1151-1160.

[25]. "Robot interactions using speech synthesis and recognition with lip synchronization," R.C. Luo, et al., Proc. IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society, IEEE, 2011, pp. 171-176.

