*International Journal of Engineering and Technology Innovation, vol. 7, no. 2, 2017, pp. 78 - 88*

# Analysis of Feature Extraction Methods for Speaker Dependent Speech Recognition

Gurpreet Kaur[1, 2,*], Mohit Srivastava[3], Amod Kumar[4]

[1] I.K Gujral Punjab Technical University, Kapurthala, India.

[2] University Institute of Engineering & Technology, Panjab University, Chandigarh, India.

[3] Chandigarh Engineering College, Landran, Mohali, India.

[4] Central Scientific Instruments Organisation, Chandigarh, India.

Received 26 October 2016; received in revised form 17 December 2016; accepted 27 December 2016

## Abstract

Speech recognition is about what is being said, irrespective of who is saying. Speech recognition is a growing field. Major progress is taking place on the technology of automatic speech recognition (ASR). Still, there are lots of barriers in this field in terms of recognition rate, background noise, speaker variability, speaking rate, accent etc. Speech recognition rate mainly depends on the selection of features and feature extraction methods. This paper outlines the feature extraction techniques for speaker dependent speech recognition for isolated words. A brief survey of different feature extraction techniques like Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding Coefficients (LPCC), Perceptual Linear Prediction (PLP), Relative Spectra Perceptual linear Predictive (RASTA-PLP) analysis are presented and evaluation is done. Speech recognition has various applications from daily use to commercial use. We have made a speaker dependent system and this system can be useful in many areas like controlling a patient vehicle using simple commands.

**Keywords:** speech recognition, feature extraction, mel-frequency cepstral coefficients, linear predictive coding coefficients, perceptual linear production, RASTA-PLP, isolated words

## 1. Introduction

Speech processing is the study of speech and its processing methods which include automatic recognition of speech signal and extraction of characteristics of the speaker [1-2]. Another purpose of speech processing is a representation of it for reproduction and transmission. There are different application areas of speech processing like human-computer interfaces, telecommunication, assistive technologies and security [3]. Important fields of speech processing are synthesis, recognition and coding of speech signals [4]. Recognition itself is a wide topic consisting of three areas i.e. speech recognition, speaker recognition and language recognition. As the name tells, speech recognition is to recognize the spoken words, language recognition is to recognize the language being spoken and speaker recognition is the identification/verification of the speaker. Speaker recognition field is divided into two categories i.e. text dependent and text independent. In text dependent speaker recognition mode, the speaker is to speak same words which are known to the system but in text independent mode, speaker can speak any set of words. Though speech recognition and speaker recognition are different fields, but the feature extraction methods in both the fields are overlapped [5]. The combined area of speaker recognition and speech recognition encompasses speaker dependent speech recognition [6]. In this paper, our main concern is for speaker dependent speech recognition. Speech

* Corresponding author. E-mail address: regs4gurpreet@yahoo.co.in

recognition system is divided into various types depending on types of speech mode, types of speaker mode, types of vocabulary size etc. Speech mode is divided into three parts: isolated, connected and continuous. In isolated speech recognition, speaker has to speak into the system word-by-word; connected speech recognition occurs when speaker speaks a number of words without stopping and continuous speech recognition is like human. Types of speaker mode can be speaker dependent recognition or speaker independent recognition [7]. Speaker dependent recognition systems are designed for one speaker who has trained the system, for example voice dialing and Speaker independent recognition systems are designed for all users without prior training, for example, Siri in i-phone. The development of speaker independent system is difficult, they are expensive and the recognition rate is low as compared to speaker dependent system [8].

Types of vocabulary can be small, medium and large [9]. Small vocabulary deals with tens of words, medium vocabulary deals with hundreds of words, large vocabulary deals with thousands of words.

When the air from lungs passes through vocal cords, throat, mouth, and nasal tract, the speech is produced. Different sound patterns are created because of different positions of lips, tongue and palate. Speech signal consists of different attributes like loudness, voiced/unvoiced, pitch, fundamental frequency, spectral envelope, formants etc. Voiced speech is produced when the vocal cords vibrate during the pronunciation, for example, a, e, i, o, u, whereas unvoiced speech is produced when there is no vibration of vocal cords, for example, /t/, /p/. Speech features can be classified as physical, perceptual and signal features. Physical features are dependent upon physical properties of speech signal like power, fundamental frequency etc. The amplitude of speech signal specifies power in it. More power means louder signal. Silence zones in speech signal can be discovered using power measurement. Another physical feature is fundamental frequency. The range of fundamental frequency of female is 165 Hz to 255 Hz and that of a male is from 85 Hz to 155 Hz, so male and female voices can be separated using this attribute. Pitch and prosody are perceptual features. Pitch analysis can tell about the emotion of the speaker. Signal features are characteristics of a speech signal. They are extracted by mathematical analysis of the signal. Speech can be analyzed both in time and frequency domain. Short time speech features are energy, amplitude, amplitude difference, zero crossing (ZC) rate and autocorrelation. Analyzing speech in the time domain often requires simple calculation and interpretation. The analysis in frequency domain consists of short time Fourier Transform, spectrogram of speech signals, Filter bank implementation, cepstral analysis etc. Spoken words are not exactly equal to what we write because of background noise, body language, channel variability, speaker variability, speaking style, dialects etc. All mentioned factors are responsible for accuracy of speech recognition [10].

## 2. Literature Survey

There is tremendous growth in speech recognition field from daily use to commercial applications like phone programs used by information system of banks, airlines, Siri in mobile phones, dragon dictate to healthcare sectors etc. In 1940s, AT & T Bell laboratories developed a device that could recognize speech. In 1960s researchers started the work on larger speech recognition system. The speech recognition technology is still being refined in terms of recognition rate, noise robustness, larger vocabulary, continuous speech etc. [11]. Table 1 shows different techniques of speaker dependent speech recognition and its recognition rate [12-19].

Speaker dependent methods usually involve training a system to recognize each of the vocabulary words uttered single or multiple times by a specific set of speakers. Various features have been used singly or in combination with others to model the speech signals, ranging from Linear Predictive Coding (LPC), Dynamic Time Wrapping (DTW), Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing with Peak Amplitude and Relative Spectra Filtering (RASTA) [20-23]. In last few years, the

MFCC and LPCC features have become reasonable leaders in the recognition systems. A 98.5 % rate was achieved recognizing isolated words in a small vocabulary using MFCC. The highest 95.47 % recognition rate was achieved using LPC and artificial neural network in a small vocabulary system. Recognition efficiency decreases in noisy environment. The recognition rate is more for the English language as compared to other languages by using conventional feature extraction methods.

Table 1 Different techniques of speaker dependent speech recognition

| Sr. No. | Title | Authors | Technique used | Language | Accuracy |
|---|---|---|---|---|---|
| 1. | Speech and Speaker Recognition System Using Artificial Neural Networks & Hidden Markov Model | Niladri Dey *et al.*, 2012 | Artificial Neural Network (ANN)and Hidden Markov Model (HMM) | English words | 90% |
| 2. | Isolated Word Speech Recognition Using Dynamic Time Warping | Rajesh Makhijani, *et al.*, 2013 | Mel-Frequency Cepstral Coeff. (MFCC) | English Words | 90.1% |
| 3. | A novel isolated speech recognition method based on Neural Network | Fu Guojiang, 2011 | Linear predictive Cepstral Coeff. MLP/RBF | English Words | 95.47% |
| 4. | DWT Feature performance analysis for automatic speech recognition of Urdu | Hazaral ali *et al.*, 2014 | Discrete Wavelet Transform (DWT) MFCC | Urdu Word | DWT-40% MFCC-70.67% |
| 5. | Speech recognition using Artificial Neural Network | Tiago P., Nascimento *et al.*, 2011 | Artificial Neural Networks/ Hidden Markov Method | English Word | HMM-96% ANN-99% |
| 6. | Analysis Of Different Feature Extraction Techniques for speaker recognition system | Yoghesh Dawande *et al.*, 2015 | MFCC, LPC and PLP in noisy environment | English words | MFCC 96.5% LPC 65.8% PLP 78.5% |
| 7. | Real Time Speaker Recognition System using MFCC and Vector Quantization Technique | Roma Bharti *et al.*, 2015 | Mel-Frequency Cepstral Coeff. (MFCC) and VQ | English Words | MFCC 91% at 20dB SNR. |
| 8. | Robust Speech Recognition System Using Conventional and Hybrid Features of MFCC, LPCC, PLP, RASTA-PLP and Hidden Markov Model Classifier in Noisy Conditions | Veton Z. Kepuska *et al.*, 2015 | MFCC,LPCC,PLP, RASTA-PLP | English Words | MFCC+ Δ + ΔΔ 97.92%, LPCC+ Δ + ΔΔ 97.63%, PLP+ Δ + ΔΔ 98.35%. RASTA-PLP+ Δ + ΔΔ 95.93% at 20dB SNR |
| 9. | DWT and LPC based feature extraction method for isolated word Recognition | Navnath S Nehe *et al.*, 2012 | LPCC and MFCC | Marathi Digits | LPCC-78.9%, MFCC-84.5% |

## 3. Speech Recognition System

The speech recognition system consists of digital acquisition of the speech signal and then pre- processing of speech signal is done. Preprocessing consists of silence removal, pre-emphasis, framing and windowing of the speech signal. After pre-processing, features are extracted from the speech signal. It is a process of retaining useful information in a speech signal while discarding the unwanted and redundant information. There are various techniques for feature extraction such as Mel Scale frequency Cepstral Coefficient (MFCC), Perceptual linear predictive (PLP), Linear Predictive Coding (LPC), RASTA - PLP, etc. After extracting features, matching/modelling is done. Basically, there are three approaches for this i.e. acoustic phonetic approach, pattern recognition approach and artificial intelligence approach [24-28]. Acoustic phonetic approach deals with the phonetic units present in speech and it is dependent upon acoustic properties of these units. Acoustic properties vary according to the co-articulation effect and speaker. The technique is implemented by carrying out the speech spectral analysis. After feature extraction, the acoustic properties of different phonetic units are defined. Then, segmentation and labelling is done. Nowadays, this approach is not used. Pattern recognition approach deals with pattern matching. It consists of four steps - feature

extraction, training, classification and decision. Feature extraction gives the unique features for every word. Then, reference patterns are made using the sound samples. Classification is done to match the reference pattern and test pattern and accordingly decision is made. We have used pattern recognition approach in our system. The artificial intelligence approach is knowledge-based approach. Artificial neural network comes under this approach. Fig. 1 shows the speech recognition system[29].

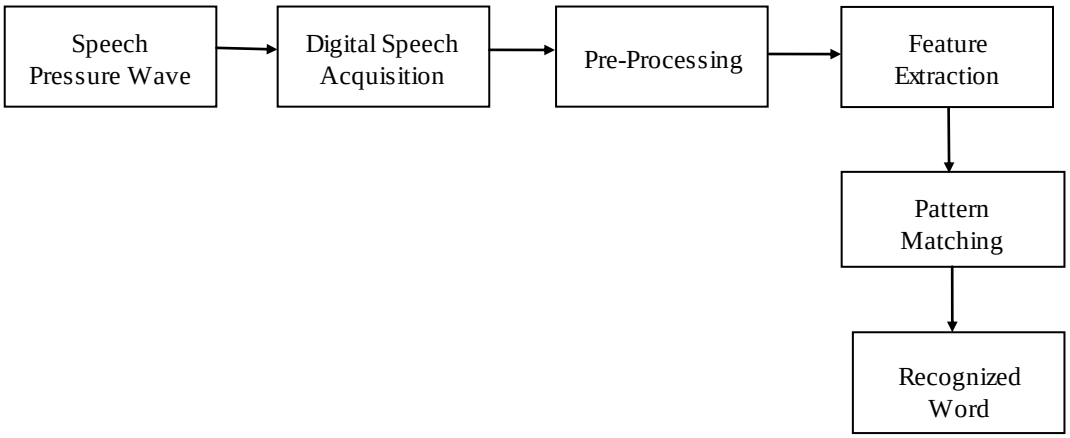


Fig. 1 Speech recognition system

Fig. 1 shows the methodology used in speaker dependent speech recognition system. Speech is acquired by sound recorder with the help of headphone. Then preprocessing of each word is done to enhance the properties of the signal. For the classification, pattern matching is used. The pattern matching stage is for modeling. In this technique, the model consists of a template which is a feature vector from a fixed phrase. In this Dynamic Time Warping Algorithm is used to compute the distance between the input speech and stored reference patterns. The word is recognized based on minimum distance.

## 4. Feature Extraction Analysis

Feature extraction techniques like Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), Perceptual linear predictive (PLP) analysis and Relative Spectra Perceptual linear Predictive (RASTA- PLP) analysis are implemented for isolated word recognition. The speech signal of duration 30ms is taken in .wav file. Then preprocessing of signal is done. Preprocessing consists of silence removal, pre-emphasis, framing and windowing of the speech signal. The steps are as below.

### *4.1. Silence Removal*

The silence removal in speech signal is done because silence contains no information in it. It makes the original signal larger and hence more space and more time will be consumed to process it. We can remove the silence signal by using signal energy estimation. A threshold is set and below this value, the signal is truncated.

### *4.2. Pre-emphasis*

In a speech signal, lower frequencies contain more energy as compared to higher frequencies. Thus, in order to boost the energy level for higher frequencies, a pre-emphasis filter is used. It is done by a first order high pass filter as

emphasis signal(n)= signal(n)-a* signal(n-1)

In z domain

$H(z) = (1-a*z^{-1})$

The value of a has been taken between 0.9 to 1.0 in the literature, we have taken a=0.93

### 4.3. Framing and windowing

Speech signal is a non-stationary signal. However, for short period of time, it may be regarded as stationary signal. Therefore speech signal is taken in frames. This is done with the help of window. The features are extracted for every N ms called frame rate for the window size M ms. M should be bigger than N. Therefore, two consecutive frames have overlapping area. We have used hamming window in this experiment.

### 4.4. Feature extraction

Feature Extraction is the process of retaining useful information in a speech signal while discarding the unwanted information. It is nothing but parameterization of speech signal. There are various techniques for feature extraction as explained below:

#### 4.4.1. Mel frequency Cepstral Coefficient (MFCC)

Mel frequency Cepstral analysis is a method which models the vocal tract system [30]. The vocal tract articulation equivalent filter is shown by the equation:

$$S(\omega) = G(\omega)H(\omega) \tag{1}$$

The equivalent logarithm of $S(\omega)$ is

$$Log|S(\omega)| = Log|G(\omega)| + Log|H(\omega)| \tag{2}$$

The cepstrum C (τ), or cepstral coefficients, is the inverse Fourier transform of the $Log|S(\omega)|$.

$$C(\tau) = F^{-1}Log|S(\omega)| = F^{-1}Log|G(\omega)| + F^{-1}Log|H(\omega)| \tag{3}$$

This equation gives the fundamental frequency and spectral envelope.

#### 4.4.2. Linear Predictive Cepstral Coefficients (LPCC)

This technique is a powerful tool which can also be used to conduct pitch and formant detection on speeches. The term linear predictive is based on the fact that the present sample value S[n] can be linearly predicted using the previous sample values S[n-k] [31].

$$S[n] = \sum_{k=1}^{p} \alpha_k S[n-K] \tag{4}$$

This linear prediction will introduce errors into the sequence of speech samples. This error is known as the residual error e [n]. It is represented by the following equation:

$$e[n] = s[n] - \sum_{k=1}^{p} \alpha_k S[n-K] \tag{5}$$

This equation is then transformed into z-domain, and is then expressed by

$$E(z) = (-\sum_{k=1}^{p} \alpha_k z^{-k?}) S(z) \tag{6}$$

The steady-state system function of the articulation transfer filter is found to be

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{(1 - \sum_{k=1}^{p} a_k Z^{-k})} \tag{7}$$

where the gain G of the filter is found to be

$$G = \sqrt{(R_n(0) - \sum_{k=1}^{p} a_k R_n(k))} \tag{8}$$

From Eqs. (7) and (8), it can be seen that the speech signal obeys the LPC model equation exactly. Thus, A(z), also called the inverse filter, will be identical to U(z) in the transfer filter. Based on these equations of linear predictive coding, the LPC analysis will minimize the error e[n] by adjusting the LP co-efficient $a_k$. Using covariance and the autocorrelation method, we can estimate the LP coefficients.

### *4.3.3. Perceptual linear predictive (PLP) analysis*

In Perceptual linear predictive (PLP) analysis of speech, auditory spectrum is represented by an autoregressive all pole model. The real and imaginary parts of FFT of speech signal are squared to get power spectrum.

$$P(\omega) = Re\left[S(\omega)\right]^2 + Im\,[S(\omega)]^2 \tag{9}$$

Then power spectrum is warped into Bark frequency (Ω) represented as

$$\Psi(\Omega) = \begin{cases} 0 & \text{for } \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & \text{for } -1.3 \le \Omega \le -0.5 \\ 1 & \text{for } -0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega-0.5)} & \text{for } 0.5 \le \Omega \le 2.5 \\ 0 & \text{for } \Omega > 2.5 \end{cases} \tag{10}$$

The convolution of $\Psi(\Omega)$ with $P(\omega)$ gives

$$\theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\ \Psi(\Omega) \tag{11}$$

This process reduces the spectral resolution. As non-equal perception of loudness at different frequencies is to be compensated, pre emphasis is done by an equal loudness curve.

$$E(\omega) = \frac{(\omega^2 + 56.8*10^6)\ \omega^4}{(\omega^2 + 6.3*10^6)^2 * (\omega^2 + 0.38*10^6)} \tag{12}$$

There is nonlinear relationship between the intensity of sound and perceived loudness:

$$\phi\ (\Omega) = (\Omega)^{0.33} \tag{13}$$

The last operation of PLP analysis $\phi(\Omega)$ is approximated by the spectrum of an all pole model using autocorrelation method.

### *4.3.4. Relative Spectra Perceptual linear Predictive (RASTA- PLP) analysis*

This technique is similar to the PLP technique. PLP technique is based upon the short term spectrum of speech and it leads to linear spectrum distortions. Therefore, RASTA -PLP technique is used [26]. Here, short term spectrum of speech is replaced by a spectral estimate in which at zero frequency, each frequency channel is band pass filtered by a filter whose transfer function is given as

$$H(z) = 0.1 * (2 + z^{-1} - z^{-3} - 2z^{-4}) / [z^{-4} * (1 - 0.98z^{-1})] \tag{14}$$

The low cutoff frequency is ignored in the output and high cutoff frequency, which determines the fastest spectral change, is preserved.

## 5. Experiment and Implementation of the System

Implementation is done on MATLAB software. It is divided into two parts. The first part is the training stage and second is the testing stage. Since it is a speaker dependent speech recognition system, so training is required for every speaker. Training means preparing a database for each speaker, then preprocessing of signal is done and features are extracted to create a speaker model. The speech is acquired by sound recorder with the help of headphone at 16 KHz frequency at room environment in mono format. In our case, database is prepared for four speakers of age 27-34, two females (F1, F2) and two males (M1, M2). The words recorded are FORWARD, BACKWARD, LEFT, RIGHT and STOP. Each word is recorded 10 times and hence fifty words are recorded for each speaker creating a database of 200 words. All the recoded files are stored as .wav file. In the testing stage, the process is same as training stage in terms of preprocessing the signal and feature extraction. The addition is pattern matching stage for modeling. In this technique, the model consists of a template which is a feature vector from a fixed phrase. Then for recognition, match score is produced by using DTW algorithm to measure the similarity between the stored template and the input speech. Graphic User Interface (GUI) is designed in MATLAB to make it easy to use.

Fig. 2 shows one of the examples of training stage of speaker dependent recognition with extracted features. Training is done by the speaker (Female 3) for the word RIGHT and MFCC feature extraction method is used. Fig. 3 shows the testing stage of speaker dependent speech recognition system with recognized word with speaker. (GKaur (F1) is speaker and the recognized word is RIGHT).

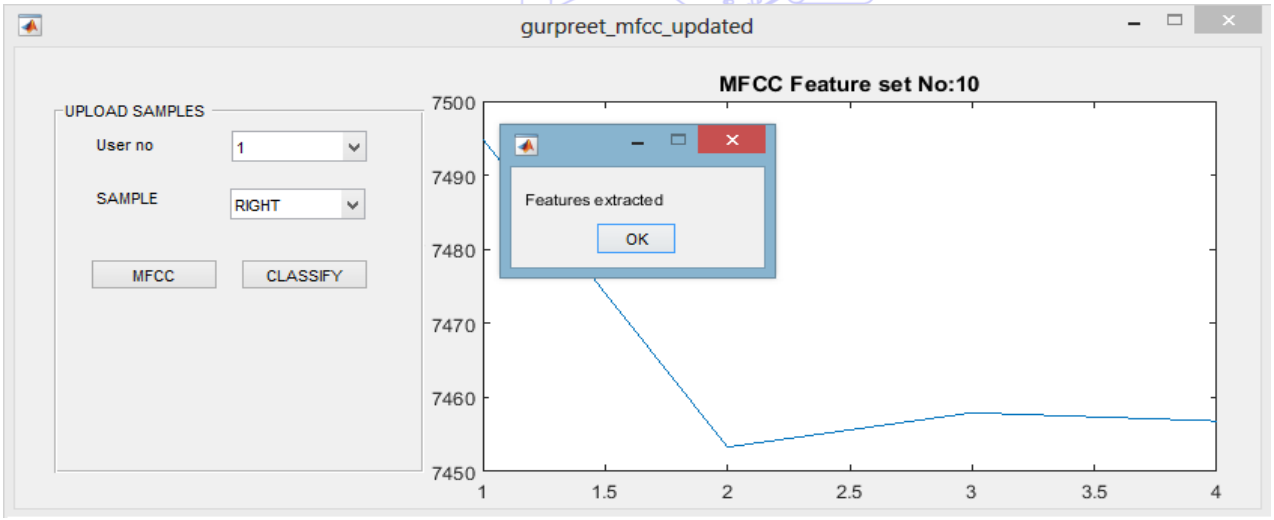


Fig. 2 Snapshot of speaker dependent speech recognition system with extracted features

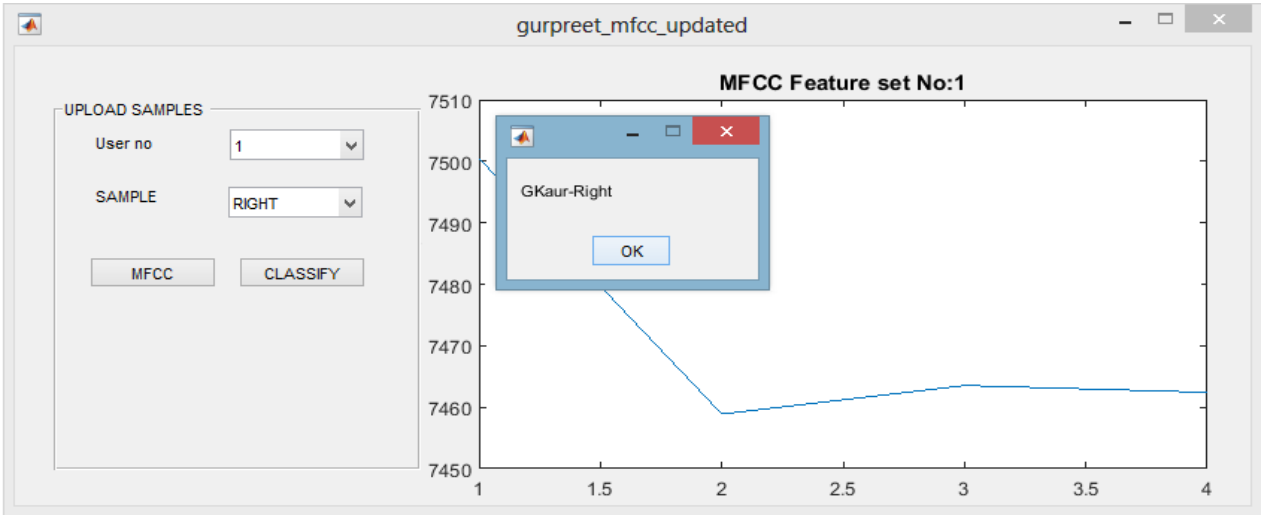


Fig. 3 Snapshot of speaker dependent speech recognition system with recognized word.

## 6. Results and Discussion

We have implemented speaker dependent speech recognition with all feature extraction methods i.e. Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding Coefficients (LPCC), Perceptual Linear Prediction (PLP) and Relative Spectra Perceptual linear Prediction (RASTA-PLP). Accuracy is calculated in clean as well as by adding White Gaussian Noise (WGN) in the speech samples as shown in Tables 2-5:

Table 2 shows that average speech recognition by using MFCC feature extraction algorithm for clean environment is 87.82% but by adding WGN to the speech signal, the rate of recognition decreases to 73.98%. In Table 3 the feature extraction method used is LPCC and average speech recognition rate for clean environment is 87.70% but by adding WGN to the speech signal, the rate of recognition decreases to 73.56%.

Table 2 Accuracy (%) in clean and with WGN using MFCC Technique for speech recognition for two males (M1, M2) and two females (F1,F2)

| | M1 | M2 | F1 | F2 |
|---|---|---|---|---|
| Words | Accuracy | Accuracy | Accuracy | Accuracy |
| Backward | 86.20 | 85.61 | 90.10 | 92.51 |
| Backward* | 70.24 | 70.04 | 70.77 | 70.59 |
| Forward | 88.95 | 88.12 | 86.11 | 88.11 |
| Forward* | 74.11 | 71.69 | 73.74 | 74.01 |
| Left | 88.30 | 85.90 | 88.22 | 88.11 |
| Left* | 73.92 | 71.01 | 71.52 | 73.74 |
| Right | 89.23 | 87.39 | 86.26 | 86.22 |
| Right* | 71.79 | 74.37 | 73.51 | 74.52 |
| Stop | 90.54 | 86.34 | 87.72 | 86.54 |
| Stop* | 70.81 | 70.41 | 70.36 | 70.16 |

* words with WGN added

Table 3 Accuracy (%) in clean and with WGN using LPCC Technique for speech recognition for two males (M1,M2) and two females (F1,F2)

| | M1 | M2 | F1 | F2 |
|---|---|---|---|---|
| Words | Accuracy | Accuracy | Accuracy | Accuracy |
| Backward | 88.92 | 86.41 | 85.33 | 86.11 |
| Backward* | 69.96 | 71.13 | 70.74 | 71.62 |
| Forward | 89.35 | 87.10 | 85.51 | 86.33 |
| Forward* | 71.25 | 69.78 | 71.90 | 70.99 |
| Left | 89.78 | 89.89 | 89.85 | 89.45 |
| Left* | 69.78 | 73.40 | 71.81 | 72.57 |
| Right | 90.45 | 87.79 | 84.81 | 84.79 |
| Right* | 71.39 | 74.23 | 72.38 | 72.05 |
| Stop | 86.48 | 89.71 | 87.56 | 88.48 |
| Stop* | 70.93 | 70.57 | 73.95 | 74.22 |

* words with WGN added

Similarly, Table 4 shows the average speech recognition rate using PLP feature extraction algorithm for clean environment as 87.46% but by adding WGN to the speech signal the rate of recognition decreases to 73.10%, and last Table 5 depicts that average speech recognition rate by using PLP- RASTA feature extraction algorithm for clean environment is 87.72% but by adding WGN to the speech signal the rate of recognition decreases to 73.58%. Thus it can be seen that MFCC technique is the best feature extraction method.

Table 4 Accuracy (%) in clean and with WGN using PLP Technique for speech recognition for two males (M1,M2) and two females (F1,F2)

| | M1 | M2 | F1 | F2 |
|---|---|---|---|---|
| Words | Accuracy | Accuracy | Accuracy | Accuracy |
| Backward | 88.75 | 86.44 | 87.16 | 86.44 |
| Backward* | 71.59 | 69.52 | 73.20 | 72.44 |
| Forward | 88.78 | 85.84 | 85.40 | 85.53 |
| Forward* | 71.21 | 71.69 | 71.59 | 72.76 |
| Left | 86.72 | 89.24 | 88.37 | 87.54 |
| Left* | 72.30 | 71.95 | 69.81 | 71.22 |
| Right | 88.16 | 90.25 | 86.12 | 88.11 |
| Right* | 72.10 | 71.91 | 69.03 | 71.52 |
| Stop | 87.11 | 88.36 | 87.94 | 87.11 |
| Stop* | 71.83 | 74.03 | 74.50 | 74.74 |

* words with WGN added

Table 5 Accuracy (%) in clean and with WGN using RASTA-PLP Technique for speech recognition for two males (M1,M2) and two females (F1,F2)

| | M1 | M2 | F1 | F2 |
|---|---|---|---|---|
| Words | Accuracy | Accuracy | Accuracy | Accuracy |
| Backward | 89.37 | 85.83 | 85.75 | 85.57 |
| Backward* | 69.92 | 73.52 | 72.48 | 72.11 |
| Forward | 86.61 | 87.38 | 89.96 | 86.11 |
| Forward* | 70.04 | 69.35 | 71.49 | 72.75 |
| Left | 88.89 | 90.23 | 85.87 | 85.49 |
| Left* | 72.85 | 72.12 | 71.73 | 71.56 |
| Right | 89.61 | 89.62 | 86.90 | 87.91 |
| Right* | 70.37 | 74.31 | 72.21 | 72.58 |
| Stop | 86.52 | 88.29 | 86.92 | 87.43 |
| Stop* | 71.46 | 73.89 | 72.47 | 72.89 |

* words with WGN noise added

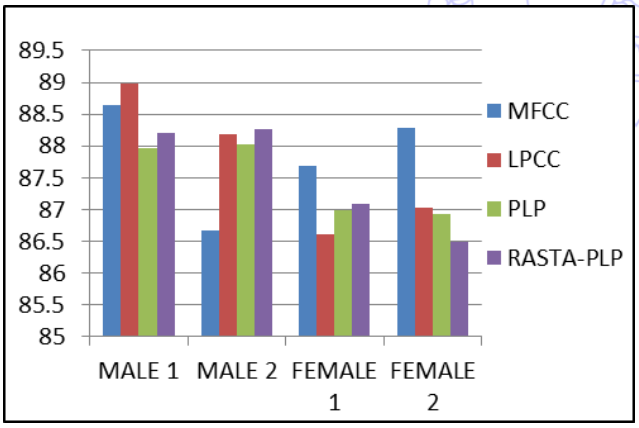


Fig. 4 Percentage accuracy in clean environment

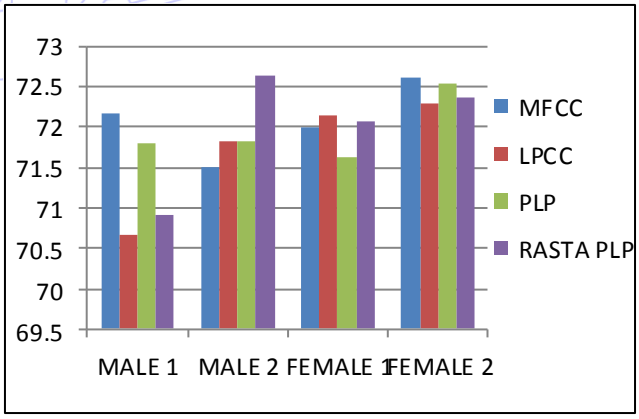


Fig. 5 Percentage accuracy by adding WGN

Fig. 4 and Fig. 5 show the results of Table 2-5 i.e. accuracy percentage in clean environment and by adding WGN in signal respectively.

The implemented speaker dependent (SD) speech recognition system can be used for text dependent speaker verification and identification. The template stored for each speaker can be used for verification or identification of the speaker based on the matching score. A threshold can be set for the decision of accepting or rejecting the claimed speaker. For speaker independent (SI) speech recognition system, templates from multiple speakers should be considered for recognition. This task is more difficult due to the wide variations of speaker characteristics. Therefore, recognition accuracy decreases three to four times than speaker dependent speech recognition. Siri speech recognition in apple i phone series is speaker independent recognition. In this, the

system is trained with multiple speakers. An experiment has done with Siri speech recognition software of apple i phone, taking a data list of 12 words. The tested words are 'forward', 'backward', 'left', 'right', 'start', 'stop', 'slow', 'fast', 'open', 'close', 'here', 'there' and comparison is done between speaker dependent and speaker independent mode. The recognition accuracy is less in case of speaker independent mode as shown in Table 6.

Table 6 Accuracy(%) of SI and SD mode

| Speaker Mode | Average Accuracy (%) |
|---|---|
| Speaker Independent (SI) | 63 |
| Speaker Dependent (SD) | 86 |

The recognition accuracy is highly dependent upon speaking style in case of speaker independent recognition. For example, many times 'forward' is recognized as 'for what' and 'backward' as 'back wood'. Recognition accuracy improves in room environment and when speaker follows some typical speaking style.

## 7. Conclusion

In this paper, the existing techniques of feature extraction like LPCC, PLP, RASTA PLP and MFCC for isolated words speech recognition are investigated, and these techniques are implemented for five words recorded by four persons in clean and noisy environment. The result shows that out of four techniques, MFCC gives the best results in clean as well as in noisy environment. Average accuracy (%) for combined speaker and speech recognition in clean environment is 87.82% and by adding WGN in signal, it is 73.98%. Using these results, applications of speech recognition can be implemented like voice operated patient vehicle. The results can be improved by adding deep neural network for classification. Comparison between speaker dependent and independent mode shows that recognition accuracy is less in case of speaker independent mode.

## References

[1] D. Shaughnessy, "Invited paper: automatic speech recognition: history, methods and challenges," Pattern Recognition, vol. 41, no. 10, pp. 2965-2979, 2008.
[2] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, et. al, "Automatic speech recognition and speech variability: a review," Speech Communication, vol. 49, no. 10-11, pp. 763-786, 2007.
[3] N. Singh, R. A. Khan and R. Shree, "Applications of speaker recognition," Procedia Engineering, vol. 38, pp. 3122-3126, 2012.
[4] S. Furui, "50 years of progress in speech and speaker recognition," ECTI Transactions on Computer and Information Technology, vol. 1, no. 2, 2005.
[5] N. Alee, P. Ehkan, R. Badlishah Ahmad, and N. Sabri, "Speaker recognition system: vulnerable and challenges," International Journal of Engineering and Technology, vol. 5, no. 4, pp. 3191-3195, 2013.
[6] M. E. Ayadi, M. S. Kamel, F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572-587, 2011.
[7] V. Fontaine and H. Bourlard, "Speaker-dependent speech recognition based on phone-like units models-application to voice dialing," Proceeding IEEE International Conference on Acoustic Speech, Signal Processing, vol. 2, pp. 1527-1530, 1997.
[8] X. Huang and K. F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," IEEE Transaction on Speech audio Processing, vol. 1, no. 2, pp. 150-157, 1993.
[9] P. Cerva, J. Silovsky, J. Zdansky, J. Nouza, and L. Seps, "Speaker-adaptive speech recognition using speaker diarization for improved transcription of large spoken archives," Speech Communication, vol. 55, no. 10, pp. 1033-1046, 2013.
[10] L. Wang, J. Wang, L. Li, T. F. Zheng, and F. K. Soong, "Improving speaker verification performance against long-term speaker variability," Speech Communication, vol. 79, no. C, pp. 14-29, 2016.
[11] O. Scharenborg, "Reaching over the gap: a review of efforts to link human and automatic speech recognition research," Speech Communication, vol. 49, no. 5, pp. 336-347, 2007.
[12] N. S. Dey, R. Mohanty, and K. L. Chugh, "Speech and speaker recognition system using artificial neural networks and hidden markov model," International Conference on Communication System and Network Technology, pp. 311-315, 2012.

[13] R. Makhijani, and R. Gupta, "Isolated word speech recognition system using dynamic time warping," International Journal of Engineering sciences and emerging Technologies, vol. 6, no. 3, pp. 352-367, 2013.
[14] F. Guojiang, "A novel isolated speech recognition method based on neural network," International Conference on Networking and Information Technology, vol. 17, pp. 264-269, 2011.
[15] H. Ali, N. Ahmad, X. Zhou, K. Iqbal, and S. M. Ali, "DWT features performance analysis for automatic speech recognition of Urdu," Springerplus, vol. 3, pp. 1-10, 2014.
[16] B. C. Kamble, "Speech recognition using artificial neural network," International Journal of Computing, Communications & Instrumentation Engineering, vol. 3, pp. 1-4, 2016.
[17] R. Bharti, "Real time speaker recognition system using MFCC and vector quantization technique," International Journal of Computer Applications, vol. 117, no. 1, pp. 25-31, 2015.
[18] V. Z. Kepuska and H. A. Elharati, "Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and hidden markov model classifier in noisy conditions," Journal of Computer and Communications, vol. 3, no. 6, pp. 1-9, June 2015.
[19] N. S. Nehe and R. S. Holambe, "DWT and LPC based feature extraction methods for isolated word recognition," EURASIP Journal of Audio, Speech, Music Processing, vol. 2012, no. 1, pp. 1-7, 2012.
[20] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition from features to supervectors," Speech Communication, vol. 52, no. 1, pp. 12-40, 2010.
[21] F. Zeng and H. Zhou, "Speaker recognition based on a novel hybrid algorithm," Procedia Engineering, vol. 61, pp. 220-226, 2013.
[22] O. Prabhakar and N. Sahu, "A survey on voice command recognition technique," International Journal of Advance Research in Computer Science and Software Engineering , vol. 3, no. 5, pp. 576-585, 2013.
[23] M. G. Sumithra, K. Thanuskodi, and A. H. J. Archana, "A new speaker recognition system with combined feature extraction techniques", Journal of Computer Science, vol. 7, no. 4, pp. 459-465, 2011.
[24] Y. Jeong, "Joint speaker and environment adaptation using tensor voice for robust speech recognition," Speech Communication, vol. 58, pp. 1-10, 2014.
[25] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol. 10, no. 1-3, pp. 19-41, 2000.
[26] L. Moreno, J. G. Dominguez, D. Martinez, O. Plchot, J. G. Rodriguez, and P. Moreno, "On the use of deep feed forward neural networks for automatic language identification," Computer Speech Language, vol. 40, pp. 46-59, 2016.
[27] R. Price, K. Iso, and K. Shinoda, "Wise teachers train better DNN acoustic models," EURASIP Journal of Audio, Speech, Music Processing, vol. 2016, no. 1, pp. 1-10, 2016.
[28] T. Alsmadi, H. A. Alissa, E. Trad, and K. A. Alsmadi, "Artificial intelligence for speech recognition based on neural networks", Journal of Signal and Information Processing, vol. 6, no. 2, pp. 66-72, 2015.
[29] S. Squartini, E. Principi, R. Rotili, and F. Piazza, "Environmental robust speech and speaker recognition through multi-channel histogram equalization," Neurocomputing, vol. 78, no. 1, pp. 111-120, 2012.
[30] K. Gupta and D. Gupta, "An analysis on LPC, RASTA and MFCC techniques in automatic speech recognition system," Conf. Cloud System and Big Data Engineering (Confluence), pp. 493-497, 2016.
[31] Z. Li and Y. Gao, "Acoustic feature extraction method for robust speaker identification," Multimedia Tools and Applications, vol. 75, no. 12, pp. 7391-7406, June 2016.