RESEARCH ARTICLE OPEN ACCESS

Performance Evaluation of Conventional and Hybrid Feature Extractions Using Multivariate HMM Classifier

Veton Z. Këpuska, Hussien A Elharati
Electrical & Computer Engineering Department
Florida Institute of Technology, Melbourne, FL 32901, USA

ABSTRACT

Speech feature extraction and likelihood evaluation are considered the main issues in speech recognition system. Although both techniques were developed and improved, but they remain the most active area of research. This paper investigates the performance of conventional and hybrid speech feature extraction algorithm of Mel Frequency Cepstrum Coefficient (MFCC), Linear Prediction Cepstrum Coefficient (LPCC), perceptual linear production (PLP) and RASTA-PLP through using multivariate Hidden Markov Model (HMM) classifier. The performance of the speech recognition system is evaluated based on word error rate (WER), which is given for different data set of human voice using isolated speech TIDIGIT corpora sampled by 8 Khz. This data includes the pronunciation of eleven words (zero to nine plus oh) are recorded from 208 different adult speakers (men & women) each person uttered each word 2 times.

Keywords: feature extraction, likelihood evaluation, speech recognition, Mel Frequency Cepstrum Coefficient, Linear Predictive Coding, perceptual linear production, RASTA-PLP, Hidden Markov Model, word error rate.

I. INTRODUCTION

The rapid growth in communication technology has made possible for machine to recognize human languages and interact with human instructions [1].

Automatic Speech Recognition as shown in Fig.1 usually divided into two parts Front-End and Back-End. Front-end used to extract acoustic features from input speech signal using specific feature extraction algorithm, while Back-End matches this features with reference model to generate the recognition result using templet or classifier technique [2].

Feature extraction algorithm used to extract input speech signal into several short segments typically 10 to 30 ms. A number of unique coefficients are calculated and combined to produce a set of features.

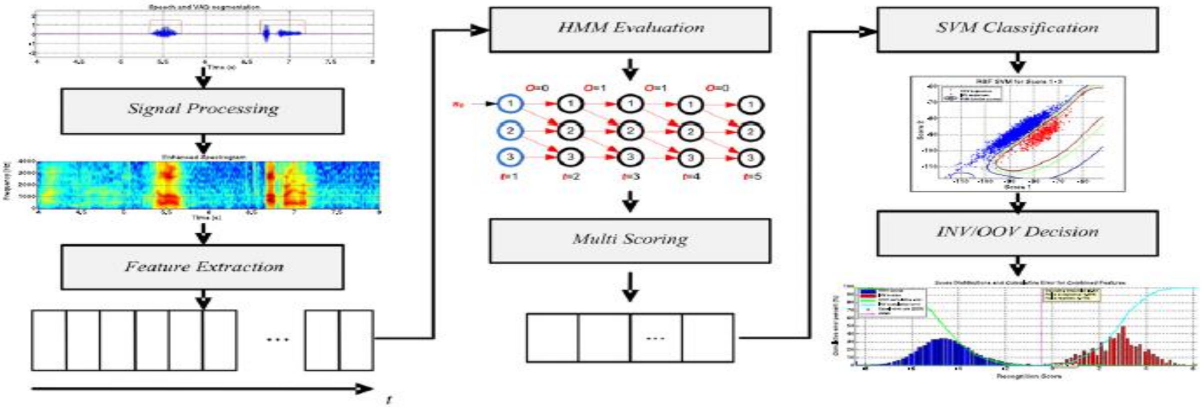

Figure 1: Speech recognition system

A new frame is overlapping to its previous frame typically ~ 10 ms. As a result, sequence of feature vectors is generated depending on speech length. On the other hand, Back-End applies statistical processes on those feature vectors which used to calculate the maximum likelihood based on reference models and selects the most likely sequence of words or phonemes[2, 3].

In this paper, we will evaluate the recognition performance of speech recognition system based on four different Front-End feature extractions algorithm, MFCC, LPCC, PLP, and RASTA-PLP.

For more robust recognition, a new hybrid feature extractions algorithm is claimed to have a good recognition rate. It is more useful to study the performance of hybrid Front-End and make a comparison with the conventional methods. Hidden Markov Model (HMMs) with Gaussian mixture emission pdfs also used as an isolated word classifier through this research.

The rest of this proposal are organized as follows. Section 2 provides a background on front end proposed analysis. Back-end analysis and a development of HMM classifier are provided in section 3. Section 4 is devoted to describe the results and analysis the data. Conclusion and references are given in sections 5 and 6 respectively.

II. FRONT-END ANALYSIS

Front-End analysis is responsible for converting speech acoustic signal into a sequence of acoustic feature vectors. The feature extractions, MFCC, LPCC, PLP, and RASTA-PLP are used to evaluate the performance of proposed automatic speech recognition system.

2.1. Preprocessing

Several common steps were taken onto speech signal in order to be ready for feature extraction calculations, include pre-emphasis, frame blocking and windowing[4].

2.1.1. Pre-emphasis

Pre-emphasis process is applied on input speech signal before extracting the features using high pass FIR filter by applying Equation (1) on input speech signal in order to flatten speech spectrum and compensate the unwanted high frequency part of the speech signal.

$$Y[n] = x[n] - A\, x[n-1] \qquad (1)$$

Where x[n] is the input speech signal, x[n-1] is the previous speech signal, and A is a pre-emphasis factor which chosen as 0.975.

2.1.2. Frame Blocking and Windowing

In order to minimize the signal discontinuities at the beginning and the end of each frame, hamming windows typically 25 ms long with a 10 ms shift is applied on pre-emphasized signal y[n] using Equation (2) as shown in Fig. 2.

$$w(n)=0.54 - 0.46 \cos (2\pi n / (N-1)) \qquad 0\leq n \leq N-1 \qquad (2)$$

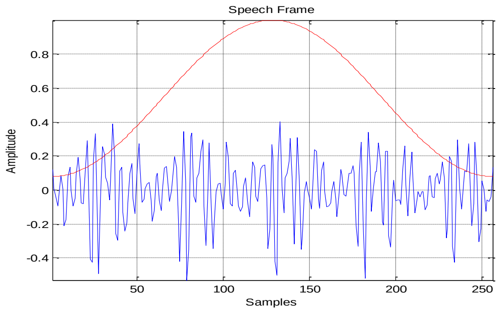

Figure 2: Hamming Window

2.2. Feature extraction

Feature extraction is a sequence of feature vectors carries a good representation of the input speech signal used to classify and recognize unknown words in speech recognition system [5]. In this research several feature extractions algorithm were designed using Matlab to extract 12 static parameters and 1 log power parameter with 13 first derivative and 13 second derivative dynamic parameter coefficients from each frame of input speech signal.

2.2.1 Mel Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) was used to extract spectral features from frames sequence using Fast Fourier Transform (FFT) which apply on each frame in order to obtain 256-point certain parameters, converting the power-spectrum to a Mel-frequency spectrum, and finally taking the logarithm of that spectrum and computing its inverse Fourier transform as shown in Fig. 3.

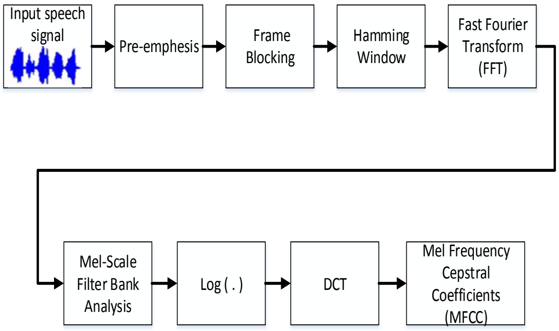
Input speech signal
Pre-emphasis
Frame Blocking
Hamming Window
Fast Fourier Transform (FFT)
Mel-Scale Filter Bank Analysis
Log (.)
DCT
Mel Frequency Cepstral Coefficients (MFCC)

Figure 3: MFCC feature extraction algorithm

2.2.2 Perceptual Linear Prediction (PLP)

PLP is used for deriving a more auditory-like spectrum based on linear LP analysis of speech, and calculate several spectral characteristics to match human auditory system using autoregressive all-pole model as shown in Fig. 4. This kind of feature extraction is reached by making some estimations of the psychophysical attributes of human hearing process [6].

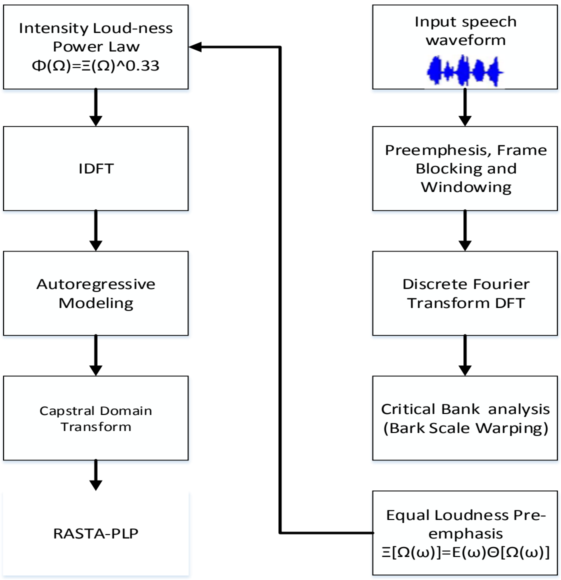
Intensity Loud-ness Power Law Φ(Ω)=Ξ(Ω)^0.33
IDFT
Autoregressive Modeling
Capstral Domain Transform
RASTA-PLP
Input speech waveform
Preemphesis, Frame Blocking and Windowing
Discrete Fourier Transform DFT
Critical Bank analysis (Bark Scale Warping)
Equal Loudness Pre-emphasis Ξ[Ω(ω)]=E(ω)Θ[Ω(ω)]

Figure 4: PLP feature extraction algorithm

2.2.3 Linear Prediction Coding Coefficients (LPCC)

LPC works at low bit-rate, which represents an attempt to mimic the human speech by compute a smoothed version of cepstral coefficients in automatic speech recognition system. Linear Prediction Coding coefficients were computed using auto-correlation method and Levinson-Durbin recursion by approximate the current sample as a linear of past sample as shown in Equation (3), and then convert LPC parameter into cepstral coefficients [7] as shown in Fig. 5.

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

Where N_w is the Length of the window and s_w is windowed segment

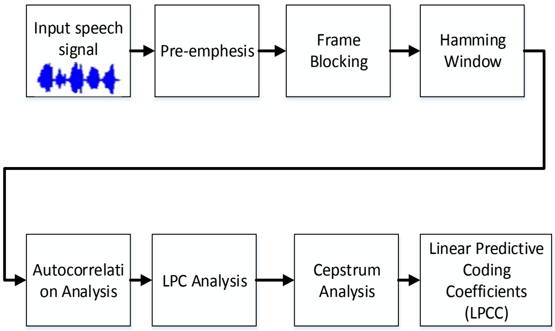

Figure 5: LPCC feature extraction

2.2.4 RASTA-PLP

RASTA-PLP is achieved by filtering the time trajectory in each spectral component. RASTA speech analysis technique is an improvement of the traditional PLP method that applies a special band-pass filter using Equation (4) to each frequency subband in order to smooth over short-term noise variations and to remove any constant offset in the speech channel [8] as shown in Fig. 6.

$$H(z) = 0.1 * \frac{2+z^{-1}-z^{-3}-2z^{-4}}{z^{-4}(1-0.98z^{-1})} \qquad (4)$$

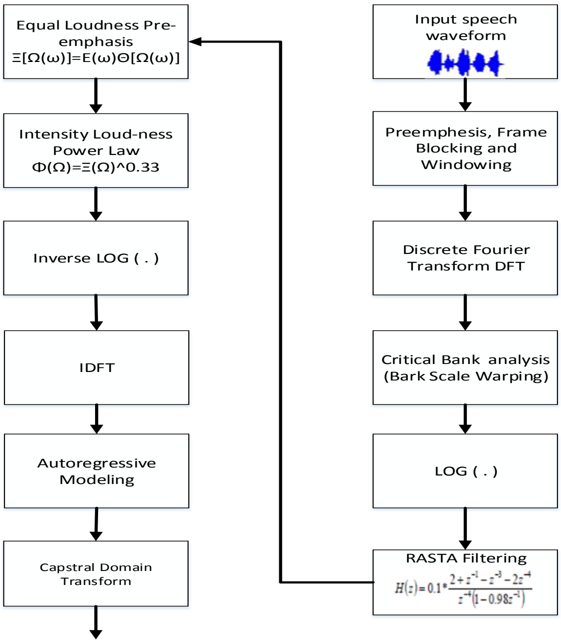
Equal Loudness Pre-emphasis Ξ[Ω(ω)]=E(ω)Θ[Ω(ω)]
Intensity Loud-ness Power Law Φ(Ω)=Ξ(Ω)^0.33
Inverse LOG (.)
IDFT
Autoregressive Modeling
Capstral Domain Transform
Input speech waveform
Preemphesis, Frame Blocking and Windowing
Discrete Fourier Transform DFT
Critical Bank analysis (Bark Scale Warping)
LOG (.)
RASTA Filtering
H(z)=0.1*(2+z^-1-z^-3-2z^-4)/(z^-4(1-0.98z^-1))

Figure 6: RASTA-PLP feature extraction algorithm

2.2.5 Hybrid features

In order to attend a new feature extractions and to make a distinction between previous features, the combination of previous features MFCC, LPCC, PLP and RASTA-PLP are taken to create a new hybrid features, each feature generates 13 parameter coefficients, as shown in Fig. 7.

Each three kinds of previous features are gathered in one vector to provide a 39 parameter coefficients using the following combination:

1) MFCC, LPCC, and PLP.
2) MFCC, LPCC, and RASTA-PLP.
3) MFCC, PLP and RASTA-PLP.
4) LPCC, PLP and RASTA-PLP.

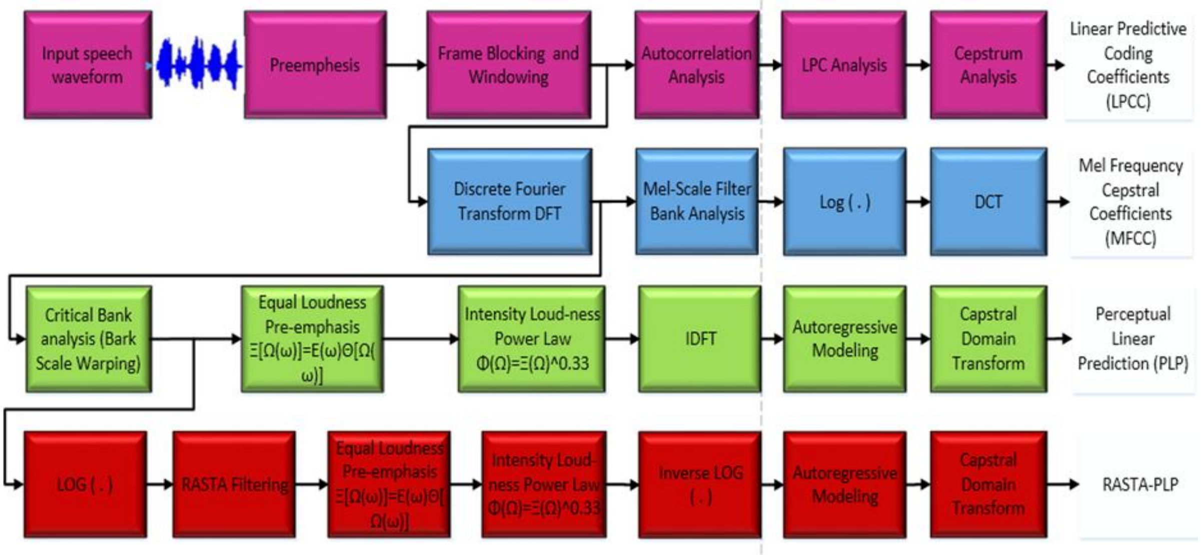
Input speech waveform
Preemphasis
Frame Blocking and Windowing
Autocorrelation Analysis
LPC Analysis
Cepstrum Analysis
Linear Predictive Coding Coefficients (LPCC)
Discrete Fourier Transform DFT
Mel-Scale Filter Bank Analysis
Log (.)
DCT
Mel Frequency Cepstral Coefficients (MFCC)
Critical Bank analysis (Bark Scale Warping)
Equal Loudness Pre-emphasis Ξ[Ω(ω)]=E(ω)Θ[Ω(ω)]
Intensity Loud-ness Power Law Φ(Ω)=Ξ(Ω)^0.33
IDFT
Autoregressive Modeling
Capstral Domain Transform
Perceptual Linear Prediction (PLP)
LOG (.)
RASTA Filtering
Equal Loudness Pre-emphasis Ξ[Ω(ω)]=E(ω)Θ[Ω(ω)]
Intensity Loud-ness Power Law Φ(Ω)=Ξ(Ω)^0.33
Inverse LOG (.)
Autoregressive Modeling
Capstral Domain Transform
RASTA-PLP

Figure 7: Hybrid feature extraction algorithm

III. BACK-END ANALYSIS

While feature extraction were used in front-end to extract the relevant characteristics from speech signal into number of feature vectors, Hidden Markov Model (HMM) are used in back-end to classify those features to generate the correct decision. HMM classifier is considered as a powerful statistical tool used in speech recognition and speaker identification systems, due to the ability to model non-linearly aligning speech and estimating the model parameters [9]. In this research mixtures of Gaussians were used to model the emission probability distribution function in each state of Hidden Markov Model (HMM).

In each iteration of HMM algorithm the observation parameters, transition probability matrix, the prior probabilities and Gaussian distribution were re-estimated in order to get good parameters in training process. This work has been performed using start_training.m Matlab function.

HMM parameters used to generate the likelihood scores which is performed using start_recognition.m Matlab function. This scores used to find the best bath between frames to recognize the unknown word.

3.1. Evaluation.

The first issue in HMM design is evaluating the probability that any sequence of states has produced the sequence of observations. Forward (α) and Backward (β) algorithms were used to find the overall result of the possible state sequence paths using Equation (5).

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

3.2. Training

Baum-Welch algorithm were used to adjusting or re-estimating the transition probability matrix and Gaussian mixture parameters (mean and covariance) that best describe the process. A multi-dimensional Gaussian PDF can be expressed using Equation (6). As shown in Fig. 8 .

$$p(x \setminus \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\Sigma^{1/2}} \exp(-\tfrac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)) \qquad (6)$$

Where d is the number of dimensions, x is the input vector, μ= E(x) is the mean vector, Σ is the covariance matrix.

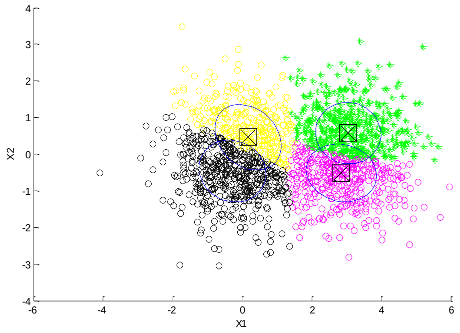

Figure 8: Four dimensional Gaussian distribution

Baum welch algorithm also used to learn and encode the characteristics of the observation sequence in order to recognize a similar observation sequence. The model can be formed as follows in Equation (7).

$$\lambda^{*} = \arg\max_{\lambda} [P(O \mid \lambda)] \qquad (7)$$

3.3. Decoding

Viterbi algorithm was used in decoding process to find the optimal scoring path of state sequence [10]. The maximal probability of state sequences is defined in Equation (8), and the optimal scoring path of state sequence selected calculated using Equation (9) as shown in Fig. 9.

$$\delta\, t(i) = \max(P(q(1), q(2),..,q(t-1);o(1),o(2),..,o(t)|\lambda) \qquad (8)$$

$$q^*_T = \arg\max_{1\le i\le N}[\delta_T(i)] \qquad (9)$$

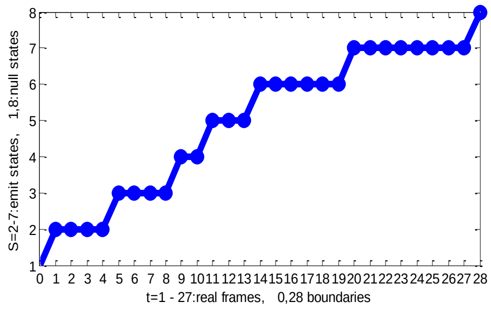

Figure 9: Viterbi trellis computation for 8-states HMM

IV. RESULTS

The performance evaluation for the proposed feature extractions based on, MFCC, LPCC, PLP, RASTA-PLP were obtained in order to find the maximum word recognition rate using Multivariate Hidden Markov Model (HMM) classifier.

The experiments are carried out using small vocabulary isolated words based on TIDIGITS corpora. This small vocabulary consists of 2072 training file and 2486 testing file, include eleven words (zero to nine and oh) recorded from 208 adult speakers male and female.

The confusion matrix of the average classification results were obtained using convenient and hybrid features. These features were trained and tested using 6, 8, 10 and 12 states and modeled by 2 to 8 multi-dimensional Gaussians Hidden Markov Model as shown in Table 1. The chart in Fig. 10 summarizes the recognition rate obtained for each feature extraction methods.

Table 1. Recognition rate with different type of feature extraction

Feature Extraction method	Wrong words											Total error count	Total correct count	Recognition rate %
	one	two	three	four	five	Six	Seven	eight	nine	Zero	Oh			
MFCC + Δ + ΔΔ	0	4	0	5	6	3	0	0	0	0	8	26	2460	98.95
LPCC + Δ + ΔΔ	0	0	0	0	0	0	0	0	0	1	0	1	2485	99.95
PLP + Δ + ΔΔ	0	0	0	0	0	0	0	0	0	1	0	1	2485	99.95
RASTA-PLP+ Δ+ΔΔ	2	0	0	0	0	0	0	3	0	0	0	5	2481	99.75
LPCC+PLP+RASTA	9	0	4	1	0	0	5	0	1	1	1	22	2464	98.93
MFCC+LPCC+PLP	4	0	2	0	1	0	3	3	6	1	5	25	2461	98.79
MFCC+LPCC+RASTA	0	2	1	0	1	0	4	3	4	3	0	18	2468	99.12
MFCC+PLP+RASTA	6	1	3	1	1	0	5	5	0	0	0	22	2464	98.93

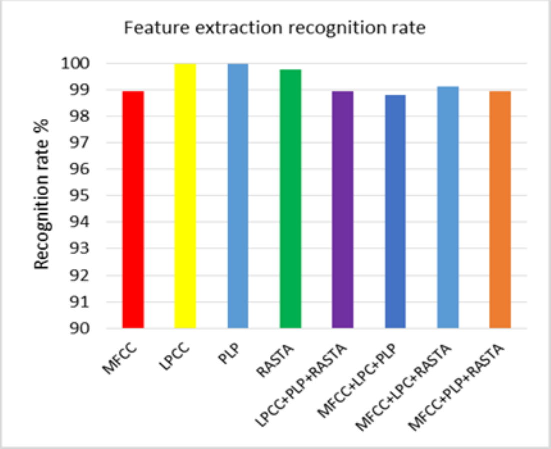

Figure 10: Overall recognition rate of conventional and hyper feature extractions

V. CONCLUSION

The objective of this research is to evaluate the performance of four feature extraction techniques MFCC, LPCC, PLP, RASTA-PLP and the combination of them is done by implementing a discrete-observation multivariate HMM-based isolated word recognizer in MATLAB.

Results as shown in Fig. 10 show that 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.

REFERENCES

[1] C. Kamm, M. Walker, and L. Rabiner, The role of speech processing in human–computer intelligent communication, Speech Communication, vol. 23, pp. 263-278, 1997.

[2] M. Kumar, R. Aggarwal, G. Leekha, and Y. Kumar, Ensemble Feature Extraction Modules for Improved Hindi Speech Recognition System, International Journal of Computer Science Issues (IJCSI), vol. 9, 2012.

[3] Q. Zhu and A. Alwan, On the use of variable frame rate analysis in speech recognition, in Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on, 2000, pp. 1783-1786.

[4] V. Këpuska and T. Klein, A novel wake-up-word speech recognition system, wake-up-word recognition task, technology and evaluation, Nonlinear Analysis: Theory, Methods & Applications, vol. 71, pp. e2772-e2789, 2009.

[5] M. Ursin, Triphone clustering in Finnish continuous speech recognition, Diplomityö, Teknillinen korkeakoulu, 2002.

[6] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, the Journal of the Acoustical Society of America, vol. 87, pp. 1738-1752, 1990.

[7] A. Abraham and S. M. Thampi, Intelligent Informatics: Proceedings of the International Symposium on Intelligent Informatics ISI'12 Held at August 4-5 2012, Chennai, India vol. 182: Springer, 2012.

[8] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, The challenge of inverse-E: the RASTA-PLP method, in Signals, Systems and Computers, 1991. 1991 Conference Record of the Twenty-Fifth Asilomar Conference on, 1991, pp. 800-804.

[9] R. Dugad and U. Desai, A tutorial on hidden Markov models, Signal Processing and Artificial Neural Networks Laboratory Department of Electrical Engineering Indian Institute of Technology, 1996.

[10] D. R. Reddy, Speech recognition by machine: A review, Proceedings of the IEEE, vol. 64, pp. 501-531, 1976.