OPTIMIZED FILTER BANKS TO IMPROVE PUNJABI SPEECH RECOGNITION IN NOISY ENVIRONMENT

Pooja Rani*[1], Ashish Saini[2], Vikas Mittal[3]

[1]Department of Computer Science & Engineering, Quantum University, Roorkee 247667 Uttarakhand, India
[2]Department of Computer Science & Engineering, Quantum University, Roorkee 247667 Uttarakhand, India
[3]Department of Electronics & Communication Engineering National Institute of Technology Kurukshetra 136119, Haryana, India

Abstract *The performance of Automatic Speech Recognition (ASR) depends on its capability to identify the test patterns best-matched with training patterns in various classes. This matching is highly dependent upon the individual feature extraction technique or combination thereof. Certain advanced feature extraction techniques such as GFCC, BFCC have been reported in the literature (with associated additional problems of accepted bandwidth and optimal number of features) in addition to the commonly used ones such as Mel Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) coefficient. MFCC is more suitable for clean environments while PLP performs better when there lies a significant mismatch between training and testing phase. Therefore, this paper proposes a minimalistic approach involving hybrid features (i.e., MFCC+PLP) to overcome shortcomings of each constituent, such as sensitivity to background noise on one hand, and avoid complexity in extracting advanced features, such as GFCC and BFCC etc. on the other hand. These hybrid features can provide favourable or comparable results as compared to those obtained using advanced features in both clean and noisy environments. The other problem of optimizing the number of filter banks for a specified bandwidth is proposed to be accomplished using an evolutionary technique like DE (Differential Evolution) to enable suitable comparisons with the existing literature. Additionally, an advanced classifier viz. Deep Neural Networks (DNN) is used as compared to ones that are more conventional such as Hidden Markov Model (HMM) used in the literature for further improvisation.*

Keywords: *Automatic Speech Recognition, MFCC, PLP, Differential Evolution and DNN.*

1. INTRODUCTION

Speech is primary mode of communication between human beings. Automatic speech recognition in its basic form is used for speech to text conversion using machines [1]. Researchers across the world have tried to minimize this gap between man and machines by the use of efficient techniques. But random behaviour of noise makes the recognition task more difficult and is still a challenge [2-5].

The right selection of language is a prime concern while designing a speech recognizer, since it is of more utility if it handles local languages such as Punjabi in Indian context [6,7]. For instance, design of interactive applications, spreading and carrying the benefits of various schemes to the lowest strata in society become easier, which would otherwise be a herculean task in foreign languages given the lack of education and computer literacy among masses [8, 9]. Hence, Hindi speech recognizers are expected to solve these sorts of problems.

Additionally, different forms of speech based on the type of utterances such as Isolated words, connected words, Continuous and spontaneous speech have a bearing on the design of a speech recognizer. As continuous speech is closer to natural way of human speaking [10,11,12], various studies considered this speech format to analyze the performance of a practical speech recognition system. The absence of efficient technique to find start and end point detection in continuous speech is still an important research gap [13]. Therefore, continuous speech in noisy environment is considered in this paper.

Another important factor for performance evaluation is noise that is an external and uncontrollable parameter. The recognition of speech is easier in clean environment as compared to the noisy one. In general, noise reduction is an important factor for designing a robust speech recognition system [14]. Noise compensation is one such technique, which requires a priori knowledge of its characteristics. Therefore, speech detection and enhancement algorithms can be used alternatively to obtain desired speech quality with maximum accuracy [15,16,17]. But this task is very difficult due to irregular noise boundaries, which can cause loss of desired signal due to improper demarcation between the two [18]. Hence, selection of proper features or their robust extraction techniques can play a vital role in improving the performance of a speech recognition system in noisy environment [19, 20].

There are various feature extraction techniques used for speech signal such as Linear Predictive Coding (LPC), Mel Frequency Cepstral coefficient (MFCC), Perceptual Linear Prediction (PLP) and Wavelet Transform etc. with their own advantages and disadvantages [21,22,23,24,25]. For instance, most of the existing speech recognition systems utilized MFCC as they perform better in clean environment and when there is no considerable mismatch between training and testing data. On the other hand, PLP features are more robust to the environment varying due to noise and other external means [26]. There are also some advanced features like GFCC and BFCC that are more robust to noise than MFCC and PLP individually, but they entail extra computational complexity [20]. However, the combination of (MFCC+PLP) features can result in performance comparable to these advanced features (GFCC & BFCC) both in clean and noisy environment. Therefore, minimalist features based on both MFCC and PLP are proposed to be used in this paper.

The second factor for robust performance is the optimal number of filters and bandwidth [20]. The assigned bandwidth for filter bank is very important to improve the performance of speech recognition in noisy environment. It has been experimented that optimized filter bank can lead to reduction in error for both MFCC and PLP for variety of tasks [27]. There are number of techniques that are used for optimization of the filterbank such as Particle Swarm Optimization (PSO), Genetic Algorithm (GA) and Differential Evolution (DE) to name a few [28,29,30]. DE is improved form of Genetic algorithm in-terms of optimization quality and speed [31]. This prompted the authors to apply this algorithm to optimize the required features in this paper.

In [32], authors proposed speech enhancement method based on Deep Neural Network (DNN). The results show that the pre-processing methods in the suggested framework enhanced voice recognition. Using the Deep Belief Network (DBN) model, Reference [33] presented a novel method for speech segregation in unlabeled stationary noisy audio signals. Music signals were successfully separated from noisy audio streams using this technique. In the meantime, reference [34] provided a thorough analysis of numerous studies that used Deep Learning in the field of automatic speech recognition and were carried out between 2006 and 2018. A N-gram modeling technique was presented [35], to recognize Punjabi continuous speech recognition using Kaldi speech recognition toolkit. In [36], authors have compared the performance of DNN over GMM and observed better performance of DNN for large training datasets. The author of [37] have applied DNN modeling to Punjabi children's speech corpus and reported a improved accuracy rate up to 87%. In [38], authors have proposed corpus optimization model on Punjabi datasets to reduce the word error rate by 5.8%.

A systematic review of 76 research paper [39], in the field of children speech recognition is discussed to analyze the potential trending of different speech recognition techniques. In [40], authors have addressed various challenges faced by the researchers in the field of Automatic Speech Recognition such as multilingual translation, emotion recognition and Human computer Interface. A language space denoising technique was discussed in [41], to design noise –robust speech recognition system using large language models.

This conversation makes it clear that there hasn't been much research done on Punjabi voice ASR in noisy settings. To improve recognition accuracy, the majority of previous research has used the Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), or their hybridizations (HMM+GMM). Thus, there is a great deal of opportunity to investigate the application of Deep Neural network

The paper is organized as follows; Second section introduces and explains the nomenclature and terms involved in this paper. Third section discusses the proposed methodology. The fourth section presents and illustrates various results for MFCC, PLP and hybrid (MFCC+PLP) feature extraction techniques in noisy environment with DE and without DE optimization followed by conclusions at the end.

2. MFCC, PLP and Differential Evolution

2.1 Mel Frequency Cepstral Coefficient (MFCC)

MFCC is a well-known feature extraction technique due to its broad coverage of distinct features of a speech signal. It approximates the characteristics of input signal following the steps as [17]:

- Pre-emphasis of speech signal is required to raise the signal to noise ratio (SNR) at high frequency component to reduce the effect of noise.
- Windowing (Hamming Window) is performed to provide continuity to the frames at edges. Mostly Hamming window is used to perform this task using the expression

$$W(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N} - 1\right) & 0 \le n \le N \\ 0 & otherwise \end{cases} \quad (1)$$

Where W(n) is Hamming Window, n are samples considered out of total N samples.

- Further Discrete Fourier Transform (DFT) is applied to transform into frequency domain as

$$X(K) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N} \quad 0 \leq n \leq N\text{-}1 \tag{2}$$

Then band pass filtering is done to approximate the power spectrum of each frequency band using

$$P(K) = \frac{1}{N} \left| X(K) \right|^2 \tag{3}$$

- Logarithmic Mel-Scaled filters bank converts frequency to Mel scale to map human auditory system as

$$M(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \tag{4}$$

Finally, Discrete Fourier Transform (DCT) is performed to generate Mel Frequency cepstral coefficients as

$$F(u) = \sqrt{\frac{2}{N}} \sum_{i=0}^{N-1} A(i) \cos \left[\frac{\Pi u}{2N} (2i+1) \right] F(i) \tag{5}$$

There are 13 MFCC features computed from above steps out of which 12 features are computed using DCT transform and one energy feature from the frame. This energy feature is important because it show the correlation between identities of frames in terms of energy. A different set of 13 MFCC features are obtained by performing first order derivative on the above derived feature vector. Therefore, a feature vector of 26 coefficient values is obtained by combining the two sets. A final feature vector of 20 coefficients is selected out of the above 26 coefficients for signal processing. These limited numbers of coefficients are used to maintain uniformity of study at reduced computational complexity. The additional features obtained using derivative method help defining non-uniform nature of the speech.

2.2 Perceptual Linear perceptron (PLP)

Based on the ideas of auditory psychophysics, the Perceptual Linear Prediction (PLP) model is a speech analysis method. By highlighting perceptually important information and eliminating extraneous material, it improves speech recognition. By using spectral modifications, PLP more accurately mimics the properties of the human auditory system than Linear Predictive Coding (LPC). The following three crucial processes are used to achieve this distinction:

key Band Analysis: To represent the human ear's capacity for frequency resolution, the speech signal is separated into key frequency bands.
Equal Loudness Curve: The spectrum has been tuned to human hearing's sensitivity to frequencies.
Intensity-Loudness Transformation: Using a power-law function, a non-linear transformation simulates the link between perceived loudness and sound intensity in contrast to LPC, which uses a linear scale to evaluate speech.

- After that Band pass filtering is done to approximate the power spectrum of each frequency band as

$$P(K) = \frac{1}{N} \left| X(K) \right|^2 \tag{6}$$

- Then audio frequency is converted to Bark Scale for better mapping of human auditory process as

$$\Omega(\omega) = 6 \ln \left[\frac{\omega}{1200\Pi} + \left[\left(\frac{\omega}{1200\Pi} \right)^2 + 1 \right]^{0.5} \right] \tag{7}$$

The frequency analysis is in line with the human auditory perception scale to the use of the Bark filter bank for spectral warping. To improve the spectrum, its output is smoothed and down-sampled. Then, with consideration for the equal loudness contours, an emphasis is placed on the smoothed spectrum to improve aural perception. To further resemble the properties of human hearing, these modified values are then increased in accordance with the power-law relationship.

- Finally, LP Model is applied to predict the feature coefficients by mapping the power spectrums P(W) and P'(W) as

$$\frac{1}{M}\sum_{m=1}^{M}\frac{P(\omega)}{P'(\omega)}=1 \tag{8}$$

Where $P(\omega)$ and $P'(\omega)$ are input and predicted power spectrum of the speech signal.

The first steps of MFCC feature extraction are the same as those in PLP, including windowing and using the Fourier transform. But PLP uses the Bark scale, whereas MFCC models the human auditory system using the Mel scale. Furthermore, PLP enhances weaker signal components by using an equal loudness function prior to linear prediction. PLP uses trapezoidal filters, as opposed to MFCC's triangle filters. Similar to MFCC, where the first-order derivative yields an extra feature vector of 13 coefficients, the first 13 coefficients of PLP features are computed using recursive cepstral calculation. A final feature vector with 26 coefficients is created by combining them with the original vector.

2.3 Differential Evolution (DE)

Differential Evolution is an optimization technique used to find optimum solution of a mathematical function. It was introduced in year 1997 by Storn and Price [42]. It is based on natural selection of elementary population to optimize its solutions. The multiple iterations are performed to find most suited solutions for a given problem. Some primary parameters are required to be initialized, before possible run of this evolutionary algorithm to optimize a filter bank for extracting features [43]. It involves the following nomenclature;

- **Population size:** The number of filters is treated as population. Higher number of filters represents larger population.
- **Population initialization:** It talks about the numbers of filters used initially.
- **Mutation:** It is a process to generate new filters for optimization.
- **Crossover**: It refers to generating trial filter by adding new filters with the help of crossover operators.
- **Selection**: It refers to ranked selection method to find filter with least fitness.

DE is an optimization technique which provides the best fitted result for the initially selected parameters and range. Here DE algorithm is adopted to optimize the number of filters. Various numbers of filters are tried in range of [80-120] but most optimized result is obtained at 100 for mutation rate of 0.06 and crossover of 0.6.

3. Proposed methodology

The proposed Automatic Speech Recognition (ASR) system utilizes hybrid features (MFCC+PLP) at front end. Differential Evolution (DE) is an evolutionary approach applied to optimize number of filter banks for better performance. Deep Neural Network (DNN) is used as a classifier to train and test the patterns of speech samples to improve the recognition rate.

3.1 MFCC and PLP Filter bank

Feature extraction is an important step in Automatic speech recognition. The robust features provide immunity to noise and environmental variations. Pre-processing is the first building block that is used for digitization, windowing and removing noise from input speech signal. The selection of a specific filter depends on the characteristics of noise.

There are various feature extraction techniques but MFCC has wide utilisation in clean environment. Mel filter bank is a type of triangular band pass filter that is linear below 1000 Hz and non-linear above it [44]. In Fig.1, frequency analysis for range 0 to 8 kHz is shown using Mel frequency filter bank. Mel filters are distributed in a band such that lower edge of a filter acts as the centre frequency for the previous filter. MFCC uses Mel-filter bank while PLP utilises bark scale to extract the features.

According to Hermansky, Bark scale filter bank are placed such that their centre frequencies are 1 Bark apart from each other [24]. Fig. 1 and Fig. 2 shows MFCC and PLP features at sampling frequency of 16 kHz.

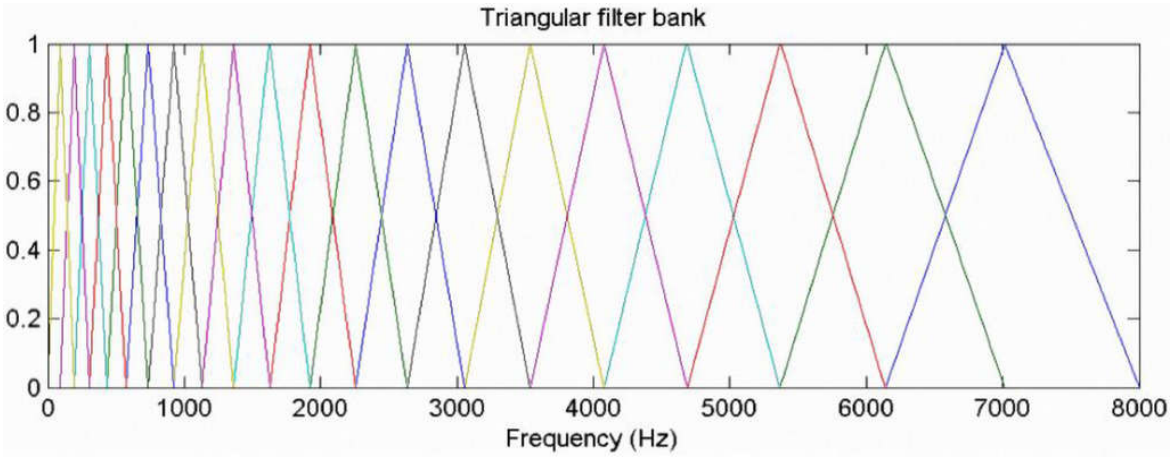

Fig. 1 Mel Frequency filter bank (MFCC)

Both MFCC and PLP show nearly similar responses since they both have the capability to adapt to non-linear region of human hearing. But PLP parameters are comparatively more robust to noise and varying environmental effects. Therefore, in this paper, both MFCC and PLP are combined to harness the benefits of each.

The purpose of DE optimization is to find optimum number of filter banks and adjust the filter spacing for both feature extraction techniques. The optimized features help in the better utilisation of accepted bandwidth for improved the accuracy of recognition.

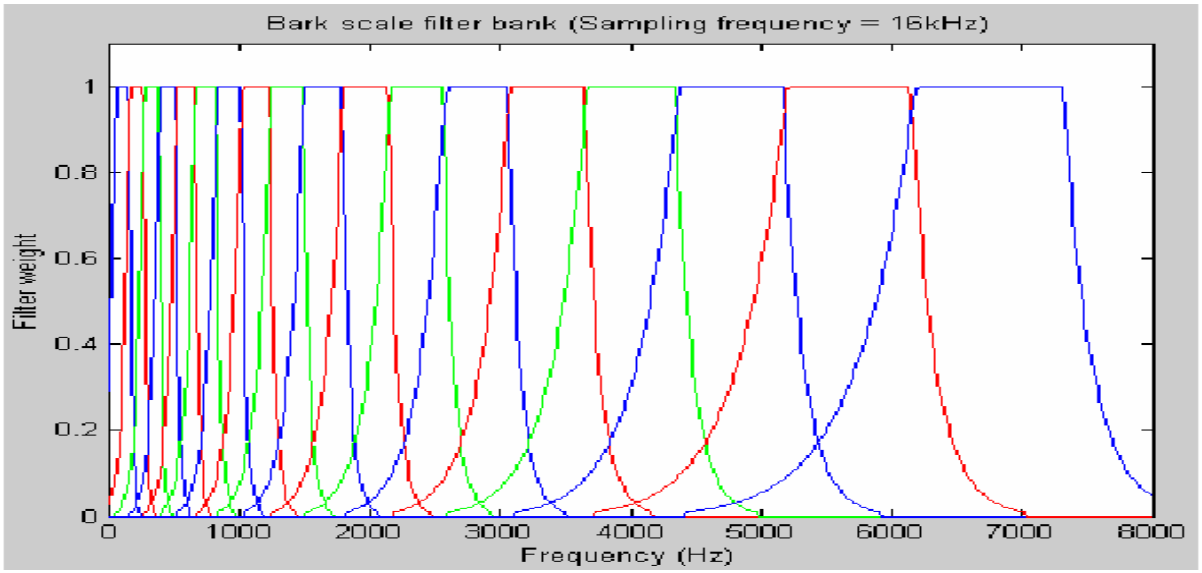

Fig.2 Bark Frequency Filter bank (PLP)

3.2 DE optimized filter bank

Optimization using DE starts with random selection of a few numbers of filters to define initial population. These are represented by chromosome to keep them distinct. The selection of chromosomes depends upon fitness value that is decided by mutation process. Further, crossover operation is performed on the population to compute fitness of newly generated filter bank. The performance of DE algorithm depends on a number of parameters like population, mutation and crossover that are to be set initially. Fig. 3 shows steps involved in extracting DE optimized MFCC and PLP features. In this paper, population size, Mutation and crossover are initialized as 100, 0.06 and 0.6 as determined after extensive experimentation.

Pre-processing of speech signal covers pre-emphasis and windowing to amplify energy of high frequency component in comparison to low frequency and remove discontinuity at edges. Then DFT is applied to compute energy within each frequency band [45]. Mel and Bark filter banks are optimized by DE for their spacing and numbers. Log-compression is used to map the loudness of human perception. Similar work is performed in PLP by the loudness compression. Inverse Discrete Fourier Transform (IDFT) is applied to get coefficients back in the time domain.

3.3 Deep Neural Network (DNN)

Deep Neural Networks (DNN) are collection of neurons interconnected to each other. The neurons are further classified into layers. The input layer corresponds to data for classification. The output is computed from weighted sum of its input.
The nature of DNN depends on types of its parameters; Generally, DNN is operated by three types of parameters: Pattern of interconnections, Training process to update the weights and activation function [46]. The purpose of training DNN is to improve the recognition up to the target value. In this research work extracted features are treated as input to DNN for the task of classification.

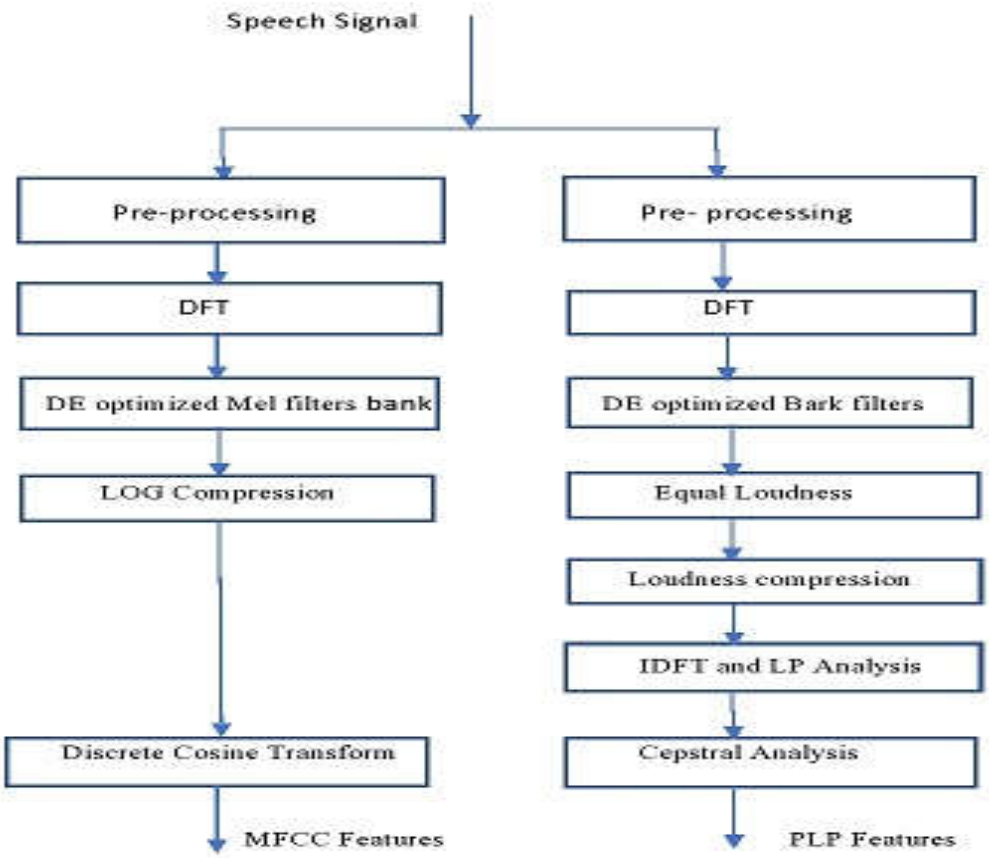

Fig. 3 Steps to extract DE optimized MFCC and PLP features

In Fig. 4 Basic building blocks of DNN based automatic speech recognition are shown. The input speech signal is processed for extracting MFCC and PLP features. Filter banks for each technique are optimized using DE evolutionary techniques beforehand. Further, Individual filters are combined to get hybrid features. Deep Neural network (DNN) is used as classifier to train and test the signal. Multilayer perceptron (MLP) architecture of DNN is adopted for classification.

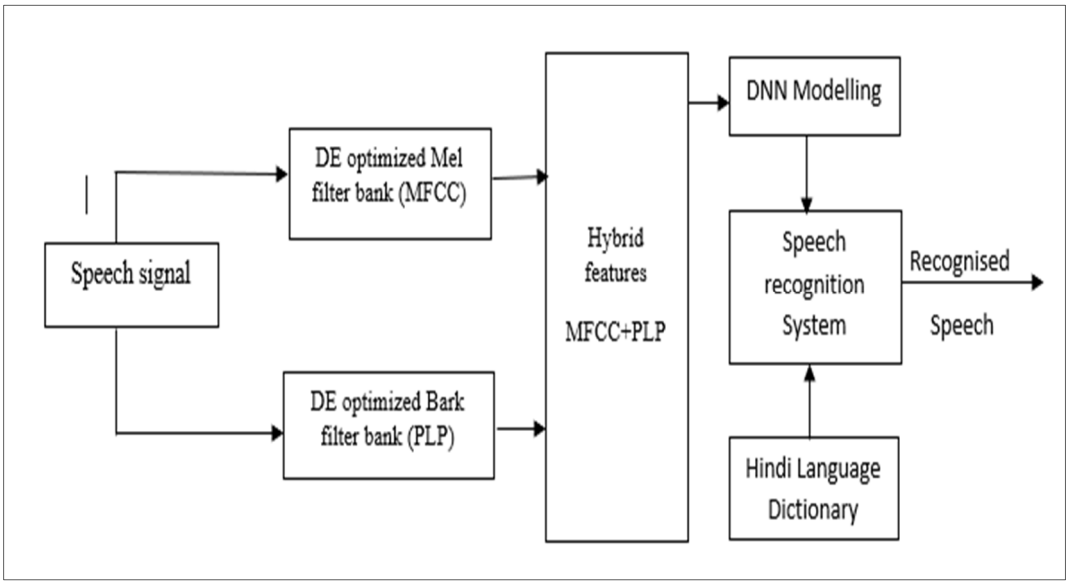

Fig. 4 Proposed Methodology

3.4 Experimentation Datasets

The first requirement for performing an experiment is input database and its broad acceptance. So, database is created in Punjabi language from peoples of different age group and gender.

3.4.1 Speech Datasets: Punjabi speech signals of different speakers are acquired by authors using Audacity 2.3.2. It is open-source software for audio editing and recording applications [47]. 3 males and 3 females of different age group are selected for the purpose. The acquired datasets are described in Table 1.

Table 1: Speech datasets

S. No.	Gender	Age	No. of samples (S)	No. of repetition (R)	Total samples (T= S*R)
1.	Male	40	100	2	200
2.	Male	17	100	2	200
3.	Male	16	100	2	200
4.	Female	17	100	2	200
5.	Female	18	100	2	200
6.	Female	30	100	2	200

A total of 1200 voice samples were recorded, which were then divided into two sets: 75% of the samples were used for training, while the remaining 25% were set aside for testing. The recordings were made using the WO Mic client interface connected to Audacity, a tool that allows for sound recording, clipping, storage, and mixing. The speech signals were captured with the following parameters:

Sampling frequency (Fc): 16 kHz

Coding technique: PCM

Recording mode: Mono

Bit rate: 16 bits/sec

The recorded speech data was mixed with various environmental noises (such as car, fan, and diesel engine sounds) to examine their impact on Automatic Speech Recognition (ASR). The noise samples were sourced from online platforms, as detailed in the following subsection.

3.4.2 Noisy Database

Noise samples from car engines, diesel engines, and fans were collected from freesound.org and NOISEX-92, which offers various sounds for research purposes. These noises were combined with the clean speech samples to create noisy speech data with Signal-to-Noise Ratio (SNR) values ranging from 0 to 15 dB in 5 dB intervals. The resulting noisy speech signals were used to train and evaluate the performance of speech recognition systems in noisy conditions.

To adjust the SNR, the noise reduction feature in the 'Effect' menu of Audacity was used. This involved opening two separate windows, one for the signal and another for the noise, and then blending portions of the noise into the signal to generate noisy speech data.

4. Results and Discussion

The performance of the proposed Automatic speech recognition (ASR) systems is analysed in three parts. The first stage comprises the study of MFCC features individually with and without optimization of filter banks using DE. The second stage utilizes PLP features individually with and without DE optimization. The final stage includes the performance of hybrid features. The Deep Neural Network (DNN) is used universally in all above stages to classify the speech signal. These three stages will help in appreciating the effects of features used individually and collectively. Also, the effect of DE optimization on the performance of proposed ASR can be assessed.

4.1 Performance analysis using only MFCC

Table 2 and 3 below display the voice recognition system's accuracy using the MFCC function with and without DE. The tables show that recognition ability improves with an increase in the signal to noise ratio. When automobile noise is present, the recognition rate is at its highest; when diesel engine noise is present, it is at its lowest. By comparing the recognition rate at 0 dB, the cause of this performance shift can be examined. Car noise is at its best and diesel engine noise is at its lowest at 0 db. Following that, there is a slight variation in the results of up to 15 dB, but the average results stay the same. Different noises are the source of this significant shift in diesel engine and automobile noise.

Table 2: Accuracy [%] without DE using MFCC

Type of noise	Signal to Noise Ratio (db)				Average (15 db to 0 db)
	15	10	5	0	
Car	85.6	66.8	43.1	23.2	54.675
Fan	84.2	66.3	41.7	22.1	53.575
Diesel Engine	82.5	61.8	38.3	11.8	48.60

Table 3: Word accuracy [%] with DE using MFCC

Type of Noise	Signal to Noise Ratio(db)				Average (15db to 0 db)
	15	10	5	0	
Car	88.5	77.8	55.8	34.8	64.22
Fan	85.8	78.3	55.8	32.8	63.17

Diesel Engine	90.8	75.3	53.2	31.8	62.77

It is observed from Table 3 that there is a considerable improvement in accuracy of speech signal using DE for all type of noises. Maximum recognition is obtained for diesel Engine noise between 10 to 15 db. It is also noticed that (0-5) dB and (5-10) db show almost similar improvement. This may be caused due to attainment of approximate signal strength for recognizer about 10 db so comparatively least margin for further improvement. From Table 2 and Table 3, it is observed that diesel engine noise using DE has maximum performance improvement of 14.17 percent. The least improvement in recognition i.e. 9.5% is noted for car noise. The fan noise has reported moderate rise in accuracy by 9.60%. On comparing Tables 2 and 3 it is also found that DE algorithm has a better impact on the noise type providing least result without it and vice versa for the reasons mentioned above.

Table 4 Accuracy without DE using PLP

Type of noise	Signal to Noise Ratio (db)				Average (15 db to 0 db)
	15	10	5	0	
Car	86.3	67.2	42.8	23.3	54.90
Fan	84.4	67.3	40.7	23.8	54.05
Diesel- Engine	83.5	62.5	38.3	12.4	49.17

Table 5 Accuracy with DE using PLP

Type of Noise	Signal to Noise Ratio (db)				Average (15 db to 0 db)
	15	10	5	0	
Car	89.5	77.5	56.8	35.6	64.85
Fan	91.5	75.6	58.9	34.5	65.12
Diesel- Engine	90.4	72.4	50.8	31.5	61.27

4.2 Performance Analysis using PLP (Perceptual Linear Perceptron)

It is observed from Tables 4 and 5 that rise in signal to noise ratio is equally effective in both cases i.e., with and without DE optimized filter bank. As signal to noise ratio increases, there is corresponding improvement in performance. DE optimized filter banks show drastic improvement in accuracy at all db values ranging between 0 to 15 db. The performance of PLP feature is slightly better than MFCC, as PLP features are more robust in noisy or time varying environments. The hike in improvement follows the patterns similar to MFCC for all three types of external noise.

4.3 Performance Analysis using MFCC_PLP

The proposed technique utilizes the hybrid features (MFCC_ PLP). It is observed from Tables 6 and 7 that accuracy of hybrid features is quite better as compared to MFCC and PLP individually It is observed from Tables 2,3,4,5 and 6,7 that hybrid technique provides better improvement in the interval 10 to 15 db due to presence of PLP features that are known to perform better in noisy environment.

Table 6 Accuracy [%] without DE using MFCC_PLP

Type of noise	Signal to Noise Ratio (db)				Average (15 db to 0 db)
	15	10	5	0	
Car	88.2	69.7	45.6	25.4	57.22
Fan	87.3	68.7	43.5	23.9	55.85
Diesel- Engine	86.3	63.6	39.2	13.6	50.67

Table 7 Accuracy with DE using MFCC_PLP

Type of noise	Signal to Noise Ratio (db)				Average (15 db to 0 db)
	15	10	5	0	
Car	95.5	79.8	54.9	33.9	66.02
Fan	94.8	78.2	56.8	34.8	66.10
Diesel- Engine	92.4	75.8	55.8	30.6	62.60

From Tables 6 and 7, it is observed that the performance shows a steep hike even at 0 db level. The improvements for car and fan noise are almost similar while diesel engine noise follows the earlier pattern. There is also improvement for fan noise which was less affected for MFCC features due to better effectiveness of hybrid features in clean and noisy environment. From Tables 3 and 5, it is observed that Performance of PLP features are slightly better than MFCC with and without DE optimization. The improvements in recognition rate follow the earlier patterns for all three type of noise and the features. The proposed hybrid feature extraction technique shows improvement in performance on an average by 12.62% compared to MFCC and 12.70 % compared to PLP individually without DE optimization.

Conclusion

Differential Evolution and hybrid features (MFCC and PLP) are shown to improve the performance of Punjabi speech recognition in noisy environments. DE algorithm has been successfully applied to optimize the number of filter banks for better performance. Hybridization of MFCC and PLP features provide robust features. Various analysis is performed to evaluate the performance of proposed system in noisy environments. The results showed that hybrid features along with DE optimization provide better result as compared to MFCC and PLP individually. This work can be further extended by integrating advanced features like GFCC and BFCC along with speech enhancement algorithms.

REFERENCES

[1] S. S. Bhabad and G. K. Kharate, An Overview of Technical Progress in Speech Recognition, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, issue 3, pp. 488-497, March 2013

[2] Keronen, S., Remes, U., Palomaki, K. J., Virtanen, T. & Kurimo, M. (2010). Comparison of Noise Robust Methods in Large Vocabulary Speech Recognition, 18th European Signal Processing Conference (EUSIPCO-2010), 1973-1977.

[3] J.M. Baker, L. Deng, J. Glass, S. Khudanpur, C.H. Lee, N. Morgan, O'Shaughnessy, Developments and directions in speech recognition and understanding, Part 1 [DSP Education], IEEE Signal Process Mag. 26 (3) (2009).

[4] G. Saon, J.T. Chien, Large-vocabulary continuous speech recognition systems: a look at some recent advances, IEEE Signal Process Mag. 29 (6) (2012) 18–33.

[5] A. Acero, Acoustical and Environmental Robustness in Automatic Speech Recognition, Springer Science & Business Media, New York, 1993.

[6] R. K. Aggarwal, M. Dave, Implementing a speech recognition system interface for Indian languages, Proceedings of the IJCNLP-08 Workshop on the NLP for less Privileged Languages, pp.105-112, Hyderabad, India, January 2008.

[7] K. Kumar, R. K. Aggarwal and A. Jain, A Hindi Speech recognition system for connected words using HTK, International Journal of Computational Systems Engineering, vol. 1, no.1, pp. 25-32, 2012.

[8] P. Saini, P. Kaur and M. Dua, Hindi Automatic Speech Recognition using HTK", International Journal of Engineering Trends and Technology, vol. 4, issue 6, pp. 2223-2229, June 2013.

[9] P. P. Shrishrimal, R. R. Deshmukh and W. B. Waghmare, Indian Languages Speech Database: A Review, International journal of Computer Applications (0975-888), vol. 47, no. 5, pp. 17-21, June 2012.

[10] S. N. S. and R. R. Deshmukh, Speech Recognition System- A review, IOSR Journal of Computer Engineering (IOSR-JCE), vol. 18, issue 4, Ver. II (July-August) 2016, pp. 01-09.

[11] P. Gautam, Survey of Speech Recognition System, International Journal of Advance Research and Development", vol. 2, issue 4, pp. 127-130, 2017

[12] P. Saini, P. Kaur, Automatic Speech Recognition: A Review, International Journal of engineering Trends and Technology, vol. 4, issue 2, pp. 132-136, 2013.

[13] P. K. Kurzekar, R. R. Desmukh, V. B. Waghmare & P. P. Shrishrimal Continuous speech Recognition System: A Review, Asian Journal of Computer Science and Information Technology (AJCSIT), vol.4, issue 6, pp. 62-66, 2014.

[14] B. T. Lilly and K. K. Paliwal, Robust speech recognition using singular value decomposition-based enhancement", Speech and Image Technologies for Computing and Telecommunications, pp. 257-260, 1997.

[15] Cui, X. & Alwan, A. (2005). Noise Robust Speech Recognition using Feature Compensation Based on Polynomial Regression of Utterance SNR, IEEE Transactions on Speech and Audio Processing, 13(6), 1161-1172.

[16] M. Khademian, M. M. Homayounpour, Factorial Speech Processing Models for Noise-Robust Automatic Speech Recognition, 23rd Iranian Conference on Electrical Engineering (ICEE), pp. 637-642, 2015.

[17] V. Ion and R. H. Umbach, A Novel Uncertainty Decoding Rule with Applications to Transmission Error Robust Speech Recognition, IEEE Transactions on Audio Speech, Language Processing, vol. 16, no. 5, pp. 1047-1060, July 2008.

[18] Y. Shao and C-H. Chang, Bayesian Separation with Sparsity Promotion in Perceptual Wavelet Domain for Speech Enhancement and Hybrid Speech Recognition, IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans, vol. 41, no. 2, pp. 284-293, March 2011.

[19] R.K. Aggarwal, M. Dave, Acoustic modeling problem for automatic speech recognition system: advances and refinements (Part II), Int. J. Speech Technol.14 (4) (2011) 309.

[20] M. Dua, R. K. Agarwal & M. Biswas, Performance evaluation of Hindi speech recognition system using optimized filter banks, Engineering Science and Technology an International Journal, 21(2018), 389-398.

[21] D.A. Reynolds, Experimental evaluation of features for robust speaker identification, IEEE Trans. Speech Audio Process. 2 (4) (1994) 639–643.

[22] H. Hermansky, S. Sharma, Temporal patterns (TRAPS) in ASR of noisy speech, in: 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. IEEE, Phoenix, AZ, 1999, pp. 289–292.

[23] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process. 28 (4) (1980) 357–366.

[24] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, J. Acoust. Soc. Am. 87 (4) (1990)1738– 1752.

[25] A. Sharma, M. Chandra, O. Farooq, Z.A. Abbasi, Hybrid wavelet based LPC features for Hindi speech recognition, Int. J. Inf. Commun. Technol. 1 (3–4) (2008) 373–381.

[26] F. Honig, G. Stemmer, C. Hacker and F. Brugnara, "Revising Perceptual Linear Prediction (PLP)", INTERSPEECH 2005, pp 2997-3000, 2005.

[27] D. Dimitriadis, A. Potamianos, On the effects of filterbank design and energy computation on robust speech recognition, IEEE Transactions on Audio, Speech and language processing, vol.19, no. 6, pp. 1504-1516, August 2011.

[28] J. Kennedy, R. Eberhart, Particle Swarm Optimization, Neural Networks, IEEE International Conference on 1995 Proceedings, Perth, WA, Australia, 1995, pp.1942–1948.

[29] S. Rainer, K. Price, Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces, J. Global Optim. 11 (4) (1997) 341–359.

[30] R.K. Aggarwal, M. Dave, Filter bank optimization for robust ASR using GA and PSO, Int. J. Speech Technol. 15 (2) (2012) 191–201.

[31] A.D. Lilla, M.A. Khan, P. Barendse, Comparison of differential evolution and genetic algorithm in the design of permanent magnet generators, in: 2013 IEEE International Conference on Industrial Technology (ICIT), IEEE, Cape Town, South Africa, 2013, pp. 266–271.

[32] H. Sun and S. Li, An optimization method for Speech enhancement based on Deep Neural Network, 3rd International Conference on Advances in Energy, Environment and Chemical Processing, Changsha, China, 2017.

[33] K. A. Qazi, T. Nawaz, Z. Mehmood, M. Rashid, & A. H. Hafiz, A hybrid technique for speech segregation and classification using a sophisticated deep neural network, *PLOS ONE,* (2018) ,1-15.

[34]. A. B. Nassif, I. Shanin, I. Attili, M. Azzeh, & K. Shaalan, Speech Recognition Using Deep Neural Networks: A Systematic Review", *IEEE Access,* 7, 2019,19143-19165.

[35] Guglani, J., & Mishra, A. N. (2018). Continuous Punjabi speech recognition model based on Kaldi ASR toolkit. International Journal of Speech Technology,21(2),211–216. https://doi.org/10.1007/s10772-018-9497-6

[36] Kadyan, V., Mantri, A., Aggarwal, R. K., & others. (2019). A comparative study of deep neural network based Punjabi-ASR system. International Journal of Speech Technology, 22(1), 111–119. https://doi.org/10.1007/s10772-018-09577-3

[37] Taniya, Bhardwaj, V., & Kadyan, V. (2020). Deep neural network trained Punjabi children speech recognition system using Kaldi toolkit. In 2020 5th International Conference on Computing, Communication and Automation (ICCCA) (pp. 374–378). IEEE. https://doi.org/10.1109/ICCCA49541.2020.9250780

[38] Kumar, Y., Singh, N., & Kumar, M. (2021). AutoSSR: An efficient approach for automatic spontaneous speech recognition model for the Punjabi language. Soft Computing, 25(3), 1617–1630. https://doi.org/10.1007/s00500-020-05248-1

[39] Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., Rehman, A. U., Shafiq, M., & Hamam, H. (2022). Automatic speech recognition (ASR) systems for children: A systematic literature review. Applied Sciences, 12(9), 4419. https://doi.org/10.3390/app12094419

[40] Chen, L. (2023). Special issue on automatic speech recognition. Applied Sciences, 13(9), 5389. https://doi.org/10.3390/app13095389

[41] Hu, Y., Chen, C., Yang, C. C. H., Li, R., Zhang, C., Chen, P. Y., & Chng, E. S. (2024). Large language models are efficient learners of noise-robust speech recognition. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2401.10446

[42] D. Lilla, M.A. Khan, P. Barendse, Comparison of differential evolution and genetic algorithm in the design of permanent magnet generators, in: 2013 IEEE International Conference on Industrial Technology (ICIT), IEEE, Cape Town, South Africa, 2013, pp. 266–271.

[43] S.P. Lim, H. Haron, Performance comparison of genetic algorithm, differential evolution and particle swarm optimization towards benchmark functions, in: 2013 IEEE Conference on Open Systems (ICOS), IEEE, Kuching, Malaysia, 2013, pp. 41–46.

[44] Stevens, S., Volkmann, J., and Newman, E., "A Scale for the Measurement of the Psychological Magnitude Pitch." Journal of the Acoustical Society of America 8: 185–190, 1937.

[45] H.P. Combrinck, E.C. Botha, On the Mel-scaled Cepstrum, Department of Electrical and Electronic Engineering, University of Pretoria, 1996.

[46] D. Fohr, O. Mella and I. Illina, "New Paradigm in speech recognition: Deep Neural Networks", IEEE International Conference on Information Systems and Economic Intelligence, April 2017, Marrakech, Morocco.

[47] S. Tilton, "e Tools Using Audacity in class room", https://www.researchgate.net/publication/299090266.