International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182, Print ISSN: 2349-5979, Impact Factor: RJIF 5.72
www.allsubjectjournal.com
Volume 4; Issue 6; June 2017; Page No. 33-38

Spectral feature extraction techniques for speech recognition

[1] Dinesh Sheoran, [2] Pardeep Sangwan, [3] Manoj Khanna

[1, 2] Maharaja Surajmal Institute of Technology, New Delhi, India
[2] BCAS, Delhi University, New Delhi, India

Abstract
In the present era, speech processing technology has made use in various applications namely "Speech enhancement", "Speech compression", "Speech recognition" and "Speech synthesis" etc. In the current work, issues related to speech recognition system and the role of feature extraction has been studied in detail. This study includes speech feature extraction using Linear Predictive Coefficients, Cepstral analysis for extensive digits database. The results presented prove the superiority of MFCC feature extraction technique.

Keywords: automatic speech recognition (ASR), mel-frequency cepstral coefficients (MFCC), linear predictive coding (LPC)

Introduction
The speech is a source of oral communication among humans for expressing thoughts and emotions. Speech is a complex signal contains rich information. Thus to understand this complex information with help of computer is known as speech recognition. The computer algorithm identifies the words spoken by any person in an ASR system, thus, enabling the communication between humans and machines in a naturally spoken language. Various important applications requiring human-machine interaction can be supported by ASR systems [1]. In today's era we have automatic call processing units in many fields where we have interaction of humans and machines via speech, so speech recognition systems find widespread applications [2, 3]. Generally, ASR systems requires to:

- Extraction of characteristic features from speech signal of known speakers.
- Creation of known speaker's feature model.
- Comparison of features of speech uttered by unknown speaker with feature models known speakers.
- Making decision after identification of unknown speaker uttering test speech signal is done.

The efficiency of ASR systems depend on factors like languages, databases used, no. of speakers, classification techniques etc. This paper presents a speech recognition system utilizing spectral features. The paper presents a brief description of spectral feature extraction techniques in section-2. Next section provides various classification methods of speech features and their comparison. In section-4, result and discussions are given and finally the paper is concluded with future scope.

2. Spectral feature extraction techniques
The general architecture of automatic speech recognition system to recognize speech signal involves different stages and is shown in figure 1. In the first stage, sound signal captured with the help of preprocessed to make them noise-free, compatible and suitable for extracting features. Four different processes are performed on signals: "End point detection", "pre-emphasis", "frame blocking" and "windowing".

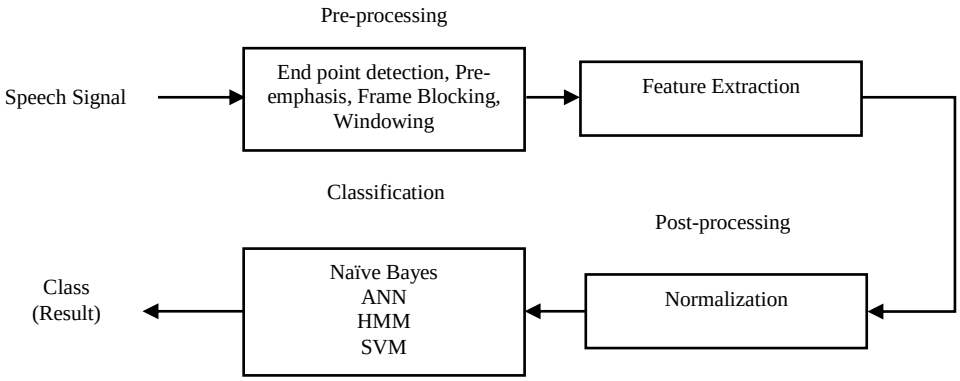


Fig 1: General block diagram for Speech recognition process

At second stage of ASR system, features are extracted from pre-processed signals. Descriptive features must be extracted from enhanced speech signals for better classification of speech signals. This step of "Feature-extraction" is required to discard huge amount of unnecessary data contained in raw speech signal. This undesired data, due to its high-

dimensionality, could make classification process unfeasible and also, results in high word-error rate. Hence, feature-extraction algorithm is required to derive feature vectors having lower-dimensionality [4, 5]. Until now, many features extraction algorithms have been developed highlighting various features of speech signals which can be broadly categorized in acoustic and linguistic features. Non-verbal vocal outbursts like sighs and laughter can be classified mainly utilizing acoustic feature whereas linguistic features are more relevant in speech recognition systems which try for transcribing linguistic messages [6]. Some of these features are "Intensity", "Linear Predictive Coding (LPC)" [7], "Perceptional Linear Predictive Coefficients (PLP)" [8], "Mel-Frequency Cepstral Coefficients (MFCC)", "Linear Prediction Cepstral Coefficients (LPCC)", "Wavelet Based Features" and "Non-Negative Matrix Factorization features" [9]. In this paper, two feature extraction algorithms namely a) LPC and b) MFCC are discussed.

a) Linear Predictive Coding (LPC)

LPC is the most commonly used technique for feature extraction from speech signals that derives fundamental parameters of speech. To get speech parameters estimated precisely, human tract structure is explored. In this methodology, a linear combination of past speech samples is utilized for approximating the given speech sample. The predictor co-efficient can be obtained by minimizing sum of squared difference of given speech sample and estimated value. These predicted co-efficient form the building block of linear predictive coding of speech [10]. The steps for feature extraction using linear predictive coding are shown in figure 2 [11].

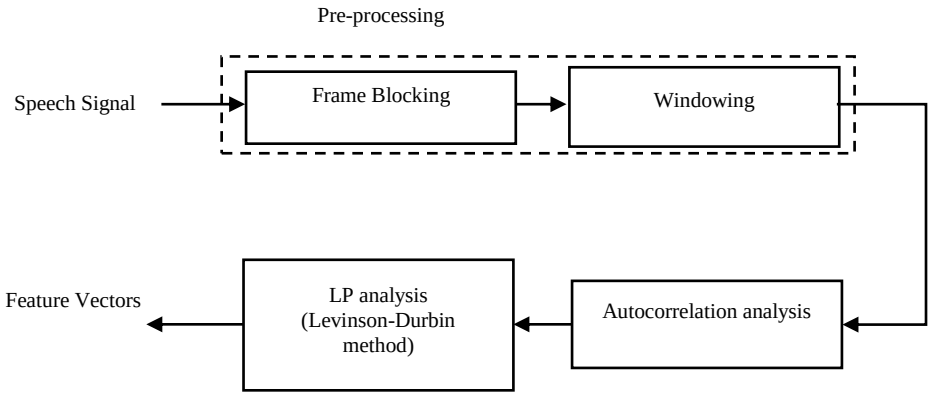


Fig 2: Feature vectors extraction using linear predictive coding

b) Mel-Frequency Cepstral Co-efficient

In speech recognition, various researchers have used MFCC for feature extraction. This can be an attempt to imitate human ear in which different frequencies across the audio spectrum are resolved non-linearly. Thus, Mel-filters are meant for distorting the frequencies so that these frequencies obey spatial relationship with characteristics of human ears. Therefore, Mel-scale means using a logarithmic-scale for frequencies higher than 1 kHz and linear-scale for frequencies lower than 1 KHz [12]. This method relies on short-term analysis in which feature vectors are computed separately for every frame. To minimize signal discontinuities, hamming window is multiplied with the input speech signal and then fast Fourier transform is applied to obtain Mel-filter bank. Then Mel-frequency warping of output is done to get number of co-efficient. Finally, cepstral co-efficient are obtained by calculating using IDFT. The numbers of coefficients are then obtained after warping [13]. This process transforms log-domain co-efficient into frequency-domain co-efficient. The steps for MFCC are shown in figure 3. MFCCs can be determined using the expression given below:

$$Mel\,(g) = 2595 \times log_{10}(1 + {}^{g}/_{700}) \quad (1)$$

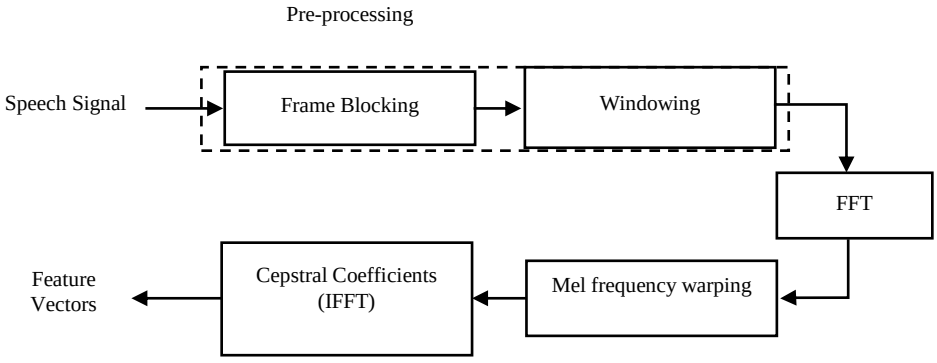


Fig 3: Feature vectors extraction using MFCC

These two presented feature extraction methods are compared in Table 1.

Table 1: Comparison of Feature Extraction methods

Feature extraction methods	
Linear predictive coding	Pros: ➢ Source to vocal tract separation possible. ➢ Lower dimensional feature vector presents spectral-envelope. ➢ Easy implementation. ➢ Precise mathematical representation. Cons: ➢ High association between feature vectors. ➢ Linear-scale is insufficient for representing speech.
Mel frequency cepstral coefficients	Pros: ➢ Not based on linear attributes; thus suitable for human auditory perception. ➢ Low association between coefficients. ➢ Necessary phonetics attributes can be gathered. Cons: ➢ Low prone to noise. ➢ Represents power spectrum while ignoring phase spectrum; therefore provides limited representation.

3. Classification from speech features

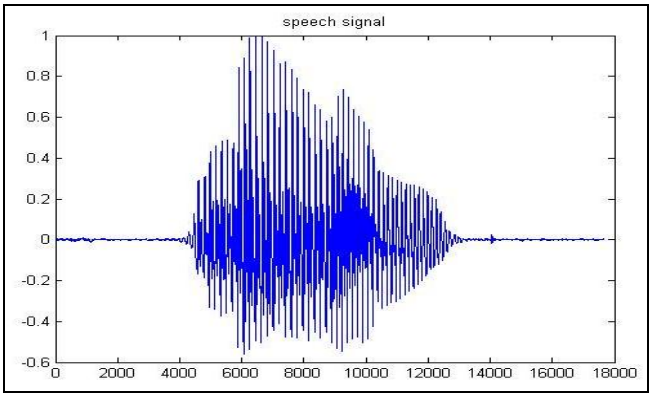
Table 2: Comparison of classification methods

Classification methods	
Naïve Based Classifier	Pros: ➢ Supervised classification. ➢ Features in one class thought to be independent of others. ➢ Ease of simple implementation and interpretation. Cons: ➢ Strong feature independence assumptions. ➢ Overfitting.
Artificial Neural Network	Pros: ➢ Capability of self-organization and self-learning. ➢ Adjustability to different environment. ➢ Suitable for pattern recognition. Cons: ➢ Requires extensive training.
Hidden Markov Models	Pros: ➢ Model time distribution of sound signal. ➢ Easy development. ➢ Support vector length input. ➢ Able to model both discrete and continuous signals. Cons: ➢ Assuming the probability of existing in specific state and independent of its previous state.
Support Vector Machine	Pros: ➢ Does not have problems like local minima and over training. ➢ Able to deal with high dimensional input vectors. Cons: ➢ Does not support variable length input. ➢ Number of classes increases computational cost. ➢ Not capable to deal with large databases.

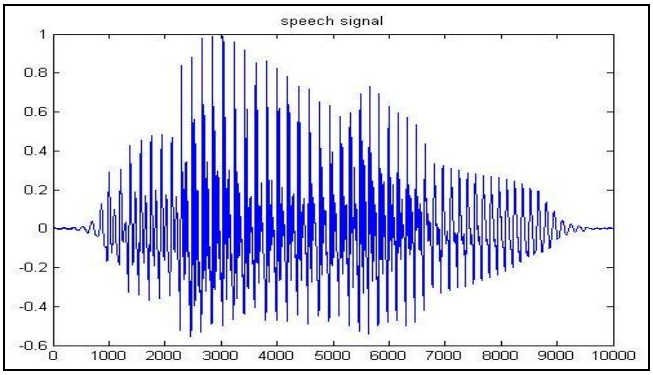
The process of partitioning of feature-space in different regions and assigning a region for every input class is called speech classification [14]. The classification is also a valuable step in ASR process as classification of speech data into different classes is done in this step which depends on supervised learning [15]. Different classification techniques are used namely ANN, Naive Bayes, SVM and HMM for recognizing speech feature sets into appropriate classes efficiently. Table 2 is presenting brief comparison of these classification methods.

4. Result and discussions

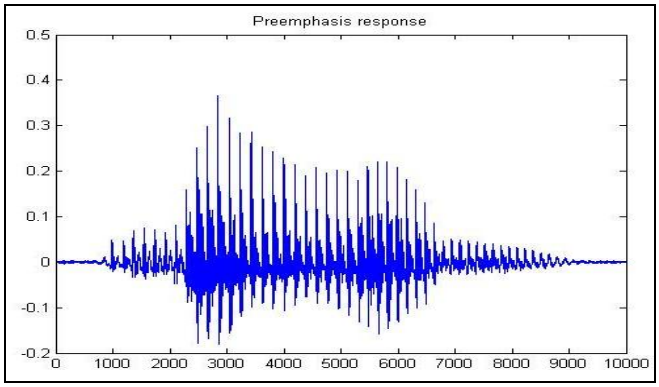
The presented method incorporates the use of preprocessing before extracting feature from speech signals. The preprocessing results of end point detection and of pre-emphasis filtering are shown in Fig 4 for digit 9 (from database).



a. Original Signal of Digit 9



b. Signal after endpoint detection



c. Signal after pre-emphasis filtering

Fig 4: Preprocessing of Speech Signal

Results Analysis for Digits Database

The database contains 2000 samples made up of utterances by two hundred speakers uttering ten digits. Total 16 experiments were conducted. Feature vectors obtained from preprocessing and feature extraction stage are classified. Table 3 and Table 4 give the results of LPC and MFCC features with four classifiers respectively. Table 5 presents comparison of recognition rate of both the features.

Table 3: Classification results for LPC of digits database (0-9)

No. of Speakers	Total Samples	Naïve Bayes		ANN		HMM		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	100	94	94.00	95	95.00	92	92.00	95	95.00
25	250	231	92.50	235	94.00	229	91.60	233	93.20
50	500	458	91.60	465	93.00	455	91.00	463	92.60
75	750	681	90.80	695	92.67	672	89.60	684	91.20
100	1000	895	89.50	918	91.80	886	88.60	909	90.90
150	1500	1320	88.00	1362	90.80	1313	87.50	1332	88.80
200	2000	1735	86.75	1774	88.70	1716	85.80	1756	87.80

Table 4: Classification results for MFCC of digits database (0-9)

No. of Speakers	Total Samples	Naïve Bayes		ANN		HMM		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	100	95	95.00	97	97.00	94	94.00	96	96.00
25	250	234	93.60	239	95.60	233	93.20	238	95.20
50	500	461	92.20	471	94.20	460	92.00	471	94.20
75	750	689	91.86	700	92.67	676	90.13	698	93.06
100	1000	906	90.60	927	92.70	892	89.20	913	91.30
150	1500	1340	89.33	1360	90.66	1326	88.40	1356	90.40
200	2000	1749	87.45	1790	89.50	1742	87.10	1778	88.90

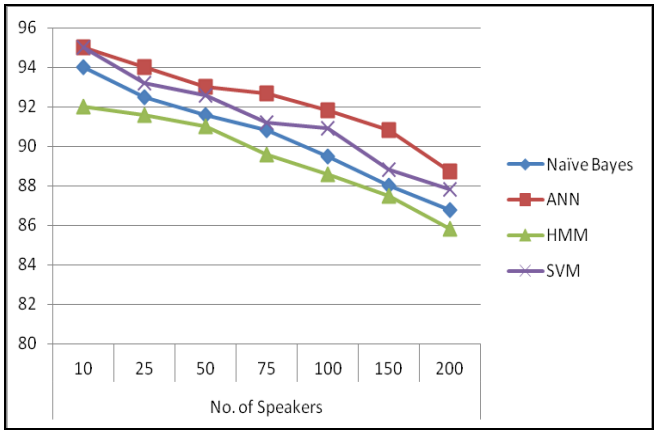


Fig 5: Accuracy on digits database with LPC

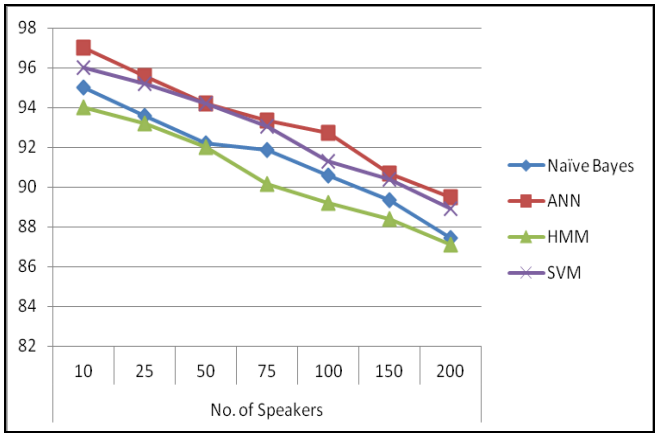


Fig 6: Accuracy on digits database with MFCC

Figure 5 and Figure 6 presents the comparison between four classifiers. It can be observed clearly that ANN is giving the best results.

Table 5: Comparison of digits database (2000 digits)

F.E. Method	No. of Features	Naïve Bayes		ANN		HMM		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
LPC	10	1735	86.75	1774	88.70	1720	86.20	1751	87.55
MFCC	12	1752	87.60	1760	89.80	1742	87.10	1773	88.65

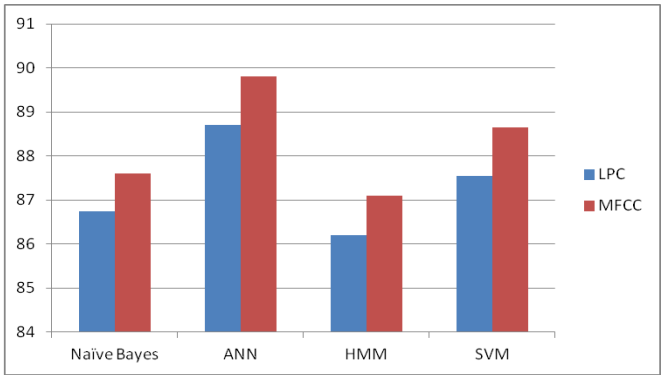


Fig 7: Comparison of speech recognition systems on Digits database

5. Conclusion and Future Directions

The paper presents the feature extraction and classification techniques for speech recognition. A comparison of several ASR techniques is made with their respective advantages and disadvantages. Based on experimental results, it is observed that MFCC performs better than LPC. Moreover, the results show that ANN has given highest accuracy rate among the four chosen classifiers. So, the best results are obtained by using MFCC for feature extraction followed by ANN as classifier. One extension of the current work can be enhancing the database of speech samples for higher accuracy rate. Also, other classifiers such as Genetic algorithm, Fuzzy logic etc. can also be utilized for analyzing the performance of the speech recognition system.

6. References

1. Anusuya MA, Katti SK. Speech Recognition by Machine: A Review, (IJCSIS) International Journal of Computer Science and Information Security. 2009; 6(3).
2. Soon Suck Jarng. HMM Voice Recognition Algorithm Coding, International Conference on Information Science and Applications (ICISA). 2011, 1-7. ISBN-978-1-4244-9222-0, IEEE.
3. Peerapol Khunarsal, *et al.* Singing Voice Recognition based on Matching of Spectrogram Pattern, Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA. 2009, 1595-1599, ISBN: 978-1-4244-3548-7, IEEE
4. Bourlard H, Hermansky H, Morgan N. towards increasing speech recognition error rates, Speech Communications. 1995; 18:205-231.
5. Shaughnesssy DO. Invited paper: Automatic speech recognition: History, methods and challenges, Pattern Recognition. 2008; 41(10):2965-2979.
6. Schuller B. Voice and speech analysis in search of states and traits, in Computer Analysis of Human Behaviour, A. A. Salah, T. Gevers (eds.), Berlin, New York, Tokyo: Springer. 2011; 9:227-253.
7. Schukat E. Talamazzini, Automatische Spracherkennung, Braunschweig, Wiesbaden: Vieweg Verlag, 1995.
8. Hermansky H, Hanson BA, Wakita H. Perceptually based linear predictive analysis of speech, Acoustics, Speech, and Signal Processing. 1985; 10:509-512.
9. Rabiner L, Juang BH. Fundamentals of speech recognition, Englewood Cliffs: Prentice-Hall International, 1993.
10. Uma Maheswari, Kabilan AP, enkatesh RV. A Hybrid model of Neural Network Approach for Speaker independent Word Recognition, International Journal of Computer Theory and Engineering, Vol.2, No.6, December. 2010, 1793-8201.
11. Vimala C, Radha V. A review on speech recognition challenges and approaches', World Computer. Sci. Inf. Technol. 2012; 2(1):1-7.
12. Sivaram GSVS, Hermansky H. Multilayer perceptron with sparse hidden outputs for phoneme recognition'.

2011 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), Prague. 2011, 5336-5339.
13. Corneliu Octavian DUMITRU, Inge GAVAT. A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language, 48th International Symposium ELMAR. Zadar, Croatia, 2006.
14. Sankar K, Pal, Pabitra Mitra. Pattern recognition algorithms for Data Mining, 1st ed. Boca Raton London, New York: Chapman and Hall/CRC, 2004.
15. Brian D. Ripley, Pattern Recognition and Neural Networks, 1st ed. Cambridge, New York: Cambridge University Press, 2008.
16. Gupta K, Gupta D. An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system, 2016 6th IEEE International Conference - Cloud System and Big Data Engineering (Confluence), Noida. 2016, 493-497.