www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN ARCHIVES OF ACOUSTICS
Vol. 40, No. 4, pp. 609–619 (2015)

DOI: 10.1515/aoa-2015-0061

Comparative Study of Visual Feature for Bimodal Hindi Speech Recognition

Prashant UPADHYAYA[(1)], Omar FAROOQ[(1)], M.R. ABIDI[(1)], Priyanka VARSHNEY[(2)]

[(1)] *Department of Electronics, AMU-Aligarh*
India; e-mail: upadhyaya.prashant@rediffmail.com, omarfarooq70@gmail.com, abidimr@rediffmail.com

[(2)] *Mindz Technology*
New Delhi, India; e-mail: priyankavarshney@gmail.com

(*received July 22, 2015; accepted September 28, 2015*)

In building speech recognition based applications, robustness to different noisy background condition is an important challenge. In this paper bimodal approach is proposed to improve the robustness of Hindi speech recognition system. Also an importance of different types of visual features is studied for audio visual automatic speech recognition (AVASR) system under diverse noisy audio conditions. Four sets of visual feature based on Two-Dimensional Discrete Cosine Transform feature (2D-DCT), Principal Component Analysis (PCA), Two-Dimensional Discrete Wavelet Transform followed by DCT (2D-DWT-DCT) and Two-Dimensional Discrete Wavelet Transform followed by PCA (2D-DWT-PCA) are reported. The audio features are extracted using Mel Frequency Cepstral coefficients (MFCC) followed by static and dynamic feature. Overall, 48 features, i.e. 39 audio features and 9 visual features are used for measuring the performance of the AVASR system. Also, the performance of the AVASR using noisy speech signal generated by using NOISEX database is evaluated for different Signal to Noise ratio (SNR: 30 dB to −10 dB) using Aligarh Muslim University Audio Visual (AMUAV) Hindi corpus. AMUAV corpus is Hindi continuous speech high quality audio visual databases of Hindi sentences spoken by different subjects.

Keywords: Aligarh Muslim University audio visual corpus, AVASR, bimodal, DCT, DWT.

1. Introduction

Automatic speech recognition (ASR) is the most ensembles' technology, which provides easy accessibility for man-machine communication. It has shown significant improvement in man-machine interaction, but the performance of ASR degrades when working under noisy environment (Potamianos, Neti, 2003). Therefore, there is a need of robust technique which can reduce the effect of noisy background condition and improve the ASR performance.

Addition of visual information not affected by noise for enhancing the robustness in ASR is reported in (Chen, 2001), where mouth height and width were selected as visual features. In (Potamianos, Neti, 2001) 24 DCT coefficients were selected as a visual feature from the region of interest (ROI), i.e. selecting a speaker's mouth as ROI and 24 MFCC coefficients as audio features. Both features were concatenated to form a single feature vector, which reported an improvement in SNR of 61% over audio only processing. It is reported by Huang *et al.* (2004) that extraction of visual information from full face video was difficult due to variations in pose, lighting and background conditions; therefore, selection of the speaker's mouth as the region of interest can be used and may provide improved recognition rate.

In another work by Carboneras *et al.* (2007), the authors reported that the visual features added to the audio features generally resulted in a small gain in accuracy. They performed experiments in two phases. In the first phase of their experiment a simple feature fusion was performed, in which 128 DCT coefficient along with 39 audio coefficients, i.e. MFFCs + Δ (delta) + $\Delta-\Delta$ (delta-delta) were selected. Due to high dimensionality and improper modelling with hidden Markovs model (HMM), it resulted in a poor recognition rate. In the second phase, the same experiment was performed with 16 low dimension DCT coefficients, which results in better recognition, i.e. outperforming the DCT feature by 2–3% with respect to 128 DCT coefficients. Experiment on phoneme recognition by Ahmad *et al.* (2008) reported that features using Linear Discriminate Analysis (LDA) perform well when using high en-

ergy coefficients. In their experiment they used DCT and DWT based visual features. Results of their experiment (with using features in different frequency region) indicated that intermediate frequencies are more informative for speech recognition than lower frequencies. Seymour *et al.* (2008) reported that robustness in AVASR system can be achieved by adding the dynamic visual features. An experiment was performed over different image transformed, i.e. DCT, Fast Discrete Curvelet Transform (FDCT), PCA and LDA. Seymour *et al.* (2008) reported that adding delta (Δ) feature to static feature resulted in a reduction in word error rate (WER) of 12.9% for DCT, 8.7% for FDCT, 9.4% for PCA and 8.1% for LDA over static feature only. A work based on audio-visual Hindi phoneme recognition was reported in Upadhyaya *et al.* (2012), where an experiment was performed using three viseme classes. In their experiment 13 audio features using MFCC and 2-D DCT based visual features were selected for the experiment. It was reported by Upadhyaya *et al.* (2012) that adding the visual information outperforms the recognition rate, especially under the noisy background condition and increase in recognition rate can be achieved by using fewer visual coefficients. The extended work of this work was reported in Varshney *et al.* (2014) in which Hindi viseme classes were increased from three to five. Recently, the work reported in (Upadhyaya *et al.*, 2013; 2014) on Hindi speech proved, that addition of dynamic visual features plays an important role in deciding the robustness of AVASR system. An overall improvement of 26.04% in word recognition is achieved with 12 low dimensional visual (LDV) DCT feature. Another approach was given by Zhou *et al.* (2014) in which latent variable model (LVM) was used to learn the compact representation of visual feature. It provides a model structure of image sequences of the same utterance by a path graph and incorporates the structural information through using the low-dimensional curve.

Due to the limited availability of audio-visual database there has been only few research in audio-visual speech processing because testing and verification of any algorithm is a difficult task. Few database reported in literature are: IBM ViaVoiceTM audio-visual (VVAV) database (Neti *et al.*, 2000); Extended M2VTS (XM2VTS) (Cardinaux *et al.*, 2003); Clemson University Audio Visual Experiments (CUAVE) (Patterson *et al.*, 2002); VidTIMIT database (Sanderson, Paliwal, 2004); TCD-TIMIT (Naomi, Eoin, 2015); AMUAV corpus (Upadhyaya *et al.*, 2013).

In this paper, we compare the three different images transformed based (DCT, PCA, and DWT) visual feature available in the literature (Potamianos, Neti, 2001; Carboneras *et al.*, 2007; Ahmad *et al.*, 2008; Seymour *et al.*, 2008). Four sets of visual feature based on Two-Dimensional Discrete Cosine Transform feature (2D-DCT), Principal Component Analysis (PCA), Two-Dimensional Discrete Wavelet Transform followed by DCT (2D-DWT-DCT) features and Two-Dimensional Discrete Wavelet Transform followed by PCA (2D-DWT-PCA) feature are reported. The audio features are extracted using Mel Frequency Cepstral Coefficients (MFCC) followed by static and dynamic feature. Overall 48 features, i.e. 39 audio features and 9 visual features, are reported for measuring the performance of the AVASR system under noisy background conditions. Performance of noisy speech signals using NOISEX (Varga, Steeneken, 1993) database is evaluated for different Signal to Noise ratio (SNR: 30 dB to -10 dB) using AMUAV corpus for 10 subjects. Different acoustic environments (white Gaussian noise, car noise, babble noise, factory noise and machine gun noise) are considered.

The rest of the paper is organized in following way: Sec. 2 deals with the importance of the Hindi language for audio-visual speech recognition system; Sec. 3 deals with feature selection techniques and proposed method for extracting the audio and visual feature for AVASR system; Sec. 4 and Sec. 5 deal with results analysis and conclusions, respectively.

2. Significance of Hindi language

Hindi language is the fourth most spoken language by number of native speakers, followed by Mandarin, Spanish and English as reported in IPA (International Phonetic Alphabet) (http://www.internationalphoneticalphabet.org). Recently, Hindi language has become more popular worldwide and that is the reason that most of speech enable technologies are building the Hindi speech interface system. Hindi as a language contains more number of phone sets than English language (Neti *et al.*, 2002). Hindi language consists of 64 phone sets, out of which 39 phone sets are common in English language. Another issue is that the International Phonetic Alphabet (IPA) has defined the phone set for labelling the speech data, but there are some sounds which are not included in IPA, i.e. DN, DXX, AWN, (Neti *et al.*, 2002) but they play an important role while building the phone model which is used for speech recognition purpose. Hindi language contains more number of fricatives which have very similar characteristics of noise. Therefore, it is very difficult to recognize speech signal under noisy environment. That is why identification of the robust feature for Hindi language has provided the opportunity for the research to work on this native language and to enhance the performance of ASR in noisy environment.

The major work in the area of speech recognition for Hindi language has been carried out at Tata Institute of Fundamental Research (TIFR), Mumbai (Samudravijaya, 2004). Some research for Hindi

speech is being reported in (Chourasia *et al.*, 2007; Farooq *et al.*, 2010; Mishra *et al.*, 2011; Pradhan *et al.*, 2012; Upadhyaya *et al.*, 2012; 2013). For this research work Hindi language has been chosen as the benchmark due to the fact that it is spoken by a large number of people internationally and very little work has been carried out on audio-visual Hindi speech recognition. The bimodal Hindi speech database (AMUAV) is being developed at Aligarh Muslim University-Aligarh, India for research purpose. AMUAV corpus (Upadhyaya *et al.*, 2013) is a Hindi continuous speech high quality audio and video database which contains 100 speakers. Each speaker in AMUAV corpus recorded 10 sentences out of which 2 sentences are common to all speakers. Recordings have been made in realistic conditions for testing robust audio visual schemes. The video was recorded at 640×380 resolutions with 25 fps in full colour. The audio was recorded in 16-bit stereo at 44.1 kHz. Hindi sentence used in the AMUAV corpus is phonetically balanced. Still more work on AMUAV corpus is under development and soon it will be available publicly for researchers working in the field of Hindi speech.

3. System description for Hindi AVASR

AVASR system consists of two important units: front-end unit and back-end unit (Abdelaziz *et al.*, 2015). The main purpose of the front-end unit is preprocessing and feature extraction. The back-end unit is used for training and classification purpose. Preprocessing is used to reduce the effect of background noise, characteristic of recording device and channel noise. Feature extraction is a dimensionality reduction stage, which tries to extract relevant features that are useful in separating different pattern/classes.

Some commonly used audio feature extraction techniques reported in literature are PCA (Seymour *et al.*, 2008); Linear Predictive Coefficients (LPC) (Lokesh, Balakishnan, 2012); Perceptual Linear Prediction (PLP) (Hoenig *et al.*, 2005); DWT (Navnath, Raghunath, 2012); Wavelet Based Features (Farooq *et al.*, 2005) and MFCC (Potamianos, Neti, 2003). MFCC is commonly used technique in ASR which uses auditory filter-bank structure with a cosine transform having a frequency separation roughly similar to the auditory system (Potamianos, Neti, 2003).

In this paper, audio features are extracted using MFCC. The relation between Mel frequency and frequency is shown in Eq. (1):

$$\text{mel}(f_m) = 2595 \times \log_{10}\left(1 + \log\frac{f}{700}\right). \qquad (1)$$

To include the temporal evolution of MFCC, additional feature Δ (delta) and $\Delta-\Delta$ (delta-delta) is computed. Finally, these feature vectors are concatenated to form as a single modality, i.e. $d_a = 39$. Detailed description of audio feature extraction can be found in (Upadhyaya *et al.*, 2014).

On the other hand, for visual front, preprocessing is done to select the required portions of a frame extracted from a video which is useful in speech recognitions, i.e. selecting a region of interest (ROI) from which the visual features are to be extracted. In case of visual speech recognition, ROI in many research (Potamianos, Neti, 2001; Carboneras *et al.*, 2007; Ahmad *et al.*, 2008; Seymour *et al.*, 2008, Upadhyaya *et al.*, 2012; 2013) is the area around the lip region. Some commonly used visual feature extraction techniques reported in literature are Geometric based and Model based approach (Neti *et al.*, 2000; Potamianos *et al.*, 2004). Both techniques require the inner and outer lip contour for feature extraction.

Procedure for extraction of visual features is shown in Fig. 1. Lip movements are more useful in conveying information. Hence, lip region is selected as the region of interest (ROI). Initially, recorded video is split into frames. For faster processing three dimensional RGB image is converted into gray scale image. Face localization on gray image is selected by using a template matching method (Lee, Park, 2008), which returns the four co-ordinate points representing the face position, as shown in Fig. 2.

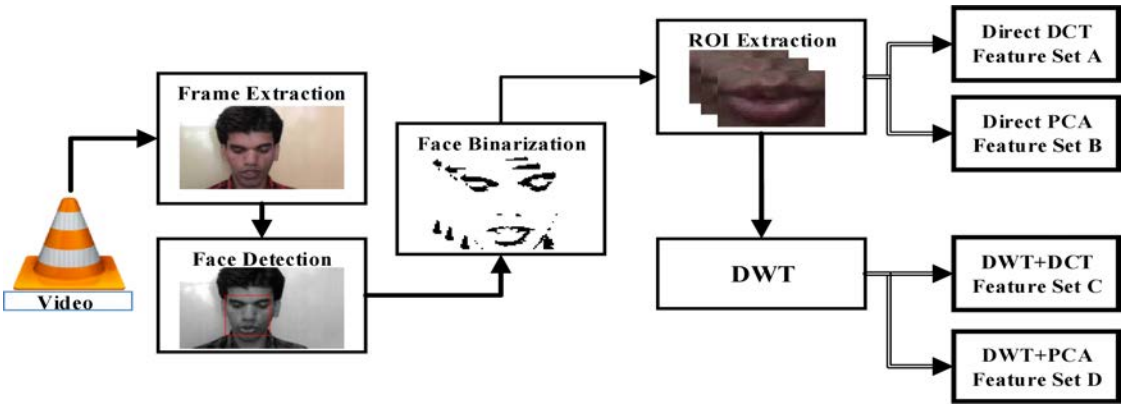

Fig. 1. Visual feature extraction technique used in AVASR.

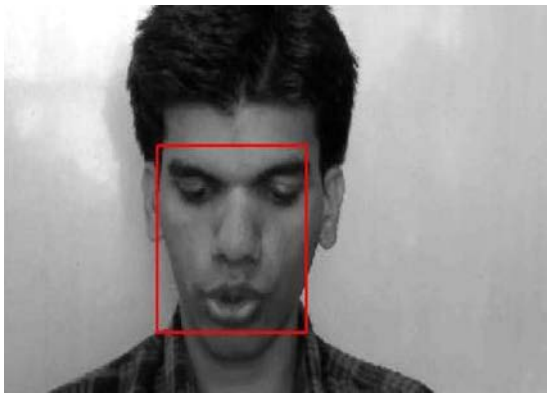

Fig. 2. Face detection from frame.

For ROI extraction, a gray scale image is converted into the binary image (Khanam *et al.*, 2010) having black and white pixels. Lip localization is chosen by counting the black pixel in the binary image. The centre of the image was chosen as the reference and then vertical and horizontal black pixel density histogram was evaluated. Finally, for obtaining robust visual features from the extracted ROI, different feature extraction technique, based on Set A (2D-DCT), Set B (PCA), Set C (2D-DWT-DCT) and Set D (2D-DWT-PCA) are applied. For extracting visual feature for Set A, the two dimensional DCT (Potamianos *et al.*, 2003) was applied on the ROI and DCT coefficients are obtained using Eq. (2) from the lowest frequency component to highest frequency component:

$$X_{ij} = a_i a_j \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} Y_{m,n} \cos \frac{\pi (2m+1) i}{2M} \cdot \cos \frac{\pi (2n+1) j}{2N}, \tag{2}$$

where

$$a_i = \begin{cases} 1/\sqrt{M}, & i = 0, \\ \sqrt{2/M}, & 1 \le i \le M-5 \end{cases}$$

and

$$a_j = \begin{cases} 1/\sqrt{N}, & j = 0, \\ \sqrt{2/N}, & 1 \le j \le N-5. \end{cases}$$

The 2-D DCT of an image returns the same size 2D matrix coefficients. However, it is found that most of the energy and discriminatory information correspond to low frequencies (Seymour *et al.*, 2008; Ahmad, 2010). Therefore, visual feature coefficient, i.e. $X_{11}, X_{12}, \ldots, X_{[M-1][N-1]}$ are obtained, order of row and column is represented by n and m respectively. Finally, 9 visual features coefficients are selected in a zigzag pattern starting from lowest frequencies where usually most of the information is confined.

For extracting visual feature for Set B, the Principal Component Analysis (PCA) (Smith, 2002; Potamianos *et al.*, 2003) was applied on the ROI. PCA is a variable reduction procedure and it is useful when obtained data have some redundancy. PCA is used to reduce the dimensionality of the data by retaining as much variation as possible with original data sets. In our case PCA is applied on the ROI, which returns the principal component coefficients. Figure 3 shows the step by step procedure for obtaining the Set B feature.

Initial step is to compute the 2-dimension mean vector (x, y) from extracted ROI. Further, covariance matrix of 2-dimensional data is computed. Finally, the eigenvalues and eigenvectors of the covariance matrix are computed and sorted from higher to lower eigenvalues. New feature vector is formed by taking n-eigenvectors and forming a feature vector matrix as shown in Eq. (3).

$$\text{Feature Vector} = (\text{eig}_1, \text{eig}_2, \ldots, \text{eig}_n). \tag{3}$$

After obtaining feature vector, visual feature is obtained by taking the transform of feature vector and multiplying it on left of the original data set, as shown in Eq. (4). Total nine visual features coefficients are selected from the Eq. (4).

$$\text{Visual Feature} = \text{row(Feature Vector)} \times \text{row(Data Adjust)}. \tag{4}$$

For obtaining Set C and Set D visual features, extracted ROI was further reduced by one-level 2D-DWT

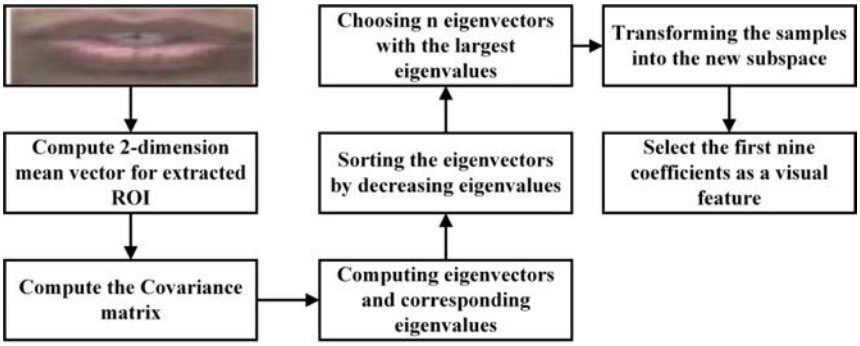

Fig. 3. PCA feature extraction for ROI.

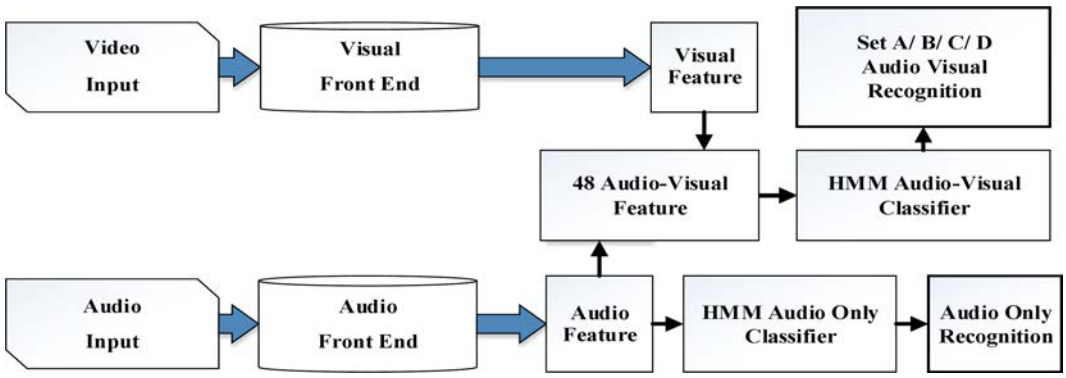

Fig. 4. Block diagram of AVASR.

using "Haar" as a mother wavelet. Haar wavelet is selected as the basis function as it is the simplest wavelet transform with compact support (Bruce *et al.*, 2002; Gundimada, Asari, 2004). Haar Wavelet Transform is mainly used for image compression and feature extraction and requires simple mathematical calculation in terms of addition and subtraction. So they are faster to compute transformation (Bruce *et al.*, 2002). The approximated coefficients which carry the majority of discriminatory information were transformed using 2D-DCT. Set C, visual features are obtained through the hybrid combination of 2D-DWT-DCT techniques. Finally, nine lowest frequency components were selected as visual features for Set C. Similarly, Set D visual features are obtained through the hybrid combination of 2D-DWT-PCA techniques. The approximated coefficients were transformed by applying PCA analysis. Finally, first nine principal components were selected as visual features for Set D.

For making the fair comparison between visual feature extraction techniques for Set A, B, C and Set D, the same number of visual features is taken into account for easy analysis. Finally, these features are concatenated using early integration (Potamianos *et al.*, 2003; 2004) techniques, i.e. audio and visual features are concatenated at initial phase before passing through the classifier. Therefore, in our experimental work thirty nine audio features ($d_a = 39$) and nine visual ($d_v = 9$) features were selected to form a 48 audio-visual feature ($d_{av} = 48$). Figure 4 shows the complete procedure for evaluating the features for AVASR systems. Finally, these features are passed through the classifier and overall recognition, for audio only recognition and audio-visual recognition for Set A, B, C and D is evaluated respectively. MFCC features are taken as a baseline feature.

4. Result analysis

The experiment was performed in Matlab version 7.12.0.635 (R2011a). Hidden Markovs model (Young, 2008) was used by calling C library module of selected HTK version 3.4.1. Three states left-right HMM model along with the one state silence model were used for modelling phoneme. The percentage of the correctly recognized words (C), and percentage of the word accuracy (W_{acc}) were computed using Eq. (5) and (6), respectively:

$$\text{Percent Correct}(C) = \frac{N-D-S}{N} \times 100, \quad (5)$$

$$\text{Percent Word Accuracy}(W_{\text{acc}}) = \frac{N-D-S-I}{N} \times 100, \quad (6)$$

where N is the total number of evaluated words, D is the total number of deletions, S is the total number of substitutions, and I is the total number of insertion error.

Total of 100 sentences with 1225 words was used for evaluation of performance. A bi-gram language model was created based on the transcriptions of the training data set. Recognition was performed using the Viterbi decoding algorithm, with the bi-gram language model. The same training and testing procedure was used for both audio-only and audio-visual automatic speech recognition experiments.

During recognition process, grammar scale factor and the word insertion penalty are used for controlling the influence of the language model over the computed probabilities. Insertion penalty is a fixed cost added to each token when it transmits from the end of one word to the beginning of the next. Grammar scale factor is the amount by which the language model probability is scaled before being added to each token as it transits from the end of one word to the beginning of the next. To make the fair comparison word insertion penalty and grammar scale factor were kept same, i.e. 0.0 and 5.0, respectively. To test the algorithm over a wide range of SNRs, noise was added to the audio signals. The experiments for SNRs from 30 dB to −10 dB were performed. Five types of noise were selected from NOISEX database, i.e. white noise, car noise, babble noise, factory noise and machine gun noise. These noises were

injected in a clean audio signal to produce the noisy environmental conditions. To match the bandwidth of speech signal and noisy signal, speech signal was down sampled to 8 kHz. As in speech signal most of the energy is concentrated up to 8 kHz. Each audio signal was then mixed with the noisy signal at a different SNRs range, i.e. from 30 dB to −10 dB.

4.1. Performance of audio only recognition under clean and adverse environment

Percentage of word accuracy for audio features along with four different categories of visual features under clean environment condition is shown in Fig. 5a. Word accuracy (W_{acc}) for clean signal is found to be 96.41% whereas for others features it is 75.02%, 95.18%, 86.53% and 93.39% for Set A, B, C and Set D, respectively under clean environment conditions. Inclusion of Set A shows a 21.39% drop in W_{acc}, where 9.88% drop in W_{acc} is reported for Set C compared to baseline feature. W_{acc} reduces due to the addition of the high dimensionality DCT visual feature which results in inadequate modelling with HTK (Carboneras *et al.*, 2007; Upadhyaya *et al.*, 2013). On the other hand, little improvement in word accuracy is seen for Set C over Set A features. Hybrid combination of DCT+DWT results in better compact structure, i.e. it contains the higher energy coefficients, and DWT allows the better localization of the signal. Set B and Set D visual features show the performance close to that of the baseline feature. Addition of visual feature with clean audio has not shown any improvement over baseline feature (MFCC) in clean audio condition. Set B shows the best performance for audio visual features. Figure 5a shows the percentage recognition of audio only under noisy acoustic conditions. There is a drop in (W_{acc}) for audio only recognition for 10 dB to −5 dB SNR. Results obtained from Fig. 5b clearly shows clean audio signal distorted more due to the presence of white noise.

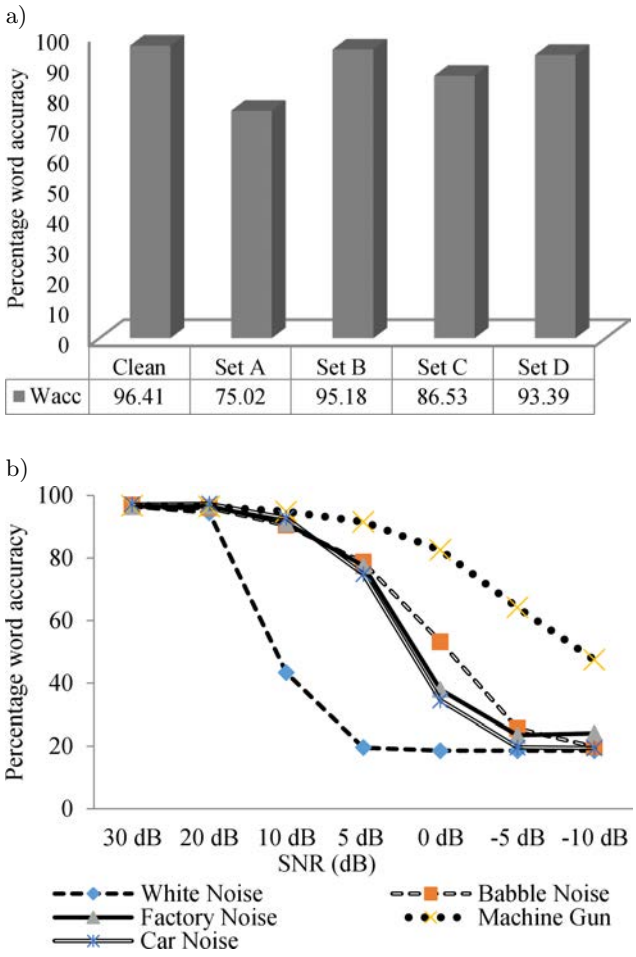

Fig. 5. a) Recognition under clean environment conditions, b) audio only recognition under noisy environment conditions.

4.2. Performance analysis of audio only versus audio visual recognition for white noise

Figure 6 shows the performance in terms of word accuracy for audio only and audio-visual using four visual sets. From the graph (Fig. 6) it can be ob-

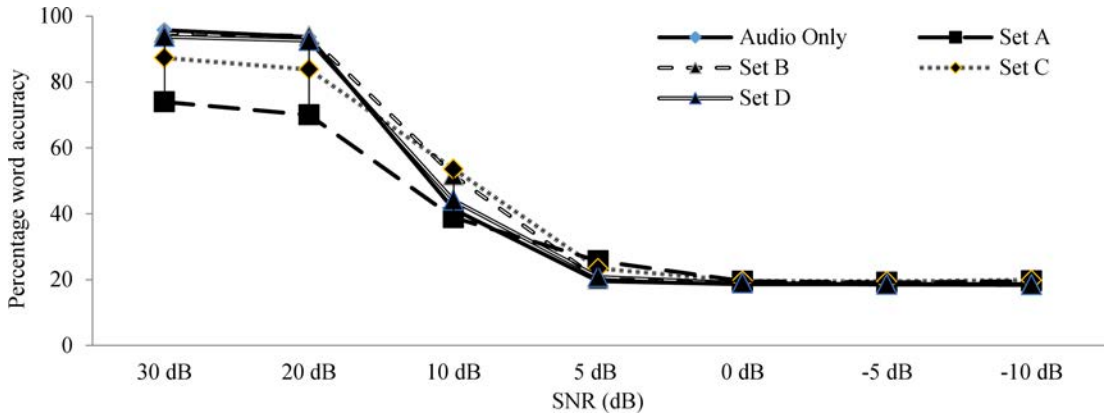

Fig. 6. Percentage of word recognition for audio only and audio visual speech recognitions using white noise.

served that concatenation of audio-visual features at 30 dB has not outperformed baseline feature. Addition of Set A feature with audio feature reported gain in accuracy from 5dB downward, whereas Set B has shown the improvement for all levels of SNR, i.e. 20 dB to −10 dB. Addition of Set C and Set D features, also reported the improvement in noisy condition for SNR value ranging from 10 dB to −10 dB.

The result obtained from Fig. 6 concludes that the addition of visual feature provides gain in accuracy in the presence of white noise. In Table 1 the detailed error distribution of the four feature sets in noisy background (white noise) along with clean acoustic conditions is shown. Table 1 indicates that for white noise, deletion error (D) increases and insertion error (I) decreases with the decrement of SNR.

For Set A feature, both substitution errors (S) as well as deletion error (D) increase simultaneously, with a decrease in SNR. On the other hand, insertion error (I) decreases with decrease in SNR, with the baseline feature (MFCC). Table 2a shows the sum of the error distribution (S+D+I) for white noise, which clearly indicates total error distribution for Set A, exceed the baseline, i.e. total sum error of MFCC at 10 dB SNR. So no improvement in accuracy is achieved for Set A at 10 dB SNR. But, as the SNR decreases (10 dB downward), the total sum error of Set A outperformed baseline feature achieving gain in accuracy. For Set C, i.e. the addition of the DWT feature with DCT decreases the substitution error (S) and insertion error (I) when compared with Set A (DCT feature). Thus, decrease in the insertion error (I) increases the gain accuracy of Set C when compared with Set A. This shows the robustness of DWT features over DCT features. From Table 2a the total error distribution sum, i.e. S+D+I error, for Set A, from 0 dB to −10 dB SNR, has approximately the same sum for error distribution when compared with baseline feature. This is the reason why constant recognition is achieved at 0 dB to −10 dB SNR, for Set A, in the presence of white noise. This results in a worse recognition performance in the presence of white noise.

Similarly, for Set B, i.e. addition of visual feature using PCA decreases the insertion error (I), which results in high gain in accuracy when compared with other visual feature. The maximum $\%W_{\text{acc}}$ under white noise is found to be 10.36%, 12.25 %, and 2.77% for Set B, Set C and Set D respectively at 10 dB SNR. Whereas for Set A maximum $\%W_{\text{acc}}$ under white noise is found at 5 dB SNR, which is found to be 6.21%. By comparing the performance of all visual set, i.e. Set A/B/C/D, we found that the performance of Set B outperforms the baseline feature from 20 dB to 5 dB SNR, and Set C outperforms the baseline feature from 0 dB to −10 dB SNR. The conclusion that can be drawn out for the feature selection method is as follows. Set A shows the poor performance at higher SNR (from 30 dB to 5 dB) but shows consistent recog-

Table 1. Error distribution of white noise for different feature extraction technique with total number of deletion (D) error, substitution (S) error and insertion (I) error for audio only and audio visual recognition.

SNR	10 dB			5 dB			0 dB			−5 dB			−10 dB		
Error / Feature	D	S	I	D	S	I	D	S	I	D	S	I	D	S	I
Audio only	**210**	484	25	**607**	378	1	**821**	176	0	**825**	172	0	**825**	172	0
Set A	**274**	435	40	**346**	535	29	**390**	569	27	**462**	513	15	**461**	510	14
Set B	**264**	321	7	**617**	359	2	**749**	244	0	**803**	193	0	**800**	197	0
Set C	**225**	321	23	**397**	519	23	**591**	388	6	**736**	248	3	**727**	254	1
Set D	**290**	374	21	**599**	367	4	**773**	219	1	**788**	208	2	**797**	200	2

Table 2. Sum of deletion (D) error, substitution (S) error and insertion (I) error for different feature extraction technique for audio only and audio visual recognition.

Noise	(a) White Noise					(b) Car Noise				
SNR	10 dB	5 dB	0 dB	−5 dB	−10 dB	10 dB	5 dB	0 dB	−5 dB	−10 dB
Error / Feature	D+S+I	D+S+I	D+S+I	D+S+I	D+S+I	D+S+I	D+S+I	D+S+I	D+S+I	D+S+I
Audio only	719	986	997	**997**	**997**	107	328	776	**945**	**961**
Set A	**749**	**910**	**986**	**990**	**985**	435	608	904	**955**	**969**
Set B	592	978	993	**996**	**997**	125	388	772	**941**	**965**
Set C	569	939	985	**987**	**982**	208	381	776	**944**	**964**
Set D	685	970	993	**998**	**999**	118	399	833	**950**	**969**

nition due to their higher energy coefficient at lower SNR value. Set B shows the best performance with respect to all sets due to a dimension reduction technique of PCA. Set C performs well because combination of DCT+DWT results in better compact structure, i.e. it contains the higher energy coefficients and DWT allows the better localization of the signal. Thus, addition of visual feature in noisy environments outperformed the audio-only recognition. Hence, the addition of visual feature at clean condition has not shown any improvement over baseline (MFCC) feature but shows better performance in noisy condition. This result proves the robustness of visual features for noisy speech.

4.3. Performance analysis of audio only versus audio visual recognition for car noise

Table 2b shows the error sum, i.e. S+D+I error for car noise. Detailed error distribution of car noise for different feature set is reported in Table 3. Due to additive car noise deletions error (D) increases as the SNR level decreases when compared with baseline (MFCC) feature as shown in Table 3. Substitution error (S) increases from 10 dB to 5 dB SNR and then there is a sudden decrease in substitution error (S) from 0 dB to −10 dB SNR. Addition of visual features along with MFCC did not perform well in noisy environments. Car noise is more prominent in the low-frequency part of the signal but decays rapidly as the frequency increases (Hansen, Zhang, 2009). Low frequency components of the speech features are corrupted more by car noise, and as the background noise changes, speech spectral changes. Thus, alteration in speech spectra results in dramatic fall during recognition. Also, from Table 2b, total sum error distribution, i.e. S+D+I, for car noise, increases as the SNR value decreases, which shows the similar behaviour as white noise, which results in worse recognition performance. None of the visual feature performed well for car noise. Therefore, visual robust feature, which can perform well in presence of car noise, is to be investigated.

4.4. Performance analysis of audio only versus audio visual recognition for babble noise

Comparison of audio-visual features performance of Set A, Set B, Set C and Set D with respect to babble noise is also evaluated. Figure 7 shows that addition of visual feature using Set A and Set C demonstrates no improvement in the accuracy. Small gain in accuracy is reported at −5 dB SNR using Set A and Set C. Similarly for Set B an improvement in recognition below 5 dB SNR is reported. Set D performed well under babble noise condition, and the improvement is achieved at all value of SNR, i.e. 10 dB to −10 dB. For babble noise Set D (2D-DWT followed by PCA feature) outperformed when compared with other visual feature. Table 4 shows the detailed error distribution of babble noise.

Table 3. Detailed error distribution for audio visual recognition performance of car noise for different feature extraction technique, where D = Deletion error, S = Substitution error and I = Insertion error.

SNR	10 dB			5 dB			0 dB			−5 dB			−10 dB		
Features	D	S	I	D	S	I	D	S	I	D	S	I	D	S	I
Set A	**173**	249	13	**247**	342	19	**578**	321	5	**768**	187	0	**799**	170	0
Set B	**43**	**72**	10	**165**	**206**	17	**523**	**242**	7	**764**	**177**	0	**798**	**167**	0
Set C	**80**	111	17	**148**	206	27	**449**	320	7	**734**	210	0	**792**	172	0
Set D	**36**	69	13	**202**	181	16	**613**	216	4	**781**	169	0	**798**	171	0

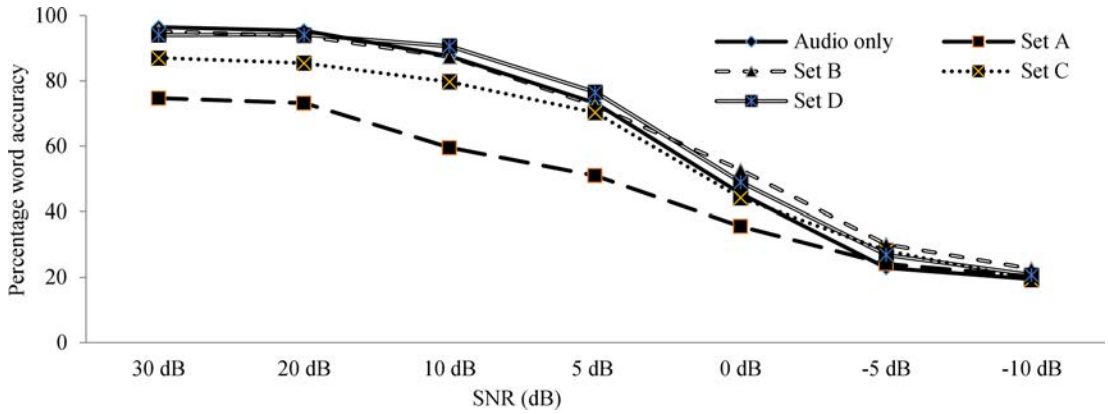

Fig. 7. Percentage word recognition for audio only and audio visual speech recognition using babble noise.

Table 4. Detailed error distribution for audio visual recognition performance of babble noise for different feature extraction technique, where D=Deletion error, S=Substitution error and I=Insertion error.

SNR	10 dB			5 dB			0 dB			−5 dB			−10 dB		
Features	D	S	I	D	S	I	D	S	I	D	S	I	D	S	I
Set A	165	**307**	24	166	**400**	35	187	**548**	54	240	**632**	57	271	**648**	57
Set B	43	**87**	25	71	**207**	60	163	**369**	47	305	**518**	33	491	**448**	9
Set C	89	**138**	22	90	**238**	36	180	**442**	62	226	**571**	83	244	**649**	95
Set D	22	**75**	18	49	**194**	45	137	**422**	66	268	**575**	53	378	**559**	34

Table 5. Percentage recognition for audio only and audio visual recognition for sets of visual features.

SNR (dB)	(a) Factory noise					(b) Machine gun noise				
	Audio	Set A	Set B	Set C	Set D	Audio	Set A	Set B	Set C	Set D
30	96.00	74.45	94.12	86.53	93.22	96.24	74.78	95.02	87.92	93.14
20	96.33	69.39	94.37	86.20	93.71	96.00	74.12	94.29	86.69	93.47
10	**89.47**	58.61	86.78	78.86	**90.69**	93.88	71.35	93.31	85.47	91.59
5	**75.02**	45.47	71.67	61.96	**75.35**	**89.71**	67.10	87.59	80.65	**89.63**
0	**36.57**	26.78	**37.55**	**39.02**	**42.12**	**79.35**	57.31	**80.73**	69.47	**84.65**
-5	**23.51**	21.06	22.61	19.67	**23.76**	**60.24**	47.67	**67.02**	57.47	**69.55**
-10	**24.16**	20.24	22.86	19.84	22.86	**40.65**	37.31	**46.29**	**42.04**	**47.92**

From the Table 4, it can be easily observed that, as the SNR value decreases, all the three errors, i.e. substitution error (S), deletion error (D) and insertion error (I), increase. Among all three errors, contribution of substitution error (S) is more, which gradually increases with the decrease in SNR. The same effect is shown in the insertion error (I). Babble noise contains the spectral peaks of the voice which are distributed over both time and frequency (JÜRGENS *et al.*, 2013). When two speech signals get overlapped, there are chances of spectral smoothing between the babble noise and speaker's utterance, especially during silence and pause period. Hence, if their correlation (matching) between features is stronger, i.e. noise can resemble speaker's utterance, then the performance of ASR increases, and if not it will result in poor performance. Therefore, for Set A and Set C, from Table 4, error distribution, i.e. (S+D+I), increases as the SNR value decreases.

On the other hand, addition of visual features of Set A and Set C has not shown any significant improvement. Similarly, for Set B, error distribution is reduced when compared with Set A and Set C, resulting in better recognition. Set B features outperformed baseline feature (MFCC) from 0 dB to −10 dB SNR. Similarly, Set D outperformed baseline feature from 10 dB to −10 dB SNR. The results prove the robustness of visual features for noisy speech recognition. Maximum $\%W_{\text{acc}}$, for Set A/C/D is reported as 1.22%, 5.24%, and 3.91% respectively at −5 dB SNR and for Set C it is 7.27% at 0 dB SNR. Finally, the experiment was performed for remaining two noises, i.e. factory noise and machine gun noise as shown in Table 5.

Table 5b shows the percentage recognition, in terms of word accuracy, for audio only and audio visual speech recognition in presence of factory noise and machine gun noise, respectively. Thus, in the presence of factory noise the speech signal is more corrupted and addition of Set A/B/C visual feature also did not perform well under noisy condition. Set D feature provides the robustness to system, thereby outperforming the baseline feature from 10 dB to −5 dB SNR. From Table 5, we observe that improvement in the recognition is achieved at lower values of SNR. Thus, addition of visual feature over MFCC in a noisy environment proved to be robust. Also, comparing all the visual feature extraction technique proposed in this paper, Set D outperformed the MFCC features in noisy condition.

5. Conclusion

This paper reports the Hindi bimodal technique for increasing the robustness of an ASR system under noisy background conditions. Performance of noisy speech signals using NOISEX database is evaluated for different Signal to Noise ratio (SNR: 30 dB to −10 dB) using AMUAV corpus for 10 subjects under different acoustic environments (white Gaussian noise, car noise, babble noise, factory noise and machine gun noise). As experimental results reported in Sec. 4 clearly show, the performance of an ASR system degraded when injected noise power is more than the source signal and additional of visual features along with the audio features help to increase the robustness. Word recognition using four set of visual feature, i.e. Set A: Two-Dimensional Discrete Cosine Transform

feature (2D-DCT) feature, Set B: Principal Component Analysis (PCA) feature, Set C: Two-Dimensional Discrete Wavelet Transform followed by DCT (2D-DWT-DCT) features and Set D: Two-Dimensional Discrete Wavelet Transform followed by PCA (2D-DWT-PCA) have been reported for extracting the visual feature and adding only 9 visual dimension feature have reported an increase in the word recognition rate. Also, it has been concluded that the Set B and Set D have shown better performance when compared with other visual features. Hence, one can conclude that the performance of audio visual is highly dependent on the type of visual features as well as the type of acoustic noise under which the performances is to be measured.

References

1. Abdelaziz A.H., Zeiler S., Kolossa D. (2015), *Learning Dynamic Stream Weights for Coupled-HMM-Based Audio-Visual Speech Recognition*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, **23**, 5, 863–876.
2. Ahmad N., Mulvaney D., Datta S., Farooq O. (2008), *A Comparison of Visual Features for Audio-Visual Automatic Speech Recognition*, Acoustic, Paris, pp. 6445–6449.
3. Bruce L.M., Koger C.H., Li J. (2002), *Dimensionality Reduction of Hyperspectral Data Using Discrete Wavelet Transform Feature Extraction*, IEEE Transactions on Geoscience and Remote Sensing, **40**, 10, 2331–2338.
4. Carboneras A., Gurban M., Thiran J. (2007), *Low Dimensional Motion Features for Audio-Visual Speech Recognition*, Proceeding of the 15th European Signal Processing Conference (EUSIPCO), Poznan, Poland, pp. 297–301.
5. Cardinaux F., Sanderson C., Sebastien M. (2003), *Comparison of MLP and GMM Classifiers for Face Verification on Xm2vts*, IDIAP Research Report (IDIAP 3–10), pp. 1–9.
6. Chen T. (2001), *Audio visual speech processing*, IEEE Signal Processing Magazine, pp. 9–31.
7. Chourasia V., Samudravijaya K., Ingle M., Chandwani M. (2007), *Hindi Speech Recognition under Noisy Conditions*, International Journal of Acoustic Society India, pp. 41–46.
8. Farooq O., Datta S., Vyas A. (2005), *Robust isolated Hindi digit recognition using wavelet based denoising for speech enhancement*, Journal of Acoustical Society of India, **33**, 1–4, 386–389.
9. Farooq O., Datta S., Shrotriya M. (2010), *Wavelet sub-band based temporal features for robust Hindi phoneme recognition*, International Journal on Wavelets and Multiresolution Information Processing, **8**, 6, 847–859.
10. Gundimada S., Asari V. (2004), *Face detection technique based on rotation invariant wavelet features*, Information Technology: Coding and Computing, **2**, 157–158.
11. Hansen J., Zhang X. (2009), *Analysis of CFA-BF: Novel combined fixed/adaptive beamforming for robust speech recognition in real car environments*, Speech Communication, **52**, 134–149.
12. Hoenig F., Stemmer G., Hacker C., Brugnara F. (2005), *Revising Perceptual Linear Prediction (PLP)*, Proceeding of 9th European Conference on Speech Communication and Technology, Interspeech 2005, pp. 2997–3000.
13. Huang J., Potamianos G., Connell J., Neti C. (2004), *Audio-visual speech recognition using an infrared headset*, Speech Communication, **44**, 4, 83–96.
14. Jürgens T., Brand T., Clark N.R., Meddis R., Brown G.J. (2013), *The robustness of speech representations obtained from simulated auditory nerve fibers under different noise conditions*, JASA Express Letters, Journal of the Acoustical Society of America, **134**, 3, 282–288.
15. Khanam R., Mumtaz S.M., Farooq O., Datta S., Vyas A.L. (2010), *Audio Visual Features For Stop Recognition From Continuous Hindi Speech*, National Symposium on Acoustics.
16. Lee J.S., Park C.H. (2008), *Robust audio visual speech recognition based on late integration*, IEEE Transactions on Multimedia, August, **10**, 5, 767–779.
17. Lokesh S., Balakrishnan G. (2012), *Robust Speech Feature Prediction Using Mel-LPC to Improve Recognition Accuracy*, Information Technology Journal, **11**, 1, 1644–1649.
18. Mishra A., Chandra M., Biswas M., Sharan S. (2011), *Robust Features for Connected Hindi Digits Recognition*, International Journal of Signal Processing, Image Processing and Pattern Recognition, **4**, 2, 79–90.
19. Naomi H., Eoin G. (2015), *TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech*, IEEE Transactions on Multimedia, **17**, 5, 603–615.
20. Navnath S., Raghunath S. (2012), *DWT and LPC based feature extraction methods for isolated word recognition*, EURASIP Journal on Audio, Speech, and Music Processing 2012, pp. 1–7.
21. Neti C., Rajput N., Verma A. (2002), *A Large Vocabulary Continuous Speech Recognition System For Hindi*, Proceeding of Works. Multimedia Signal Process, pp. 475–481.
22. Neti C., Potamianos G., Luettin J., Matthews I., Glotin H., Vergyri D., Sison J., Mashari A., Zhou J. (2000), [in:] Technical report on Audio Visual Speech Recognition, Center for Language and Speech Processing, The John Hopkins University, Baltimore.
23. Patterson E., Gurbuz S., Tufekci Z., Gowdy J. (2002), *CUAVE: A new audio-visual database for mul-*

timodal human-computer interface research, proceeding of the IEEE International Conference of Acoustics, Speech, and Signal Processing, **2**, 2017–2020.

24. Potamianos G., Neti C. (2001), *Automatic Speech reading for Impaired Speech*, Proceedings of the Audio Visual Speech Processing Workshop.
25. Potamianos G., Neti C. (2003), *Audio visual speech recognition in challenging environments*, Proceedings of the European Conference on Speech Communication and Technology, Geneva, Switzerland, pp. 1293–1296.
26. Potamianos G., Neti C., Gravier G., Garg A., Andrew W. (2003), *Recent Advances in the Automatic Recognition of Audio visual Speech. Invite paper*, Proceedings of the IEEE, **91**, 9, 1306–1326.
27. Potamianos G., Neti C., Luettin J., Matthews I. (2004), *Chapter to appear* [in:] Issues in Visual and Audio-Visual Speech Processing, MIT Press, 2004.
28. Pradhan G., Haris C., Prasanna S., Sinha R. (2012), *Speaker verification in sensor and acoustic environment mismatch conditions*, International Journal of Speech Technology (Springer), **15**, 3, 381–392.
29. Samudravijaya K. (2004), *Variable Frame Size Analysis for Speech Recognition*, [in:] Proceedings of the International conference on Natural Language Processing, Dec 19–22, Hyderabad, pp. 237–244.
30. Sanderson C., Paliwal K. (2004), *On the use of speech and face information for identity verification*, IDIAP research report (IDIAP–RR 04–10), pp. 1–33.
31. Seymour R., Stewart D., Ming J. (2008), *Comparison of image transform-based features for visual speech recognition in clean and corrupted videos*, Journal on Image and Video Processing, EURASIP, Hindawi Publishing Corporation, **2008**, 1–9.
32. Smith L.I. (2002), *A tutorial on Principal Components Analysis*, pp. 2–8.
33. Upadhyaya P., Farooq O., Varshney P. (2012), *Comparative study of viseme recognition by using DCT feature*, Proceeding of the International Symposium Frontier Research on Speech and Music(FRSM), Gurgaon, Haryana, India, pp. 171–175.
34. Upadhyaya P., Farooq O., Varshney P., Upadhyaya A. (2013), *Enhancement of VSR Using Low Dimension Visual Feature*, IEEE Proceeding of the International Conference on Multimedia Signal Processing and Communication Technologies (IMPACT), AMU, Aligarh, India, pp. 71–74.
35. Upadhyaya P., Farooq O., Abidi M.R., Varshney P. (2014), *Performance Evaluation of Bimodal Hindi Speech Recognition under Adverse Environment*, Advances in Intelligent Systems and Computing, Springer International Publishing, **328**, 347–355.
36. Varga A., Steeneken H.J.M. (1993), *Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems*, Speech Communication, **12**, 3, 247–251.
37. Varshney P., Farooq O., Upadhyaya P. (2014), *Hindi viseme recognition using subspace DCT features*, International Journal of Applied Pattern Recognition (IJAPR), **1**, 3, 257–272.
38. Young S. (2008), *HMMS and related speech recognition technologies*, [in:] Handbook on Speech Processing and Speech Communication, Springer-Verlag Berlin Heidelberg, pp. 539–558.
39. Zhou Z., Hong X., Zhao G., Pietikainen M. (2014), *A compact representation of visual speech data using latent variables*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **36**, 1, 181–187.
40. http://www.internationalphoneticalphabet.org.