How to Cite:
Srikanth, G. N., & Venkatesha, M. K. (2022). Word recognition through mapping of lip movements from speech utterance using audiovisual fusion and MLP. *International Journal of Health Sciences*, *6*(S2), 4533–4545. https://doi.org/10.53730/ijhs.v6nS2.6078

Word recognition through mapping of lip movements from speech utterance using audiovisual fusion and MLP

Srikanth G N
Department of Electronics and Instrumentation Engineering, R N S Institute of Technology, Affiliated to Visveswaraya Technological University, Bengaluru, Karnataka, India
Email: srikanth.g.n@rnsit.ac.in

M K Venkatesha
R N S Institute of Technology, Affiliated to Visveswaraya Technological University, Bengaluru, Karnataka, India
Email: principal@rnsit.ac.in

***Abstract*---**Speech has information more than text, but under noisy environment speech sufferance from disadvantage of not properly decoded by humans and same is true with machines. speech being bimodal along with audio features if we augment visual features specifically related to lip movements. the degree of speech recognition can be improved. The objective of this work is to use audio and visual features to aid word recognition. In this work we extracted MFCC features for audio and Geometrical features of lip movements together is used in machine learning algorithm to predict the word utterances. Videos related to word utterances are extracted from TIMID database. With the statistical information related to audio and corresponding visual features from lip movements is extracted to form input feature vector to machine learning algorithm (Multi-layer perceptron). The experimental results show that using MLP we have obtained a word recognition accuracy of 91% and using KNN Classifier the accuracy attained is 61%. The results presented here have important implications for applications in HMI communication and helps hearing impaired.

***Keywords*---**human machine interface, automatic speech, recognition systems, multilayer perceptron, rectified linear unit.

Introduction

Speech is considered to be one of the major means of communication between humans as well in human computer interaction (HCI). Speech is bimodal in nature encompassing both audio and visual modality. Every system or an application is always affected by some kind of noise either initial stage of signal extraction or in subsequent stages of processing, many a times with the absence of audio utterance being corrupted, still human can interpret the information by perceiving the lip movements, of course the knowledge of the prior information gathered will also aid the utterance recognition, i.e, even though the auditory modality is dominant in speech perception it has been shown that the visual modality increases speech intelligibility [17][18][19]. Over decades continuous research effort is gone in to ASR in order to cater several needs of HMI and IoT related applications, at the same time increase for demand in recognizing robust significant features from speech utterance. If Audio speech is supplemented with visual lip movements that is by combining the audio-visual features together, when it is used to train the model then recognition rate can be improved effectively even when speech is captured under adverse conditions. It is known that there is no one to one mapping exists between speech and the corresponding lip movements always, for some vocabulary example homophones mapping is even many to one in such cases word recognition cannot be performed solely by using visual information. The bimodal processed speech signal can be considered as a vector that combines acoustic and visual features.

In the visual speech recognition system, extracting facial feature is a difficult task due to the large variations in appearance among people, differences in speech production, accent, skin colour, illumination conditions and head positions variation further causes difficulties in corresponding image frame analysis to extract the features, facial feature extraction is equally important for emotion recognition that predominantly uses visual input. The visual speech analysis majorly concentrates on extraction of ROI related to variation of the lip movement in succeeding frames, though facial features also contribute to certain extent major contribution in extracting the features relay on the lip movements which can be viewed as a deformable object.

Audio feature extraction

It is known that major part of the speech information signals will be in the frequency range 200 Hz to 1.5 kHz, however the audio speech signal is considered to be roughly up to 4 kHz. Though the sampling frequency of 8 kHz is sufficient to sample and reconstruct the signal, today's modern speech processing equipment codecs use a sampling frequency of 44.1 kHz or 48 kHz, since the audio range is varying in the range between 20 hertz to 20kHz. Speech signal is preprocessed to remove noise to improve the SNR and preemphasis is carried out to boost the magnitude of high frequencies. Speech processing is essential to improve the quality of perception, since it is applied in ASR, Speaker identification system and HMI etc. The methods that can be applied to extract the information content from audio utterance in the form of features can be with the use of autocorrelation, Pitch detection, extracting LPC coefficients [20], Wavelet coefficients exhibiting time and frequency resolution characteristics[16], MFCC [5],[12],[13],[15] etc, and

for prediction and classification of utterance traditionally researchers used HMM modeling with Expectation Maximization algorithm [6],[8] which is related to the design of probabilistic models. Use of machine learning algorithms emerged as an alternate solution producing promising results. The algorithms like Multilayer perceptron [15] that suits for labelled input, output format usually for the limited dataset, however with the availability of large data records one can use deep neural network. We have extracted features from audio utterance using MFCC, MFCCs have higher correlation with lip movements and MFCC improves mapping between audio and lip movement [2].

Mel Frequency Cepstral Coefficients

The purpose of the system is to convert the speech wave into a parametric form to analyse and derive the information content in the form of features vectors. Speech examined over a short interval of time i.e., 20 – 30 msec, the properties of the signal is considered to be statistically stationary, however for a longer duration structural change show different sounds of speech utterance. The short time Fourier Transform is one of the most popular ways to analyse speech signals, speech segment (Frame) of 20 to 30 msec time duration with a frame interval of 10 msec contain enough samples to obtain reliable spectral data. To obtain MFCC coefficients, FFT is applied on each frame which is excerpt of the signal considering overlapping. On the obtained spectra apply Mel scale filter bank to extract energy of each bin. The Mel scale maps the perceived speech frequencies and its variations related to tone, pitch. By discriminating small pitch changes in lower frequencies than higher ones like humans hearing system. Frequency mapping from linear scale to Mel scale in expressed in (1).

$$\mathrm{M(f)} = 1125 \ln\left[1 + \frac{\mathrm{f}}{700}\right] \qquad (1)$$

'f' in hertz and 'm' in mels, the inverse equation is given by

$$\mathrm{M^{-1}} = 700[\exp^{\left(\frac{\mathrm{m}}{1125}\right)} - 1] \qquad (2)$$

Applying log on the powers at each of the Mel frequencies, in reality, this operation is mapped to human hearing perception and to obtain the cepstral features we apply DCT, that decorrelate the energies of the list of log powers in different Mel bands, resulting in generating the cepstral coefficients. Fig 1 depicts the entire process of MFCC extraction from speech utterance [1] – [4].

Using the procedure described above, for each 30msec speech segment with an overlap of 10 msec, the MFCC is computed. This set formed by extracting the first 13 elements forms the acoustic vector that gives information about the phonemes, formants, vocal tract response and ignore the higher end elements since it is related to fast changing spectral information. The very first element gives information about energy content and, we ignore and consider next 13 MFCC coefficients, if the dynamics of speech also to be investigated like in music analysis, or decoding sentence then along with generated MFCC the velocity and acceleration coefficients can be augmented to form feature vector. In applications

such as speech analysis related to children or baby crying, we can also add energy, pitch information, ZCR, spectral centroid to form acoustic feature vector

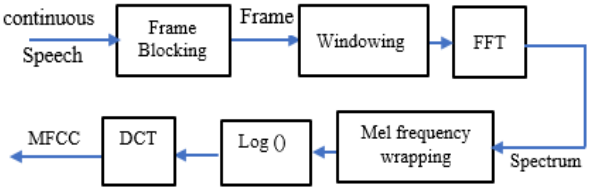

Fig. 1. MFCC for Speech Utterance

The obtained MFCC vectors combined to form covariance matrix and the diagonal elements of the covariance matrix called the variance can be used to form acoustic feature vector.

Visual feature extraction

To supplement the visual features with audio features for identification of the word utterance, as discussed, we need to incorporate the visual features specifically related to lip movements. Essentially video processing involves frame analysis in which object to be extracted is considered to be the most deformable part of the face i.e, mouth portion. Video of 30 frames per second is considered which is compromise between the object recognition and the computation complexity involved in processing and time constraint. Ultimately the work involves detection of ROI in the frames put together forming a sequence, hence a video segment of a very short duration as applicable in vowel, letter or the word utterance. Many approaches are available like using applying discrete cosine transform, Wavelets and for object recognition, use of template matching that can be used in identifying or separating the object from the scene and to study the same to extract the significant features related to the object under consideration, but use of a specific method must simplify the complexity and the reduce the time taken for processing. We have the scenario of identifying the deformable object in the scene/frame. The methods for extracting visual speech information from image tracking are image-based, motion-based, visual-geometric-based and, model-based. In an image-based approach a gray image containing an orofacial component (ROI) is used directly or after image modification as a feature vector while the visual-based approach assumes that visual movement during speech production contains appropriate speech information. Geometric-based techniques, on the other hand, consider certain measures such as length or width of the mouth opening to be important factors, and in a model-based approach, a visual representation model, usually lip contours, is designed and designed to define a small set of boundaries.

The objective of feature analysis algorithms is to search structural features of an object that exists in the scene even with varying background, lighting conditions and colour. The structural features are used to locate objects example lips in the face. These methods are proposed generally for object localization. Various feature analysis algorithms are Viola Jones method, Gabor feature extraction method etc. In audio or visual features extraction the transforms that are applied the data

either by using cosine transform (DCT) or wavelet transform, the requirement is to reduce the dimension of the features by redundancy removal and to retain most significant features by applying PCA, finally applying probability-based model for speech recognition using HMM. Automatic speech recognition is significantly improved in performance with use of neural network algorithms like MLP, DNN which takes the raw data for analysis to identify the speech.

Problem Formulation

Speech signal produced can be perceived as a multimodal information generally considered for emotion recognition where facial expression also contribute a lot, but for automatic speech recognition involving speech is seen as bimodal information using audio and the visual information pertain to lip movements. visual speech recognition is also given much more importance in automatic speech recognition (ASR), though auditory signal contribution is more but when audio is corrupted by noise usually in adverse environment visual speech processing improve the recognition rate, further when words uttered by an accented speaker, can be more easily recognized with the addition of lip movements (visual signal). There exists a strong correlation between the audio signal and the corresponding visual signal that can be detected from human mouth part of the face, specifically lip movements. To understand and to enhance the audio-visual speech recognition rate numerous methods were attempted and implemented by researchers in the field of ASR. Extracting MFCC [3],[5],[12,[13] from speech utterance. MFCC performs better compared to LPC [20] when used for speech analysis, features gathered by applying transforms on both audio and visual signal used in ASR systems. various methods and algorithms were implemented by researches in both the domains w.r.t preprocessing as well extracting the relevant significant feature and using hidden Markov model and machine learning algorithms like MLP, DNN.

DCT coefficients obtained from speech and even for extracting important features related to lip and its movements while analysing video related to word utterance. DCT being an orthogonal transform decorrelated the energies which means diagonal covariance metrics can be used to model the features. another approach applied for visual speech recognition by using lip movements includes considering edge detection algorithm for ROI extraction and GLCM (Gray Level Co-occurrence Matrix), Gabor convolve algorithm for best feature extraction of lip parameters [10]. Applying wavelet transform [16] on image frames to extract and track the features generated from deformable object like lip movements and preparing wavelet coefficients to work as input vector for machine learning.

Template matching algorithm using cross correlation method for locating and recognizing a face has been applied on various face candidates to match the template with right face candidate [9].Essentially the methods adopted must use robust significant features using both modalities of speech and to reduce the dimension of the feature vector space, which is supposed to be the input to a machine learning algorithm there by it reduces the computation complexity and time consumed for processing, methods like singular value decomposition(SVD), Principal component analysis(PCA), or extracting the statistical variations of the chosen variables like variance, entropy will contribute for improving the ASR

systems. With the availability of large database about the speech, video input for words or short sentence, we can use the DNN which can handle raw data and dimensionality reduction made possible by max pooling to arrive at limited dimension robust significant feature set, based on which predicted output about the utterance can be ascertained.

It is known facial and acoustic features are strongly interrelated [2], but many utterances will have many to one mapping between speech utterance and the corresponding lip movement, hence it is desirable to have combined features obtained from acoustic and visual features extracted from lip movements to form an input for training machine for proper recognition even in case any one of the input is corrupted due to noise usually happens in audio though it has higher level of modality. It is revealed that from [17],[18],[19] by combining audio and visual information has been shown to significantly improve the overall performance ASR, especially in adverse acoustic conditions. To extract the lip features even Active shape model (ASM) is a shape-constrained iterative fitting algorithm. The shape constraint comes from the use of a statistical shape model, also known as a point distribution model (PDM), that is obtained from the statistics of hand-labelled training data. In such application, the PDM describes a reduced space of valid lip shapes, in the sense of the training data, and points in this space are compact representations of lip shape that can be directly used as features for lipreading.

Active Appearance Models. An active appearance model (AAM) is a statistical model of both shape and gray-level appearance. In the lipreading context, it combines the gray-level analysis approaches of with the shape analysis. Further the features obtained is used to train the HMM or can be used machine learning for classification [7] For audio visual feature extraction to detect the utterance of vowels/words use of MFCC, for face and mouth detection, Viola-Jones object recognizer called haarcascade for mouth detection ROI and transforms like DWT for visual feature extraction and HMM classifier [14] from features formed by combined audio visual features. Development of neural networks in recent years, automatic learning of visual features by supervised training of convolutional neural networks (CNN) or LSTM has significantly improved performance.

Objectives of the Work

- Preprocessing the bimodal speech.
- Significant features extraction considering statistical information and its variations in both the domain.
- Recognition and classification of word utterance using machine learning algorithm aiming at higher degree of accuracy.

Problem Solution

Video files related to word utterances are prepared from TIMID dataset consisting of 460 audio and its related video files, the files related to word utterances are saved in mp4 format.Audio and corresponding video frames are separated. Audio processing is carried out as discussed to extract mel frequency cepstral coefficients, average of three observation in synchronism with video frames

determined likewise observations of generated 13 MFCC are recorded in csv file. From video, image frames are extracted, and pre processing is carriedout to filter the noise.Template matching performed to extract the orofacial area (ROI) that is the mouth portion. Region of interest for computation of visual feature will be concentrated towards the movement of lips over the time frame which is very complex as discussed. In-view of this calculating good and discriminatory visual feature of mouth plays vital role in the recognition. The basic step is to search ROI to localize and track the lips in the image and then to extract the features.

Correlation between template and scene is given by

$$R_{tf}[i,j] = \sum_{m}\sum_{n} f[m,n]\, t[m-i\,, n-j] = t \otimes f \tag{3}$$

Where $R_{tf}[i,j]$ is the cross correlation f[m,n] is the scene , t[m-i , n -j] indicates template □(i , j). t ⊗ f indicates correlation between scene and the template. Normalized Cross-Correlation given by

$$N_{tf}\,[j\,j] = \frac{\sum_m \sum_n f[m,n]\; t[m-i\,, n-j]}{\sqrt{\sum_m \sum_n f^2[m,n]}\;\sqrt{\sum_m \sum_n t^2\,[m-i\,, n-j]}} \tag{4}$$

Essentially for tracking the mouth portion experimentally it is found that we have use minimum of two templates one with lip open and other with lip closed [2], depending on the degree of normalized correlation (optimum value chosen is 0.65) exist between the template and the deformmable object (mouth portion) proper template is used to scan the entire image frame and mouth portion is extracted. Equation (3) and (4) shows the correlation and normalized cross correlation between the scene and the template. For the gray image obtained preprocessing is carried out by calculating the histogram of the gray image and result is used in global thresholding to extract lip portion in all the frames related to particular utterance.

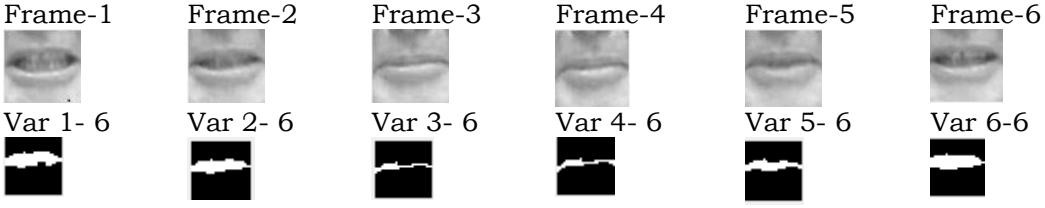

Fig. 2 Frame sequence with extracted mouth and lip (ROI) movement variation for word utterance “me”.

The extracted mouth portion is resized to 28 x 28 from every frame related to particular word utterance. The obtained images are subjected to blob analysis to identify the objects and removing the redundant information and retain relevant largest object that is solely lip portion is considered from the binary image. Fig.2 show the extracted mouth and lip portions from sequence of frames for the utterance. From the sequence of frames object details related to lip movements

that is geometrical features viz “Area”, MajorAxis”, “MinorAxis”, Eccentricity”, ratio of “Major Axis to Minor Axis” ascertained and tabulated in csv file hence reducing the dimentionality. The two matricies having audio observation vector and corresponding visual feature vectors are augmented to generate a matrix, the approach is to form a combined feature space by concatenating the two feature vectors.Covariance among the variables is determined, the diagonal elements of the covariance matrix that form the variance of the chosen variables is extracted and a complete record for the word utterance for every video subjected to feature extraction, hence a row of observation related to A-V fusion is populated in csv file with labelled input variables and known output. Thus forming 460 records consisting of significant features with reduced dimensionality is derived from the considered vidTIMIT dataset.

We have applied two classifier for word utterance prediction and classification that is Multilayer perceptron and K-Nearest Neighbour. MLP, feed forward network MLP with 3 hidden layers,back propagation for updating the weights and using tanh as an activation function, iteratively trained to get optimum parameters to minimise the error. Fig.3 depicts the complete flow used in audio visual speech processing for prediction and classification of the word utterance.
With the same extracted features, we have applied K- Nearest Neighbour which is a class of unsupervised machine learning where in the data to be predicted k is assigned to centroid cluster based on the similarity measure i.e., distance function. KNN yield highly adaptive behaviour and optimal in large sample limit. The demerit is that it is computationally intensive and require large storage requirement. KNN is a traditional classification method and does not require any training effort, it relies heavily on quality of samples, noise can easily affect the performance of the classifier. The distance function used is the most common matric Euclidean distance, depicted in (5) which is useful in low dimensional dataset.

$$D_{\text{Euclidean}}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \qquad (5)$$

where x and y are m-dimensional vectors denoted by x = (x_1,x_2,x_3 … x_n) and y = (y_1,y_2,y_3 …y_n) represents ‘n’ attributes of two records.

Table 1 and Fig.5, Fig.6 depicts the accuracies obtained considering individual and combined modalities of audio and visual information w.r.t word utterance.

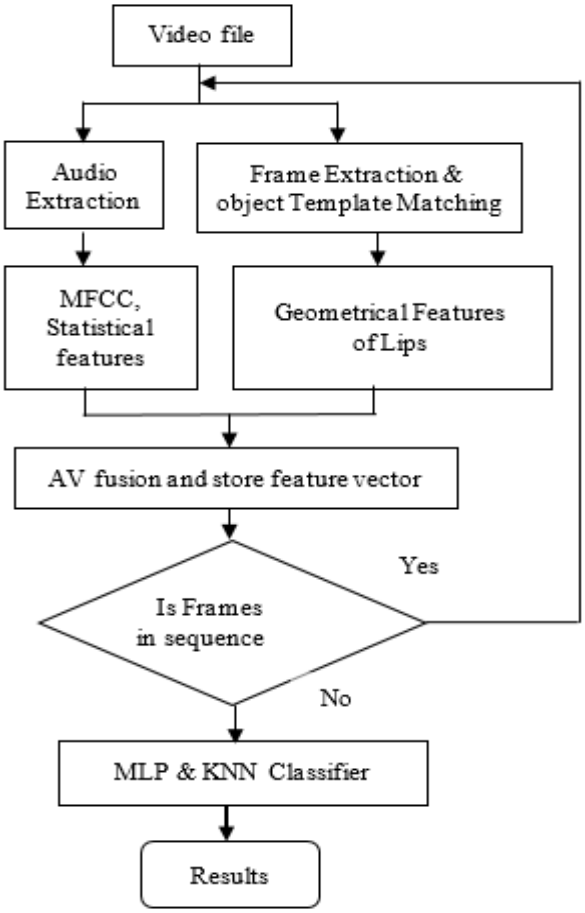

Fig. 3 Key stages of steps involved in Audio Visual speech recognition.

For the implementation of the work we have developed Program using MATLAB software to segregate audio and related image frames from videos, the audio sampling rate is 48kHz, and frame rate is 30 frames/sec.For each frame the audio feature vector contains a total of *13* real scalars (usually for better discrimination delta (velocity) coefficients can also be included but since we achieved good results using only Cepstral coefficients and we wanted to keep number of features as low as possible we did not include velocity and acceleration coefficients, and for visual feature extraction, geometrical features are made to form a visual feature vector as discussed above. MLP and KNN algorithms are implemented based on the feature set for prediction of the word utterance program is written in Python. The experimental results are shown in Table 1 and Fig. 4-7.

Dataset

TIMID database is used in our work, steps followed to clip the audio word utterance using Audacity software and using windows live movie maker image frames related to the word utterance is superimposed to sync to form video database consisting of 430 files stored in .MP4 format. Database involves a set of

word utterances like *"He", She", "Had", "The" etc*, from the different age groups from male and female (excluding children).

Results

Table 1, Performance of the classifiers for word utterance recognition on VidTIMIT dataset. The first column (M) specifies the input modalities: A, V, and AV denote, audio-only, video-only, and audio-visual models respectively, MSE.

M	Features set	Classifier	Recog Accuracy	MSE
A	MFCC	MLP	91%	0.08
A	MFCC	KNN	83%	0.61
V	Geometrical parameters	MLP	100%	0.00
V	Geometrical parameters	KNN	72%	0.77
AV	MFCC, Geometrical parameters	MLP	96%	0.04
AV	MFCC, Geometrical parameters	KNN	61%	1.33

Table 1 depicts the results in considering only audio data, video data and Audio visual combined (AV) data. Further its observed that combining audio and visual features the performance of the classifier has increased w.r.t accuracy in evident from Fig 4 and 5. Though use of KNN classifier resulted in an accuracy of 83% for audio and 72% for video, put comparatively when audio MFCC and visual lip movement geometrical information is combined only 61% accuracy recorded, this shows that MLP outperform KNN classifier. Because KNN uses each sample as independently and try to assign nearest to the centre of cluster based on similarity measure. It stores the training data using memory, noise can easily affect the performance of the KNN classifier

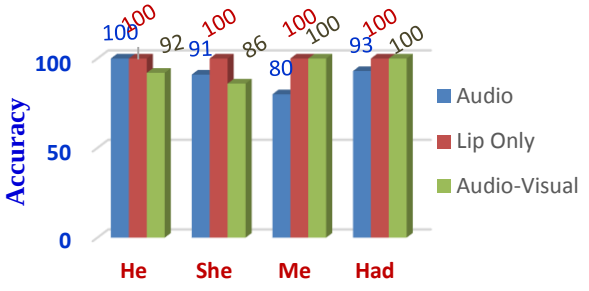

Fig. 4, Word recognition accuracy w.r.t Audio, Visual and Audio-Visual fusion using MLP classifier

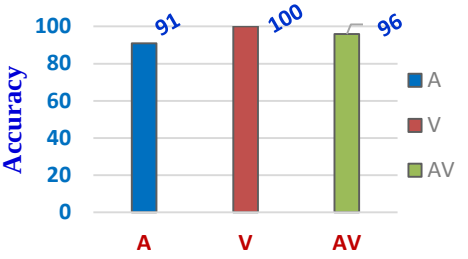

Fig. 5, Overall Accuracy of Audio, Visual, Audio-Visual using MLP

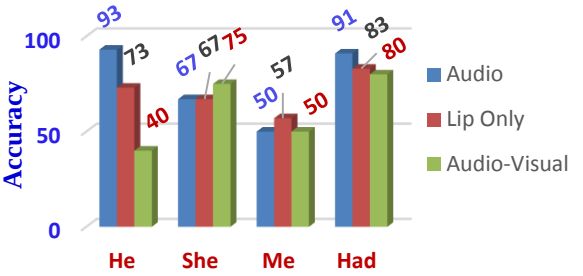

Fig. 6, Word recognition accuracy of Audio, Visual and Audio-Visual fusion using KNN classifier

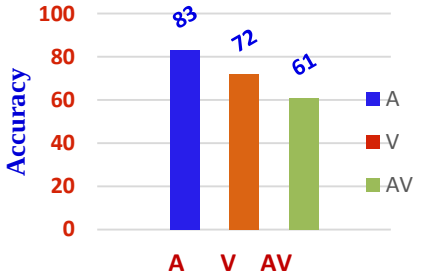

Fig. 7, Overall Accuracy of Audio, Visual, Audio-Visual using KNN.

Conclusion

We have shown that Acoustic and Visual information related to lip movements will improve the recognition accuracy of word utterances. Statistical features (variance) are extracted in both the domains. The audio information is extracted using MFCC and visual by using geometrical properties. Using ANN feedforward network Multilayer perceptron with three hidden layers and tanh as an activation function seen that the network is tuned to give considerable accuracy of 96% in classifying the word utterances. even with limited dataset (records). By experimentation considering individual modalities, combining audio visual modalities it is found that, in all the cases MLP outperform KNN classifier. The above work is carried out considering limited dataset, however use of larger dataset is a key requirement to use DNN and other methods which can contribute for the continuous speech recognition. We are grateful to RNS Institute of

Technology and Management for the support extended and providing R&D facilities.

References

1. Stéphane Dupont and Juergen Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition" *IEEE transactions on multimedia*, vol. 2, no. 3, sept 2000
2. Carlos Busso, Student Member, IEEE, and Shrikanth S. Narayanan, Senior Member, IEEE "Interrelation between Speech and Facial Gestures in Emotional Utterances: A single subject study" *IEEE transaction on audio,speech and language processing*,vol. X, no XX , July 2007
3. Lazaretto, F. "Converting Speech into Lip Movements "A Multimedia Telephone for Hard of Hearing People, *IEEE Trans. on Rehabilitation Engineering*, Vol.3, No.I, pp. 90-92 (1995).
4. Mahesh Goyani et. al "Performance Enhancement in Lip Synchronization Using MFCC Parameters", *International Journal of Engineering Science and Technology*, Vol. 2(6), 2010, 2364-2369.
5. Yamamaoto E, Nakamura S and Shikano K, " Speech to lip movements Synthesis maximizing audio-visual joint probability based on EM algorithm", .*IEEE Second workshop on Multimedia Signal Processing* (Cat.No.98EX175). pp. 53-58 (1998).
6. Iain Matthews, Timothy F. Cootes, J. Andrew Bangham, Stephen Cox, Richard Harvey, "Extraction of Visual Features for Lipreading". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 2, February 2002.
7. Yamamaoto E, Nakamura S and Shikano K, " Speech to lip movements Synthesis maximizing audio-visual joint probability based on EM algorithm", .*IEEE Second workshop on Multimedia Signal Processing* (Cat.No.98EX175). pp. 53-58 (1998).
8. Singh, Namrata, Daniel, A. K, Chaturvedi, Pooja, "Template matching for detection & recognition of frontal view of human face through MATLAB". *International Conference on Information Communication and Embedded Systems* (ICICES) -, (), 1–7. doi:10.1109/ICICES.2017.8070792
9. Amaresh P Kandagal, V. Udayashankara "Visual Speech Recognition Based on Lip Movement for Indian Languages" *International Journal of Computational Intelligence Research*.ISSN 0973-1873 Vol.13, (2017), pp. 2029-2041
10. S. L. Wang, A. W. C. Liew, W. H. Lau, and S. H. Leung, "An Automatic Lipreading System for Spoken Digits with Limited Training Data*". IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 18, No. 12, December 2008.
11. Jihyuck Jo, Hoyoung Yoo, and In-Cheol Park. "Energy-Efficient Floating-Point MFCC Extraction Architecture for Speech Recognition Systems" *IEEE transactions on very large scale integration (VLSI) systems*, 1063-8210 © 2015 IEEE.
12. Sayf a. majeed,Hhafizah husain, Salina abdul samad, Tariq f. idbeaa " Mel frequency cepstral coefficients (mfcc) feature extraction enhancement in the application of speech recognition a comparison study" *Journal of theoretical*

and Applied Information Technology. 10th September 2015. Vol.79. No.1. ISSN:1992-8645.

13. Befkadu Belete, "Audio-Visual Speech Recognition using LIP Movement for Amharic Language". *International Journal of Engineering Research & Technology* (IJERT) ISSN: 2278-0181 Vol. 8 Issue 08, August-2019
14. M. S. Daud, I. M. Yassin, A. Zabidi, M. A. Johari, M. K. M. Salleh "Investigation of MFCC Feature Representation for Classification of Spoken Letters using Multi-Layer Perceptron (MLP)" 2011 *International Conference on Computer Applications and Industrial Electronics* (ICCAIE 2011) 978-1-4577-2059-8/11 ©2011 IEEE.
15. Nitin Trivedi et al. "Speech Recognition by Wavelet Analysis" *International Journal of Computer Applications* (0975 – 8887) Vol 15– No.8, February 2011
16. C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition. *In technical report, John Hopkins University,* Baltimore, 2000.
17. G. Meyer, J. Mulligan, and S. Wuerger. Continuous audio-visual digit recognition using N -best decision fusion. Information Fusion, 5(2):91–101, 2004.
18. G. Potamianos, C. Neti, and S. Deligne. Joint audio-visual speech processing for recognition and enhancement. *In Proc. AVSP*, pages 95–104, 2003
19. K.Pavan Raju1 , A. Sri Krishna2 And M. Murali3. "Automatic Speech Recognition System Using Mfcc-Based Lpc Approach with Back Propagated Artificial Neural Networks". *ICTACT Journal On Soft Computing*, July 2020, Volume: 10, Issue: 04 , ISSN: 2229-6956.
20. Gholamreza Farahani, "Robust Feature Extraction using Autocorrelation Domain for Noisy Speech Recognition". *Signal & Image Processing An International Journal (SIPIJ)* Vol.8, No.1, February 2017.
21. N.M.Nawi et al., "The Effect of Pre-Processing Techniques and Optimal Parameters selection on Back Propagation Neural Networks," Vol.7 (2017) No. 3 ISSN: 2088-5334
22. Nitin Trivedi et al. "Speech Recognition by Wavelet Analysis" *International Journal of Computer Applications* (0975 – 8887) Vol 15– No.8, February 2011

4546