Journal of Computer and Communications, 2015, 3, 1-9
Published Online June 2015 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2015.36001

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. Këpuska, Hussien A. Elharati

Electrical & Computer Engineering Department, Florida Institute of Technology, Melbourne, FL, USA
Email: vkepuska@fit.edu, helharati2013@my.fit

Received 19 April 2015; accepted 23 May 2015; published 26 May 2015

Abstract

In recent years, the accuracy of speech recognition (SR) has been one of the most active areas of research. Despite that SR systems are working reasonably well in quiet conditions, they still suffer severe performance degradation in noisy conditions or distorted channels. It is necessary to search for more robust feature extraction methods to gain better performance in adverse conditions. This paper investigates the performance of conventional and new hybrid speech feature extraction algorithms of Mel Frequency Cepstrum Coefficient (MFCC), Linear Prediction Coding Coefficient (LPCC), perceptual linear production (PLP), and RASTA-PLP in noisy conditions through using multivariate Hidden Markov Model (HMM) classifier. The behavior of the proposal system is evaluated using TIDIGIT human voice dataset corpora, recorded from 208 different adult speakers in both training and testing process. The theoretical basis for speech processing and classifier procedures were presented, and the recognition results were obtained based on word recognition rate.

Keywords

Speech Recognition, Noisy Conditions, Feature Extraction, Mel-Frequency Cepstral Coefficients, Linear Predictive Coding Coefficients, Perceptual Linear Production, RASTA-PLP, Isolated Speech, Hidden Markov Model

1. Introduction

Automatic speech recognition (ASR) is an interactive system used to make the speech machine recognizable.

How to cite this paper: Këpuska, V.Z. and Elharati, H.A. (2015) 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*, **3**, 1-9. http://dx.doi.org/10.4236/jcc.2015.36001

ASR as shown in the block diagram in **Figure 1** consists of two main parts. The first part, the signal modeling, known as front-end is used to extract the acoustic features from input speech signal using specific feature extraction algorithm. The second part, the statistical modeling, known as back-end, is used to match these features with reference model to generate the recognition result using one templet or classifier techniques [1], such as Hidden Markov Models (HMMs), Artificial Neural Network (ANN), Dynamic Time Warping (DTW), or Vector Quantization (VQ). Front-end is used to extract input speech signal into several short frames. Typically, each frame between 10 to 30 ms length reflects a number of useful physical characteristics of the input signal. The same processes are repeated for all subsequent frames. A new frame is overlapped to its previous frame typically ~10 ms to generate sequence of feature vectors and then passes to the next back-end part to select the most likely words out of all trained words as possible words. Back-end applies statistical modelling which is used to calculate the maximum likelihood based on reference models to select the most likely sequence of words. The performance of automatic speech recognition system based on acoustic model is totally dependent on the condition of training and testing data [2]. This means that the lack of noise robustness is the largely unsolved problem in automatic speech recognition research today. Indeed, the main challenges involved in designing speech recognition system are selecting the signal modelling, statistical modelling, and noise. The focus of this study is to experimentally evaluate the effectiveness of noise on different conventional and hybrid feature extractions algorithm using MFCC, LPCC, PLP, and RASTA-PLP through using multivariate HMM classifier and TIDIGIT speech corpora. This paper is organized as follows: Section 1, introduction; Section 2 describes the speech modeling; Section 3, details of different feature extraction techniques that are discussed, followed by a description of Hidden Markov Model as statistical modeling classifier in Section 4. Sections 4 and 5 include the result and the conclusion of the comparison done on all the eight above mentioned methods of speech extraction algorithms respectively.

2. Speech Pre-Processing

Sampling, pre-emphasis, frame blocking and windowing are the common steps needed to prepare input speech signal in order to extract the features [3].

2.1. Pre-Emphasis

The input speech signal has been digitally disturbed and corrupted by adding different values of realistic noises at SNRs ranging from 30dB to5dB as shown in **Figure 2** using *v_addnoise.m* Mathlab function.

2.2. Signal-to-Noise Ratio Estimation

First order High-pass filter (FIR) was used to flatten the speech spectrum and compensate for the unwanted high frequency part of the speech signal [4]. Equation (1) describes the transfer function of FIR filter in z-domain

$$y[n] = x[n] - Ax[n-1] \tag{1}$$

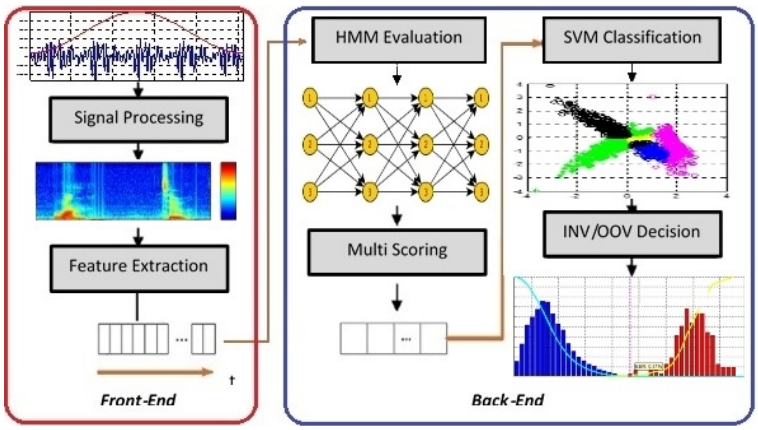

Figure 1. Speech recognition system.

$x[n]$: input speech signal.
$x[n-1]$: previous speech signal.
A: pre-emphasis factor which chosen as 0.975.

2.3. Frame Blocking and Windowing

In order to ensure the smoothing transition of estimated parameters from frame to frame, pre-emphasized signal $y[n]$ is blocked into 200 samples with 25 ms frame long and 10 ms frame shift. In addition to that hamming window as shown in Equation (2) was selected and applied on each frame in order to minimize the signal discontinuities at the beginning and the end of each frame as shown in **Figure 3**.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N} - 1\right),\ 0 \le n \ll N \tag{2}$$

n: windowed speech signal.
N: sampled speech signal.

3. Speech Feature Extraction

Feature extraction is used to convert the acoustic signal into a sequence of acoustic feature vectors that carry a good representation of input speech signal. These features are then used to classify and predict new words. To increase the feature evidence of dynamic coefficients, delta and delta delta can be devoted by adding the first and second derivative approximation to feature parameters [4]. In this research, several conventional and hybrid

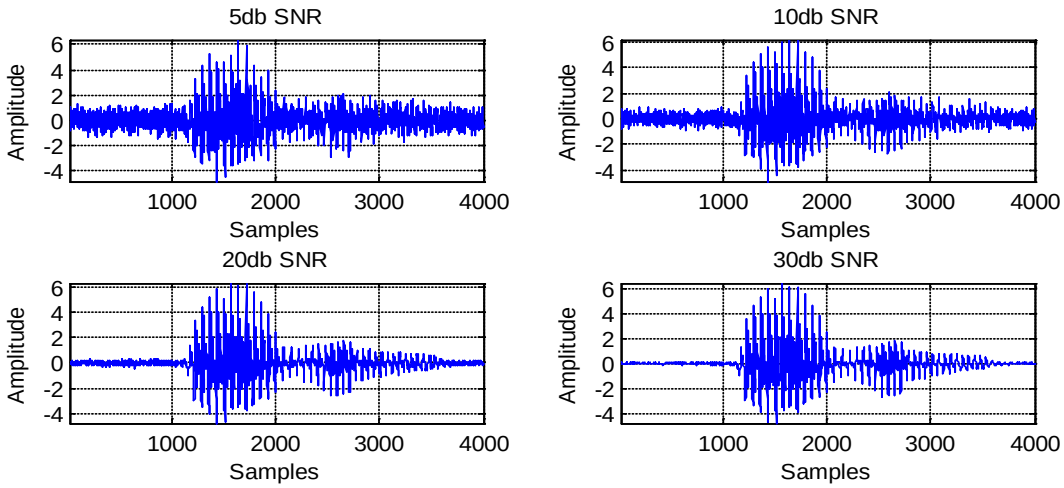

Figure 2. Word seven corrupted by different values of SNR.

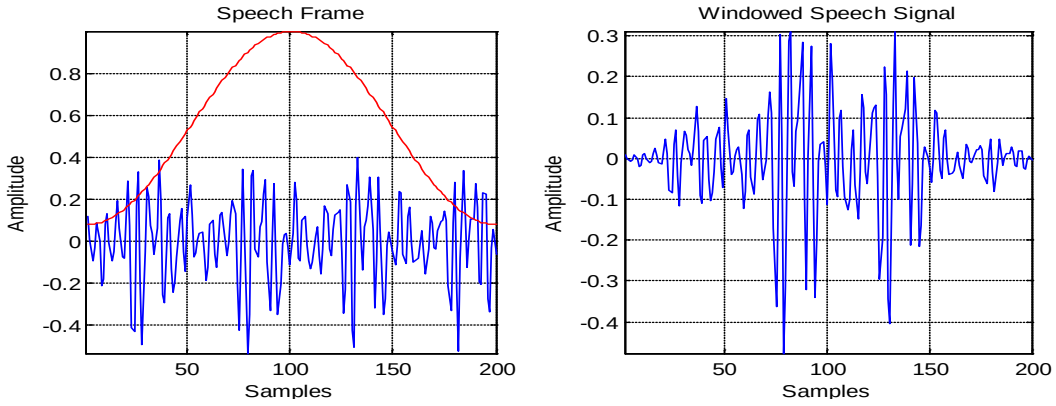

Figure 3. Hamming window.

feature extraction techniques were designed and tested using Matlab software to generate 39 parameter coefficients.

3.1. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is the most dominant method used to extract spectral features. MFCCs analysis is started by Appling Fast Fourier Transform (FFT) on the frame sequence in order to obtain certain parameters, converting the power-spectrum to a Mel-frequency spectrum, taking the logarithm of that spectrum, and computing its inverse Fourier transform [5] as shown in **Figure 4**.

3.2. Linear Prediction Coding Coefficients (LPCC)

LPCC is one of the earliest algorithms that worked at low bit-rate and represented an attempt to mimic the human speech and was derived using auto-correlation method [6]. Autocorrelation technique is almost an exclusively used method to find the correlation between the signal and itself by auto-correlating each frame of the windowed signal using Equation (3) as shown in **Figure 5**.

$$R(i) = \sum_{n=1}^{N_w - 1} s_w(n) s_w(n-i), \quad 0 \le i \le p \tag{3}$$

N_w Length of the window.
s_w Windowed segment.

3.3. Perceptual Linear Prediction (PLP)

Several spectral characteristics were calculated in order to match human auditory system. PLP computation was used as an autoregressive all-pole model to derive a more auditory-like spectrum based on linear LP analysis of speech. This kind of feature extraction was reached by making spectral analysis, frequency band analysis, equal-loudness pre-emphasis, intensity-loudness power law, and autoregressive modeling [7] as shown in **Figure 6**.

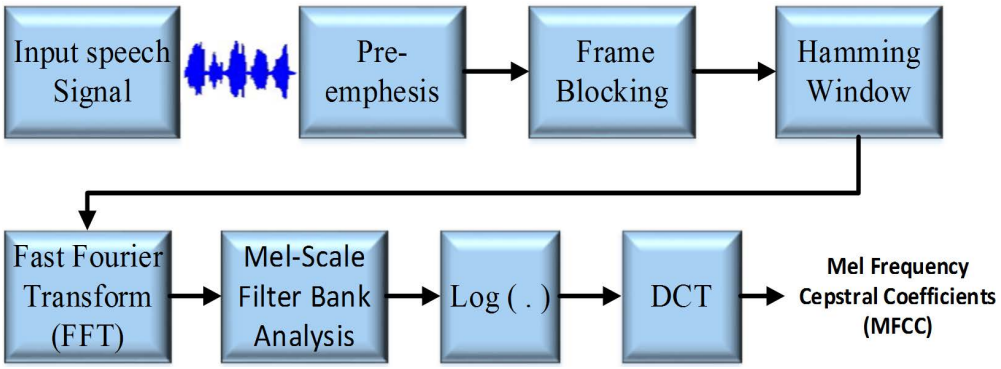

Figure 4. Mel Frequency Cepstral Coefficients (MFCC).

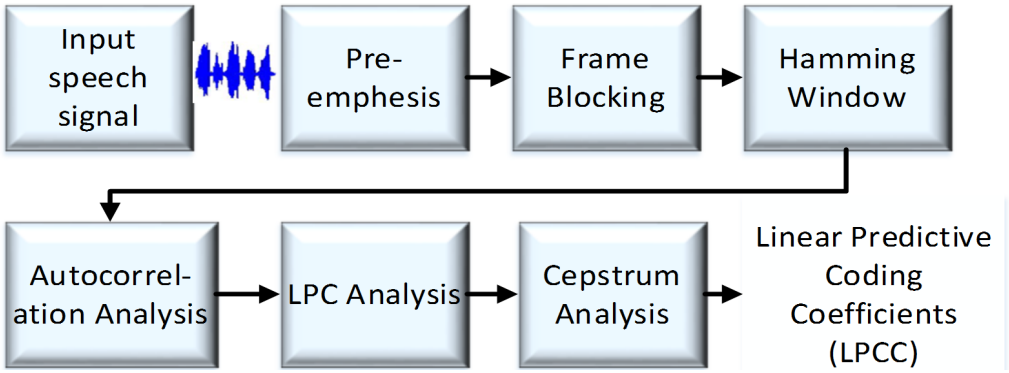

Figure 5. Linear Prediction Coding Coefficients (LPCC).

3.4. RASTA-PLP

A special band-pass filter was added to each frequency sub-band in traditional PLP algorithm in order to smooth out short-term noise variations and to remove any constant offset in the speech channel. **Figure 7** shows the most processes involved in RASTA-PLP which include calculating the critical-band power spectrum as in PLP, transforming spectral amplitude through a compressing static nonlinear transformation, filtering the time trajectory of each transformed spectral component by the band pass filter using Equation (4), transforming the filtered speech via expanding static nonlinear transformations, simulating the power law of hearing, and finally computing an all-pole model of the spectrum, as in the PLP [8].

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}\left(1 - 0.98z^{-1}\right)} \tag{4}$$

3.5. Hybrid Feature Extraction.

In order to obtain new features, hybrid algorithms are developed using a combination of previous feature extraction methods MFCC, LPC, PLP, and RASTA-PLP. Each of the previous features were designed to generate 13 coefficient parameters as shown in **Figure 8**. In each experiment, three kind of feature extractions were selected to provide 39 coefficient parameters in one vector as follows:

1) 13 MFCC + 13 LPC + 13 PLP.
2) 13 MFCC + 13 LPC + 13 RASTA-PLP.
3) 13 MFCC + 13 PLP + 13 RASTA-PLP.
4) 13 LPC + 13 PLP + 13 RASTA-PLP.

4. Statistical Modeling

Powerful statistical tools are used to test the previous feature extraction algorithms. HMM classifier is selected due to the ability of modeling non-linear aligning speech and estimating the model parameters [9] is to classify feature vectors and to predict unknown words based on evaluation, learning, and decoding processes. HMM is a finite-state machine characterized by a set of parameters hidden states, observations, transition probabilities,

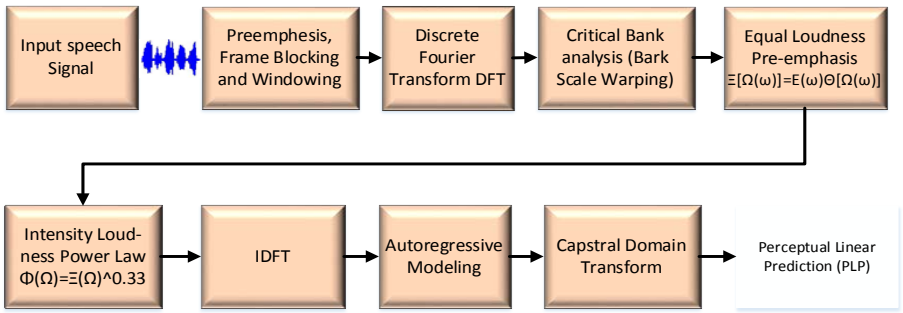

Figure 6. Perceptual Linear Prediction (PLP).

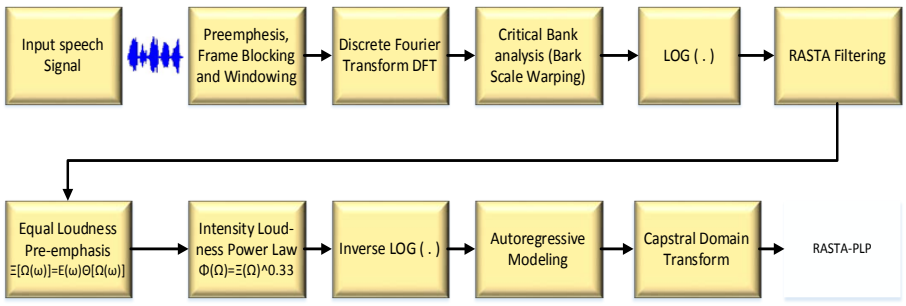

Figure 7. RASTA-PLP.

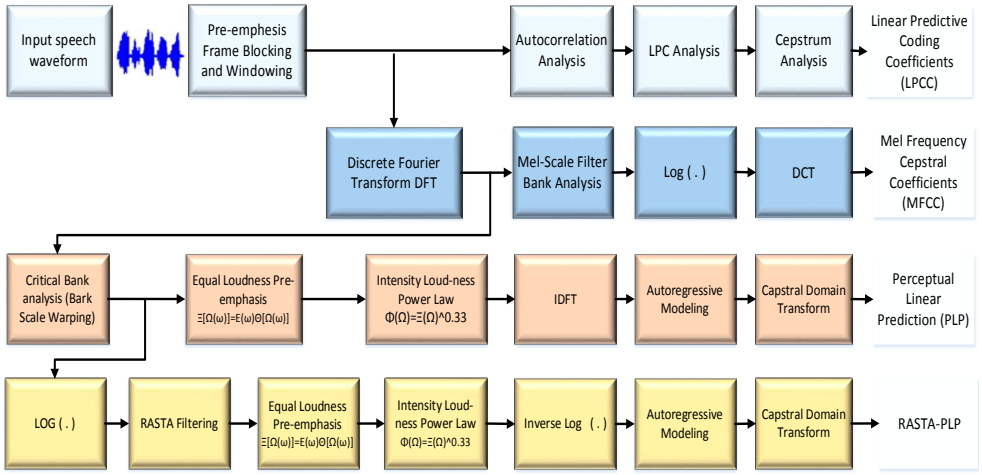

Figure 8. Hybrid feature extraction algorithm.

emission probabilities, and the initial state probabilities.

4.1. Evaluation

Probability of the observation sequence given in the model was computed using forward-backward dynamic programming. This algorithm was used to compute the probability that any sequence of states has produced the sequence of observations using Equation (5) as shown in **Figure 9**.

$$P(O \setminus \lambda) = \sum_{i=1}^{N} P(O_t q_t = \lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \tag{5}$$

4.2. Learning

In this step, all the model parameters (λ), mean, variance, transition probability matrix, and Gaussian mixtures were re-estimated using Baum-Welch algorithm as shown in **Figure 10**. Baum-Welch is used to learn and encode the characteristics of the observation sequence that best describes the process in order to recognize a similar observation sequence in the future [9]. The training model can be formed as Equation (6).

$$\lambda^* = \arg\max_{\lambda} \left[P(O|\lambda) \right] \tag{6}$$

4.3. Decoding

In order to find the state sequence that is most likely to have produced an observation sequence, Viterbi algorithm was used to find the optimal scoring path of state sequence as shown in **Figure 11**. The maximum probability of state sequences was defined in Equation (7), and the optimal scoring path of state sequence selected was calculated using Equation (8).

$$\delta t(i) = \max\left(P(q(1), q(2), \cdots, q(t-1); o(1), o(2), \cdots, o(t)|\lambda) \right) \tag{7}$$

$$q_T^* = \arg\max_{1 \le i \le N} \left[\delta_T(i) \right] \tag{8}$$

5. Results

The performance evaluation for the proposal speech recognition model was obtained. This system includes conventional and new hybrid feature extractions of MFCC, LPCC, PLP and RASTA-PLP, was trained and tested in clean [10] and noisy conditions in order to find the maximum word recognition rate through using Multivariate Hidden Markov Model (HMM) classifier. A number of experiments are carried out in different conditions using small vocabulary isolated words based on TIDIGITS corpora. The data consist of 2072 training file and 2486

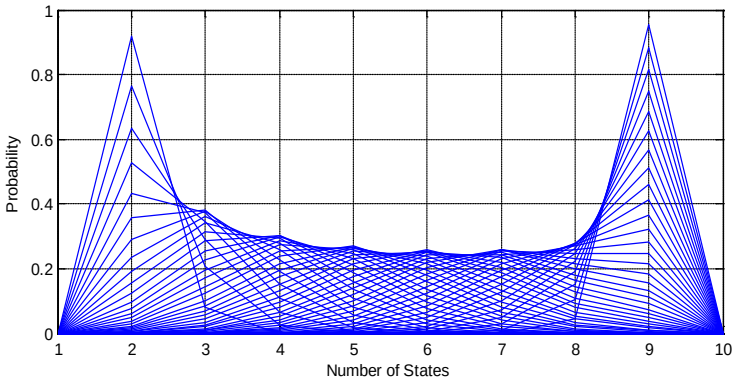

Figure 9. Forward α and Backward β probabilities in each state.

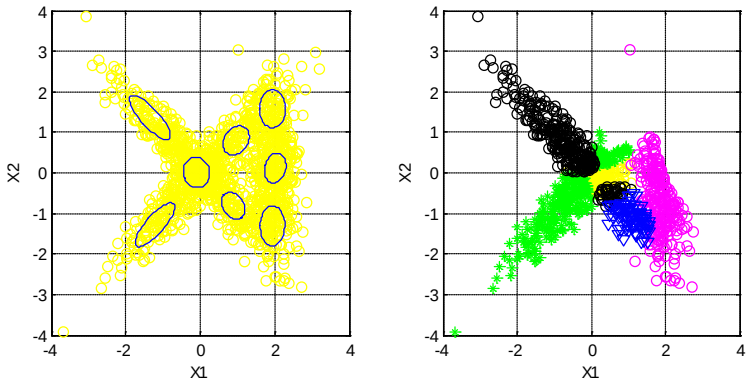

Figure 10. Eight dimensional Gaussian distribution.

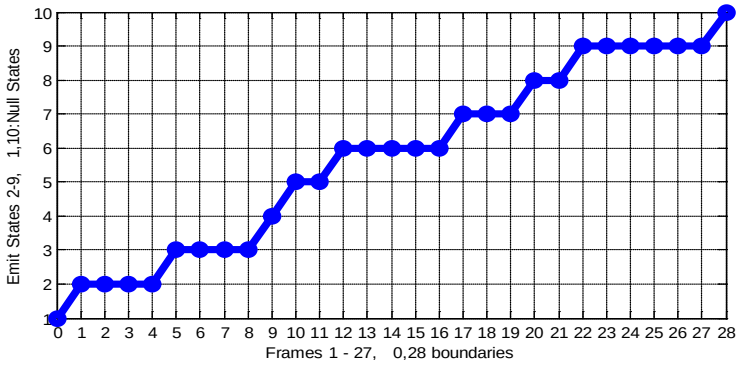

Figure 11. Viterbi trellis computation for 8-states HMM.

testing file, including eleven words (zero to nine and the letter o) recorded from 208 adult speaker males and females. For the purpose of fair comparison, all experiments were repeated using the same pre-reemphasis factor 0.975, covered by 25 milliseconds hamming window, and 10 milliseconds overlapping. 256-point Fast Fourier Transform (FFT) was applied to transforming 200 samples of speech from time to frequency domain. The resulting confidence level intervals for the recognition rate obtained in decoding process are listed in **Table 1**. All training data were modeled using 6, 8, 10 and 12 states. Each state has 2 to 8 multi-dimensional Gaussians Hidden Markov Model. The chart in **Figure 12** summarizes the recognition rate obtained for each feature extraction methods.

6. Conclusions

The objective of this research is to evaluate the performance of four feature extraction techniques MFCC, LPCC,

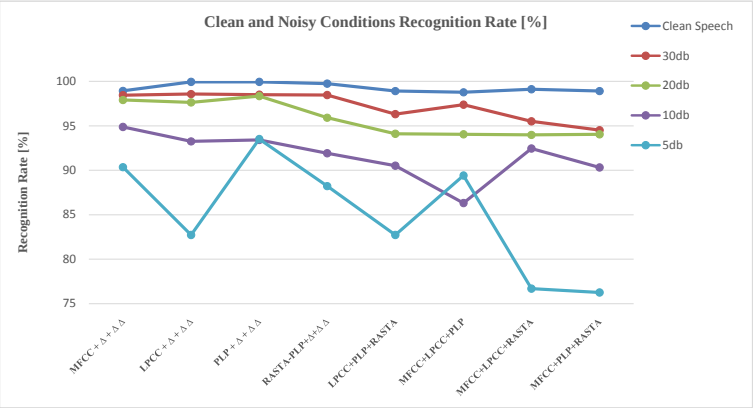

Figure 12. Recognition rate of conventional and hypered feature extractions.

Table 1. Recognition rate of different type feature extractions.

Feature Extraction Methods	Word Accuracy [%]				
	Clean Speech	SNR [dB]			
		30	20	10	5
MFCC + Δ + ΔΔ	98.95	98.45	97.92	94.87	90.37
LPCC + Δ + ΔΔ	99.95	98.59	97.63	93.27	82.73
PLP + Δ + ΔΔ	99.95	98.50	98.35	93.42	93.52
RASTA-PLP + Δ + ΔΔ	99.75	98.46	95.93	91.92	88.24
LPCC + PLP + RASTA	98.93	96.32	94.10	90.52	82.73
MFCC + LPCC + PLP	98.79	97.92	94.05	86.31	89.41
MFCC + LPCC + RASTA	99.12	95.50	94.00	92.45	76.69
MFCC + PLP + RASTA	98.93	94.53	94.05	90.32	76.25

PLP, RASTA-PLP and the combination of them is done by implementing a discrete-observation multivariate HMM-based on isolated word recognizer in MATLAB.

In clean speech, as shown in **Table 1** and **Figure 12**, the acoustic signals extracted using the individual algorithms LPCC and PLP give the best recognition rate. At 99.95%, LPCC and PLP separately provide the highest rate of recognition rate using 12 states and 4 Gaussian mixtures. Followed by the combination of MFCC, LPCC, and RASTA which provides a 99.12% recognition rate using the same number of states and Gaussian mixtures, the hybrid combination of LPCC, PLP, and RASTA represents the third highest recognition rate at 98.93% using 10 states and 3 Gaussian mixtures. Trailed by the combination of MFCC, LPCC, and PLP with a recognition rate of 98.79% using 10 states and 3 Gaussian mixtures, the lowest of the group, MFCC, provides a 98.95% recognition rate using 12 states and 4 Gaussian mixtures. When adding 30 db of realistic noises at SNR range to the input speech signal, individual LPCC method provides the best recognition rate by 98.59%. With the addition of 20 db, PLP provides the best recognition rate at 98.35%. When adding either 10 db or 5 db, individual MFCC provides the best rate of recognition at 94.87%, and 90.37% respectively.

References

[1] Këpuska, V. and Klein, T. (2009) A Novel Wake-Up-Word Speech Recognition System, Wake-Up-Word Recognition Task, Technology and Evaluation. *Nonlinear Analysis*: *Theory, Methods & Applications*, **71**, e2772-e2789. http://dx.doi.org/10.1016/j.na.2009.06.089

[2] Veisi, H. and Sameti, H. (2013) Speech Enhancement Using Hidden Markov Models in Mel-Frequency Domain.

Speech Communication, **55**, 205-220. http://dx.doi.org/10.1016/j.specom.2012.08.005

[3] Zhu, Q. and Alwan, A. (2000) On the Use of Variable Frame rate Analysis in Speech Recognition. 2000 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, **3**, 1783-1786.

[4] Rabiner, L. R. and Juang, B.-H. (1993) Fundamentals of Speech Recognition. Vol. 14, PTR Prentice Hall, Englewood Cliffs.

[5] Chetouani, M., Gas, B. and Zarader, J. (2002) Discriminative Training for Neural Predictive Coding Applied to Speech Features Extraction. *Proceedings of the* 2002 *International Joint Conference on Neural Networks*, **1**, 852-857. http://dx.doi.org/10.1109/ijcnn.2002.1005585

[6] Dave, N. (2013) Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition. *International Journal for Advance Research in Engineering and Technology*, **1**.

[7] Hermansky, H. (1990) Perceptual Linear Predictive (PLP) Analysis of Speech. *The Journal of the Acoustical Society of America*, **87**, 1738-1752.

[8] Hermansky, H., Morgan, N., Bayya, A. and Kohn, P. (1991) The Challenge of Inverse-E: The RASTA-PLP Method. 1991 *Conference Record of the* 25*th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, 4-6 November 1991, 800-804. http://dx.doi.org/10.1109/acssc.1991.186557

[9] Dugad, R. and Desai, U. (1996) A Tutorial on Hidden Markov Models. Signal Processing and Artificial Neural Networks Laboratory, Department of Electrical Engineering, Indian Institute of Technology, Bombay Powai, Mumbai, 400 076, India.

[10] Këpuska, V.Z. and Elharati, H.A (2015) Performance Evaluation of Conventional and Hybrid Feature Extractions Using Multivariate HMM Classifier. *International Journal of Engineering Research and Applications* (*IJERA*), **5**, 96-101.