Proceedings of the Jangjeon Mathematical Society
24 (2021), No. 4, pp. 429 – 438

www.jangjeon.or.kr
http://dx.doi.org/10.17777/pjms2021.24.4.429

CRITICAL ANALYSIS OF DIFFERENT FEATURES FOR HINDI ASR USING ANN

ASHOK KUMAR AND VIKAS MITTAL

Abstract. Features play a vital role in Automatic Speech Recognition (ASR). They also affect the performance of speech recognizers in all environments to an extent. There are many types of features used in automatic speech recognition. Each feature type has its own significance depending upon its unique characteristics. LPC (Linear Predictive Coding), LPCC (Linear Predictive Cepstral Coefficients), Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP) and Relative spectral Analysis (RASTA) are some of the features that are generally used. Better results can be obtained by combining coefficients of different features. For instance, MFCC and PLP show better results when they are jointly used to recognize a speech signal. This paper aims to compare the performance of different features using ANN for same number of speech datasets. The result shows that proposed technique increases the recognition rate by 8.33% as compared to MFCC alone.

2010 Mathematics Subject Classification. 11T23, 20G40, 94B05.

Keywords and phrases. Automatic speech recognition, Feature extraction, Linear Predictive Coding, Mel Frequency Cepstral Coefficients, Perceptual Linear Prediction.

1. Introduction

Automatic speech recognition (ASR) is a practice to convert a sequence of words spoken by human being into text by means of machines. Speech is an acoustic signal which is generated in the human brain as a result of a thought process. Automatic speech recognition has many speech formats to be recognized. Isolated words and continuous words speech formats are generally used as speech datasets. The continuous words speech recognition is complex than Isolated one due to improper boundary detection in the former. Another way of ASR classification is based on the speaker. The system is treated as speaker dependent if same dataset is used for training and testing purpose and vice versa. Speaker independent speech recognition system is more difficult to design. The input for automatic speech recognition is a speech signal that is processed to explore its benefits. Most of the speech recognizers use datasets that are available in foreign languages. The speech recognition can be done in any language for performance evaluation, but its utilization increases manifold when it is designed for native languages. In country like India most people belongs to villages and they are comfortable to communicate in Hindi because it gain

Corresponding Author: Ashok Kumar, National Institute of Technology, Kurukshetra, INDIA. Email iD-ashokgiri010@gmail.com.

wide acceptance in terms of understanding. The information received from acoustic speech signal is used for signal processing to extract its various features. Speech processing is considered at many levels. Frame processing, Phoneme Processing and processing at word level is performed to recognize the speech correctly. The accuracy of speech recognition generally depends on two factors; firstly, on right selection of features and secondly on selection of suitable classifiers. The combined efforts applied at feature extraction and training the speech recognizer with suitable parameters can enhance the performance of an automatic speech recognizer. Nowadays automatic speech recognition has wide applications like voice dialing, automatic appliance control, voice messaging and automation etc. There have been various developments in the field of speech recognition in last few years. Artificial Neural Network (ANN) with different features is proposed to be used here to assess the effects of different features on and to improve recognition rate of speech utterances. In [3], Mel frequency cepstral coefficients (MFCC) was introduced by Devis and Mermelstein that showed better performance due to better representation of perceptually relevant aspects of the short -term speech spectrum. The authors in [12], presented a new technique, perceptual linear predictive (PLP) technique for the analysis of speech based on three concepts of psychophysics of human hearing systems: (a) Critical band spectral resolution, (b) The equal loudness curve and (c) The intensity-loudness power law. In [14], authors compared different features and found that MFCC features provide better results as compared to that provided by other features. In [16], authors mentioned that MFCC features were more relevant and precise as compared to other time domain parameters. In [9], authors showed that MFCC parameters were more rigid to the theory of critical bandwidth of hearing as compared to PLP parameters. In [4], authors compared recognition rate using different feature extraction techniques (LPC, PLP and MFCC) and found that MFCC provided best recognition rate out of these three. The authors in [15], developed a speech recognizer for Bangla language using LPC features and different structure of Artificial Neural Network as classifier. In [1], authors showed that use of wavelet transforms increased the accuracy of speech recognition in clean as well as noisy environment. In [5], authors found that wavelet decomposition and reduced order linear predictive coding (LPC) could improve the speech recognition rate. In [6], the authors highlighted the progress made in feature extraction methods and overview of technological scope of an Automatic speech recognition. In [2], authors showed that Mel Frequency cepstral coefficients[MFCC] and perceptually linear perceptron [PLP] are more accurate to map speech features as compared to linear predictive Cepstrum as they use logarithmic scale despite linear scale used in LPC. The authors in [11] , explained about different features extraction techniques in sense of their application in field of speech recognition and concluded that MFCC can match human auditory system more accurately as compared to other feature extraction techniques. In [13], authors compared various features extraction techniques like Linear predictive cepstral coefficients (LPCC), Mel frequency cepstral coefficients (MFCC)

and Relative Spectral Analysis perceptual linear perceptron(RASTA-PLP) and found that recognition rate obtained with MFCC is better. In [7], author provided a broad coverage of use of Artificial Neural Network (ANN) in Speech recognition and compared their advantages and disadvantages. The authors in [8], compared the accuracy rates for speaker dependent and speaker independent using MFCC-VQ and MFCC-GMM and found that MFCC-GMM is comparatively better technique in case of both type of these recognitions. In [12], authors outlined speech recognition and different feature extraction techniques used in it. Comparison between different feature extraction technique showed that Mel frequency cepstral coefficients (MFCC) is more near to human auditory perception. The authors in [10], showed the comparative study of different features extraction techniques (LPC, MFCC and Spectrogram) using Artificial Neural Network (ANN) and noticed that MFCC-ANN provides the best result.

It is observed from above that most of the work in automatic speech recognition is done on isolated words despite recognizing speech. Foreign languages datasets are used for training rather than local Indian dialects and major languages. This motivated the present authors to create their own datasets in Hindi language for this research. The paper is organized as follows; 1st section covers the introduction, types and recent development in the field of speech recognition. 2nd section provides knowledge of various parameters considered for the study and proposed methodology adopted. 3rd section presents the result and discussion followed by conclusion at the end.

2. MATERIALS AND METHODS FOR AUTOMATIC SPEECH RECOGNITION SYSTEM (ASR)

This section provides general information regarding different features used in this research work with their importance and applications. It also discusses about the different types of neural networks and their use according to complexity of task. The method adopted here to get required result is explained under subheading proposed methodology. It also includes algorithm and flow that for better understanding of aim and procedures.

2.0.1. *Mel Frequency Cepstral Coefficients (MFCC).* Mel Frequency Cepstral Coefficients (MFCC) are widely used features, used in Automatic speech recognition. MFCC features are computed in frequency domain using Mel Scale. Mel scale is well tuned to human auditory system. Mel Scale frequency features are more precise as compared to other time domain features. Human speech signal cant be plotted in terms of linear scale of frequencies. So, it is required to convert it in Mel Scale. Mel scale is almost like, human auditory system. This scale is linear below 1Khz and logarithmic above it [12]. The given frequency in Hertz can be converted in Mel scale by using this formula(1)

$$F(Mel) = 2595 \log(1 + F/700) \tag{1}$$

Where F(Mel) is frequency in Mel scale and F is frequency in Hertz

2.0.2. *Perceptual Linear Perceptron (PLP).* The Perceptual Linear Prediction (PLP) model was developed by Hermansky in 1990. The concept of this model is based on psychophysics of hearing. PLP rejects the irrelevant data and so process the required one. This characteristics of PLP helps to improve the recognition rate. PLP is like LPC with a difference that it uses its spectral characteristics to map the characteristics of human auditory system. PLP approximates the human auditory system under three steps: 1. Critical Band Analysis 2. Equal Loudness Curve 3. Intensity Loudness (Power -Law Relation) . The main difference between PLP and Linear Predictive Coding (LPC) is that LP modelling technique considers all -pole transfer function of vocal tract for known number of resonances in given band. This way its equally approximates the power distribution at all frequencies. But speech signal cannot be estimated on linear scale thats why PLP is preferred over LPC [2].

2.0.3. *Linear Predictive coding (LPC).* In Linear Predictive Coding (LPC) , speech samples are approximated by combination of previous speech samples. A set of predictor coefficients can be obtained by minimizing the sum of squared difference between actual speech samples and linearly predicted ones. In this approach speech is modeled as the linear output excited by voiced or unvoiced speech. It uses all-pole model to find set of predictor coefficients that minimizes the mean squared error for only short span speech signal . For voiced speech signal this analysis provide a good approximation but for unvoiced speech it is less effective[6]. The predictor coefficients obtained this way describes the formants. Formants are frequencies at which resonant peak occurs [2].

2.0.4. *Artificial Neural Network (ANN).* Artificial Neural Network (ANN) is electronic model that is like neural structure of human brain. As human brain learns from past experiences similarly inter-connected neurons like human brains in artificial neural networks can classify speech data. A complex communication network can have hundreds of simple processing units that are wired together. Each simple unit is like real neuron which work according to the priority of work. There are mainly four types of neural network.

(1) Feedforward Network : Feedforward network is the simplest form of ANN. This network transfers the information in forward direction from input node to output node. This network does not involve any loop to send information from input to output.
(2) Recurrent Neural Network: It is a neural network that operates in time. It takes the in-vector form and respectively updates in hidden state by using nonlinear activation function to predict output. In this network output obtained is multiplied with a weight and feedback to input with a delay. RNN provides improved recognition rate as compared to MLP(Multi-layer Perceptron)but more complex.
(3) Modular Neural Network (MNN): As its name indicates, it has many modules. Each neural network is associated with a sub task of the

global task of the network. The global task is final application of neural network.

(4) Kohonen Self Organizing Maps: This type of network requires no supervision. So, it is also known as Self Organizing. They are also called as Maps because they try to map their weight according to input data. These types of network can learn from their own unsupervised competitive learning.

2.1. **PROPOSED METHODOLOGY.** First, speech dataset is created to input the speech recognizer. For this purpose, speech signal is accepted from person of different age groups and gender. The data set is prepared in Hindi language. The recorded speech signal is converted into digital form for signal processing. Then feature extraction is performed to get different features of speech signal like LPC, MFCC and PLP. After that features are classified using artificial neural network (ANN). The result is estimated using maximum likelihood function.

2.1.1. *Speech Datasets.* The Hindi language sentences from 6 different users are collected and each sentence is repeated 10 times, so we get $6 \times 6 \times 10 = 360$ speech samples. Speech samples are collected from peoples of different age group and gender of society to get robust speaker independent speech recognition system. Here, dataset is prepared in Hindi language due to lack of data availability in this language. For recording purpose, WO Mic client application is used to provide mobile- laptop Interface for Audacity . The sound is recorded by specifying the Audacity recording parameters to given value like sampling frequency = 16 KHz, Mono mode, type of coding is 16-bit PCM. Following table provided below describes the different specification:

TABLE 1. Speech Datasets

Sr. No.	Speaker Sex	Speaker Age (yrs.)	No. of Sentences Spoken(S)	No. of Repetition(R)	Total(T) = S*R
1	Male	40	6	10	60
2	Female	17	6	10	60
3	Male	17	6	10	60
4	Male	16	6	10	60
5	Female	18	6	10	60
6	Female	30	6	10	60

2.1.2. *Pre-Processing on speech samples.* Following steps were involved in the preprocessing on the acquired speech samples.

(1) Silence Removal
   (a) First, speech sample was divided into smaller frames of 16ms duration.
   (b) Then, short term energy was calculated for each frame.
   (c) Then all frames energy was normalized so that all frames energy comes in the range of (0, 1). This was obtained by dividing each frames energy by the maximum energy of frames.

(d) Then by thresholding operation, only those frames were selected which have energy greater than 0.02.
(e) Speech samples were reconstructed from the remaining available frames.

(2) Framing
Speech sample received after silence removal was divided into frames of 20ms for extracting the features from each frame.

(3) Frames Pre-processing
Prior to extracting the feature from frames, each frame was preprocessed by following steps.
(a) Windowing: Each frame was windowed by Hamming window.
(b) Pre-emphasis: Each frame was pre-emphasized by passing through a high pass filter with the following transfer function.

$$H(z) = 1 - 0.95z^{(-1)} \tag{2}$$

2.1.3. *Feature Extraction.* Finally, features are extracted from each frame. Different type of feature is selected for this study like LPC, MFCC, and PLP. To get the better performance individual features of MFCC and PLP are added to each other . Following table gives the information about length of feature vector for different studies.

Table 2. Different features with their lengths

Sr. No.	Features Name	Length of Features
1	LPC	10
2	MFCC	13
3	PLP	21
4	MFCC + PLP	(13+21) = 34

2.1.4. *Target Vector.* There was a target vector of 6x1 for each frame. In target vector number of rows represents the number of sentences to be identified. Further, row corresponding the sentence number contains 1 and remaining rows contain 0.

2.1.5. *ANN.* Separate ANN was created for studying the effect of different feature vector sets. Input layer of ANN contains neurons equal to the length of feature vector. Hidden layer and output layer of each ANN contained same number of neuron 151 and 6 respectively. ANN was trained using the built-in application in Matlab with trainlm training algorithm.

2.1.6. *Testing of ANN.* After training the ANN, each ANN was tested for the classification of sentences from each user. In order to check the efficiency of the ANN it was tested against the original data base. Following table gives the efficiency of ANN for different kind of feature selection method.
The algorithm for Hindi speech recognition using different feature extraction technique followed by ANN classification is given below:

2.2. **Algorithm for proposed methodology.**

Step 1: Initialize the different parameter used in speech processing e.g. sampling frequency, mode of recording, coding techniques, number of frames per samples and separation between frames.

Step 2: Record the speech samples using microphone for input datasets.

Step 3: Store the data in different folder for training and testing purpose to distinguish it easily.

Step 4: Remove silence zone from speech signal to improve accuracy of speech recognition system.

Step 5: Resolve the signal into manageable frames.

Step 6: Windowing is done to remove discontinuity between the frame using Hamming window.

Step 7: Pre-emphasis is performed using a high pass filter of specific transfer function.

Step 8: Feature extraction is done using various feature extraction techniques (LPC, PLP, MFCC).

Step 9: Compute target vector for each frame corresponding to feature extraction technique.

Step 10: Artificial neural network is used for classifying the samples for different features.

Step 11: Maximum likelihood ratio is calculated for training and testing.

Step 12: Result is defined in terms of recognition rate.

In the above algorithm input speech recorded from microphone is analog in nature so it is digitizing for further processing steps. Then different feature is extracted for their comparison. Artificial Neural Network(ANN) is used as a classifier with different feature extraction technique(LPC-ANN, PLP-ANN, MFCC-ANN) for comparison. Training and testing is performed on same text spoken by different speakers to find the accuracy for speaker independent system. Maximum Likelihood ratio is computed to get the result. The flow chart showing the individual step is given below.

2.3. **Flow chart of the proposed methodology.** Flow chart of the proposed methodology is shown in FIGURE 1.

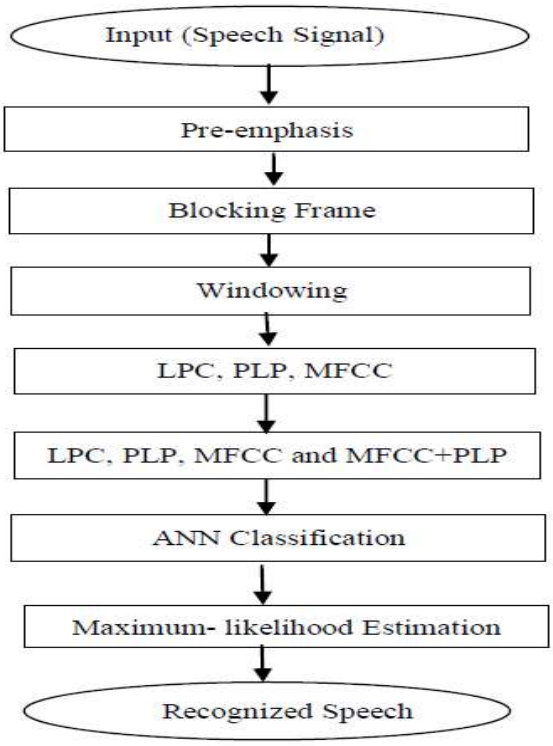

FIGURE 1. Flow chart for proposed methodology

3. RESULT AND DISCUSSION

The LPC_ANN based speech is analysed first for the given dataset. It is utilizing 10 linear coefficients of speech and the recognition efficiency reported is 55.55%. Secondly MFCC, 13 coefficients are used to find recognition rate for same datasets and its performance is better than LPC. The 21 coefficients of PLP parameter are used with ANN to model the recognizer that show the efficiency, 75% that lie in between LPC and MFCC. The hybrid technique(MFCC+ PLP)_ANN has provided the significant result of 94.44% by using 34 features 13 for MFCC and 21 for PLP.

TABLE 3. Efficiency obtained by different feature extraction technique

Sr. No.	Name of Features	Efficiency (%)
1	LPC	55.55
2	PLP	75
3	MFCC	86.11
4	MFCC + PLP	94.44

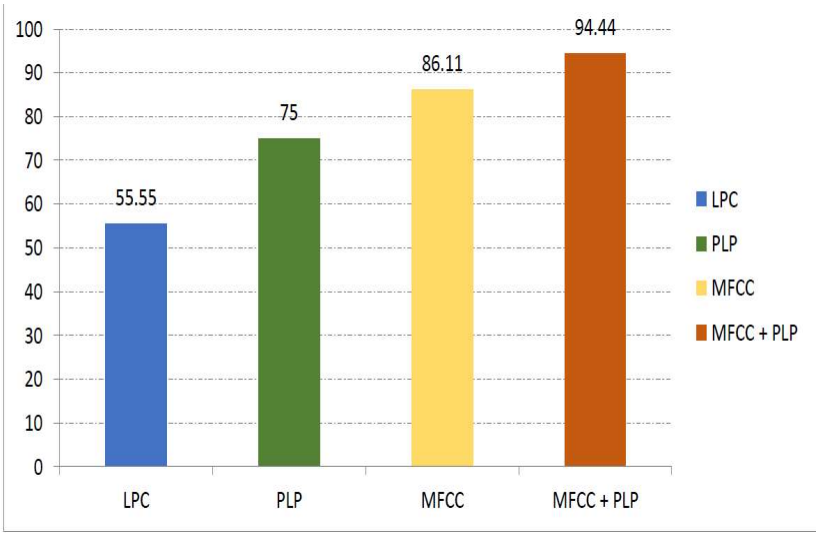

FIGURE 2. Efficiency of different features with ANN

From the study of TABLE 3 and FIGURE 2 it is observed that the result of recognition is maximum when MFCC and PLP parameter are added. Each feature extraction technique has its own pros and cons. As Linear Predictive Coding (LPC) uses linear scale for mapping of speech signal but speech signal is not basically linear, so it is not effective to recognized it well. As MFCC work on Mel- scale which is linear up to 1 KHz and after that its non-linear so its closer to human auditory system. Similarly, PLP also uses concepts of psychophysics of hearing thats its rejects the irrelevant data from speech and accepts the requires one which reduces computation and increase the response. So MFCC and PLP are better recognition technique as compared to LPC , which restrict its performance on linear scale. The artificial neural network is mostly used to design the speech recognizers due to its adaptive optimization of hidden layers. In the proposed system we have composed the feature of both techniques to get the better results.

4. CONCLUSION

Feature extraction is used to find stable and robust parameters of speech signal. But each technique has its own pros and cons according to their applications and characteristics. LPC provides good result for linear range of speech signal so its use is limited in area of speech recognition. As PLP and MFCC has gained a wide acceptance due to their better match to human auditory system so proposed methodology adopted a hybrid technique by combining the feature of individuals to enhance the recognition rate. The proposed system utilizes the combination of MFCC and PLP and show the improvement in recognition by 8.33% as compared to MFCC alone.

References

1. M. Anusuya and S. Katti, *Comparison of different speech feature extraction techniques with and without wavelet transform to kannada speech recognition*, International Journal of Computer Applications **26** (2011).
2. Namrata Dave, *Feature extraction methods lpc, plp and mfcc in speech recognition*, International Journal For Advance Research in Engineering And Technology(ISSN 2320-6802) **Volume 1** (2013).
3. S. Davis and P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Transactions on Acoustics, Speech, and Signal Processing **28** (1980), no. 4, 357–366.
4. C. O. Dumitru and I. Gavat, *A comparative study of feature extraction methods applied to continuous speech recognition in romanian language*, Proceedings ELMAR 2006, 2006, pp. 115–118.
5. Hynek Hermansky, *Perceptual linear predictive (PLP) analysis of speech*, Journal of the Acoustical Society of America **87** (1990), no. 4, 1738–1752.
6. B. Jolad and R. Khanai, *International journal of engineering sciences & research technology different feature extraction techniques for automatic speech recognition: a review*, **7** (2018), no. 2, 181–188.
7. Bhushan C Kamble, *Speech recognition using artificial neural network - a review*, **3** (2016), no. 1, 1–4.
8. Ankur Maurya, Divya Kumar, and R. K. Agarwal, *Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach*, Procedia Computer Science **125** (2018), 880–887.
9. Ludk Mller and Josef Psutka, *Comparison of mfcc and plp parameterizations in the speaker independent continuous speech recognition task.*, 01 2001, pp. 1813–1816.
10. H M Mohammed, M S Alkassab, H R Mohammed, H Abdulaziz, and Ahmed S Jagmagji, *Speech Recognition System with Different*, (2018), 2003–2006.
11. Shreya Narang and Ms Divya Gupta, *Speech Feature Extraction Techniques: A Review*, International Journal of Computer Science and Mobile Computing **4** (2015), no. 3, 107–114.
12. Navnath S. Nehe and Raghunath S. Holambe, *DWT and LPC based feature extraction methods for isolated word recognition*, Eurasip Journal on Audio, Speech, and Music Processing **2012** (2012), no. 1, 1–7.
13. P Prithvi and T Kishore Kumar, *Comparative Analysis of MFCC, LFCC, RASTA-PLP*, International Journal of Scientific Engineering and Research **4** (2016), no. 5, 1–4.
14. D. A. Reynolds and R. C. Rose, *Robust text-independent speaker identification using gaussian mixture speaker models*, IEEE Transactions on Speech and Audio Processing **3** (1995), no. 1, 72–83.
15. Shakil Sumon, Joydip Chowdhury, Sujit Debnath, Nabeel Mohammed, and Sifat Momen, *Bangla short speech commands recognition using convolutional neural networks*, 11 2018.
16. R. Vergin, D. O'Shaughnessy, and A. Farhat, *Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition*, IEEE Transactions on Speech and Audio Processing **7** (1999), no. 5, 525–532.

National Institute of Technology, Kurukshetra-136119, Haryana, INDIA
E-mail address: ashokgiri010@gmail.com

National Institute of Technology, Kurukshetra-136119, Haryana, INDIA
E-mail address: vikasmittalkkr@gmail.com