International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 6 590 – 595

Speaker Identification and Spoken word Recognition in Noisy Environment using Different Techniques

Shaik Shafee
Dept. of Electronics & Communications
S. V. University College of Engineering
Tirupati, A.P, India
shafee.shaik@yahoo.co.in

Prof. B. Anuradha
Dept. of Electronics & Communications
S. V. University College of Engineering
Tirupati, A.P, India
anubhuma@yahoo.com

***Abstract* —** In this work, an attempt is made to design ASR systems through software/computer programs which would perform Speaker Identification, Spoken word recognition and combination of both speaker identification and Spoken word recognition in general noisy environment. Automatic Speech Recognition system is designed for Limited vocabulary of Telugu language words/control commands. The experiments are conducted to find the better combination of feature extraction technique and classifier model that will perform well in general noisy environment (Home/Office environment where noise is around 15-35 dB). A recently proposed features extraction technique Gammatone frequency coefficients which is reported as the best fit to the human auditory system is chosen for the experiments along with the more common feature extraction techniques MFCC and PLP as part of Front end process (i.e. speech features extraction). Two different Artificial Neural Network classifiers Learning Vector Quantization (LVQ) neural networks and Radial Basis Function (RBF) neural networks along with Hidden Markov Models (HMMs) are chosen for the experiments as part of Back end process (i.e. training/modeling the ASRs). The performance of different ASR systems that are designed by utilizing the 9 different combinations (3 feature extraction techniques and 3 classifier models) are analyzed in terms of spoken word recognition and speaker identification accuracy success rate, design time of ASRs, and recognition / identification response time .The testing speech samples are recorded in general noisy conditions i.e.in the existence of air conditioning noise, fan noise, computer key board noise and far away cross talk noise. ASR systems designed and analyzed programmatically in MATLAB 2013(a) Environment.

***Keywords* -** *Speech recognition, speaker identification, speech features extraction techniques,Hidden markov models, Learning Vector quantization neural networks,Radial basis funct neural networks.*

*****

I. INTRODUCTION

Speech recognition means converting speech into text or some standard transcribed data which can use as an input for further processing. Speech recognition also called as Automatic Speech Recognition (ASR) or computer speech recognition. Speech Recognition systems are used as voice controlled interface between human beings and artificial machines or computers. The advancement of speech or voice recognition automation process improves the interface between human beings and machine in numerous applications. Human beings are more comfortable to interact with computers or automatic machines through speech rather than other primitive interfaces such keyboards and other pointing devices. This motivated the researchers to work in automatic speech recognition since the 1950's.

ASR systems can be designed through computer programs. The designing of ASR systems mainly consist of two tasks. The first task is to extract the feature vectors from speech signals using Signal/Speech processing techniques, and the other task is designing the word/sentence /speaker models using classifiers, template matching and other model designing techniques.

ASR system accuracy depends on many factors such as Environment (the type of noise), Speaker (Sex, Age, and psychical state), and Voice tones (quiet, normal, shouted), Speed (slow, normal, and fast), Vocabulary (Characteristics of available training data: specific or generic vocabulary).The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy is measured in terms of performance accuracy which is usually rated with word error rate (WER) or Command Success Rate (%) where as speed is measured with the real time factor (Recognition or Identification time taken by ASR).

II. RELATED WORK

K. H. Davis, R. Biddulph and S. Balashek designed a spoken digit recognition circuit to deal with 10 digit series when spoken by a single talker [1]. Lawrence. R. Rabiner, Stephen .E .Levinson, Aaron. E. Rosenberg and Jay .G. Wilpon (1979) described a speaker-independent isolated word recognition system which is based on the use of multiple templates for each word in the vocabulary [2]. B. H. Juang; L. R. Rabiner (1991) published their paper on Hidden Markov Models for Speech Recognition [4].Richard P. Lippmann (1988) submitted a paper on neural network classifiers for Speech Recognition [3]. Douglas A. Reynolds (1995) experimented automatic speaker recognition using Gaussian Mixture Speaker Models [5]. Sahar E. Bou-Ghazale and John H. L. Hansen (2000) compared the speech recognition performance between traditional and the proposed features under stress [7]. Qifeng Zhu, Abeer Alwan (2003) proposed analysis based Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise [8]. Rafik Djemili, Mouldi Bedda, and Hocine Bourouba (2004) proposed an algorithm for Arabic isolated digit recognition [9]. Florian Honig, Georg Stemmer, Christian Hacker, Fabio Brugnara (2005) developed a revised processing steps for Perceptual Linear Prediction (PLP) that combines the advantages of both MFCC and PLP [10]. Manal El-Obaid, Amer Al-nassiri, Iman Abuel Maaly (2006) presented a paper on recognition of isolated Arabic speech phonemes using artificial neural networks [11]. Iosif Mporas, Todor Ganchev, Mihalis Siafarikas, Nikos Fakotakis, Department of Electrical and

Computer Engineering, University of Patras (2007) compared different feature extraction techniques for the task of speech recognition [12]. R. Schluter, I. Bezrukov, H. Wagner, H. Ney (2007) introduced an acoustic feature set based on a Gammatone filterbank for large vocabulary speech recognition [13]. Khalid Saeed and Mohammad Kheir Nammous (2007) discussed a Speech-and-Speaker (SAS) Identification System for spoken Arabic digit recognition [14]. Ji Ming, Member and Timothy J. Hazen, James R. Glass and Douglas A. Reynolds (2007) investigated the speaker identification and verification problem when speech signals are corrupted with environmental noises where the characteristics of the noise are not known [15]. Meysam Mohamad pour, Fardad Farokhi (2009) presented an advanced method for Persian language speech recognition to classify speech signals with the high accuracy at the minimum recognition time [16]. Wouter Gevaert, Georgi Tsenov, Valeri Mladenov (2010) investigated speech recognition classification performance using two standard neural networks such as Feed-forward Neural Network (NN) with back propagation algorithm and Radial Basis Functions Neural Networks [17]. Fu Guojiang (2011) proposed a Novel Isolated Speech Recognition based on Neural Networks [18]. Recognition of the words was carried out in speaker dependent mode and has used same data for both training and testing purpose. He has chosen 16 Linear Predictive cepstral coefficients with 16 parameters from each frame as feature extraction. Fatma zohra Chelali, Amar.Djeradi, Rachida.Djeradi (2011) have investigated Speaker Identification System based on PLP Coefficients and Artificial Neural Networks [19]. Mondher Frikha, Ahmed Ben Hamida (2012) compared the performance of ANN and Hybrid HMM and ANN Architectures for Robust Speech Recognition [20]. Djellali Hayet and Laskri Mohamed Tayeb (2012) described different approaches for vector quantization in Automatic Speaker Verification [21]. Addou Djamel, Selouani Sid Ahmed, Malika Boudraa, and Bachir Boudraa (2012) introduced an efficient front-end for distributed Speech Recognition over Mobile [22]. Hamdy K. Elminir, Mohamed Abu ElSoud, L. M. Abou El-Maged (2012) experimented different feature extraction techniques and analyzed the speech recognition evaluation parameters such as recognition success rate(%), training time ,feature extraction time and PCA conversion time [23]. Mahmoud I. Abdalla, Haitham M. Abobakr and Tamer S. Gaafar (2012) presented a paper on DWT and MFCCs based feature extraction method for Isolated Word Recognition [24].

Finally, we aim to analyze the comparative study for the better combination of speech features extraction and classifier techniques for spoken word recognition and speaker identification in more common noise environment (Home / Office environment).

III. FEATURE EXTRACTION TECHNIQUES

The task of feature extraction techniques is to transform a speech signal into a set of parameters that more economically represents the pertinent information in the original speech. There are several speech analysis techniques available for extracting different features of speech signals . Three different speech feature extraction techniques Mel frequency cepstral coefficients (MFCC) and Perceptual linear prediction (PLP) which are the most popular acoustic feature extraction techniques used in speech recognition and a recently proposed Gammatone frequency cepstral coefficients (GFCC) feature extraction technique which fits human auditory system are chosen for experiments.

A. Mel-Frequency Cepstral Coefficients (MFCC)

In MFCC features extraction, the speech samples are first processed through pre-emphasis which is a high pass filter to cancelling out the effect of glottis, and then the power spectrum is computed from the windowed speech signal, different types of windowing functions are available among which Hamming window is more commonly used in speech technology. Psychophysical studies of human auditory perception shown that the frequency content of speech does not follow linear scale and so there is a need to convert the liner frequency scale to non linear. A non linear transformation of the frequency called “Mel scale” frequency warping is applied as shown in equation.

$$f_{mel} = 2595.\log_{10}\left(1 + \frac{F_{hz}}{700}\right) \quad (1)$$

The Mel frequency filter bank is a series of triangular band pass filters which mimics the human auditory system. Power spectrum of each successive speech frame is effectively deformed in frequency according to the critical-band Mel scale and amplitude in usual decibel or logarithmic scale. Mel filter-bank contains typically 24 to 40 triangular filters which have a 50% overlap [10]. MFCC feature vectors are extracted by applying inverse discrete cosine transform on log magnitude on each speech frame.

B. Perceptual Linear Prediction (PLP) Coefficients

In PLP features extraction, the process starts with the computation of power spectrum from the windowed speech signal, and then the frequency warping into the trapezoidal shaped bark scale filters are applied. The combination of three steps frequency warping, smoothing and sampling are integrated into a single filter-bank called Bark filter-bank. And then equal-loudness pre-emphasis weights applied which was introduced by Hermansky as shown in equation (2) to consider the frequency sensitivity of human hearing [10].

$$E(f) = \frac{(f^2+1.44.10^6)f^4}{(f^2+1.6.10^5)^2(f^2+9.61.10^6)} \quad (2)$$

Then the equalized values are transformed according to the power law by rising to the power of 0.33. The resulting warped spectrum is further processed by linear prediction (LP) analysis and computing the predictor coefficients of an approximated signal that has this warped spectrum as a power spectrum. Finally the PLP coefficients are obtained from the predictor coefficients by a recursion that is equivalent to the logarithm of the model spectrum followed by an inverse Fourier transform. Though there are many similarities between MFCC and PLP, the following differences are considered in PLP [9], triangular shaped Mel filter-bank is replaced by a Trapezoidal Bark filter-bank, pre-emphasis is replaced by the equal-loudness weighting of the spectrum and the duplication of the first and last filter-bank value before linear prediction (LP) is dropped. The Bark-scale filter bank typically consists of 19 to 21 trapezoidal shaped filters.

C. Gammatone Frequency Cepstral Coefficients (GFCC)

Patterson and Moore proved that the gammatone function is the best fit to the human auditory system. The gamma tone function is defined in time domain by its impulse response as shown in equation (3). It was also suggested that a 4th order filter (n=4) would be a good model for the human auditory filter [6].

$$G(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi ft + \phi) \quad (3)$$

The Equivalent Rectangular bandwidth (ERB) of the auditory filter with the function has been proposed as shown in equation (4).

$$\text{ERB} = 24.7 \times \left(\frac{4.37}{1000}\, f + 1\right) \quad (4)$$

Gammatone features are extracted for every 100Hz of frequency shift (i.e. 10ms of overlapping in time domain).

All the speech waveforms are recorded with 16 KHz sampling frequency using MATLAB functions. In all the above three feature extraction techniques, feature vectors are extracted from each overlapping frame of 10ms (i.e. 160 samples). By using End point detection algorithm the unvoiced samples are removed at both the ends of speech waveforms before applying for framing/windowing. The recorded speech samples may not be having of same length though the same words are collected from the same speaker, so there is a need to process the speech waveforms to a fixed set of feature vectors (same number of feature vectors to all the speech waves) to apply for training/modeling ASR system. By using k-means algorithm with 'K' centroids, all the speech wave forms feature vectors are processed to a fixed set of k-feature vectors to each of the speech wave form.

IV. CLASSIFIER TECHNIQUES

After extracting the feature vectors from the speech signals, the next task is designing of spoken word and speaker models using classifier or pattern recognition techniques. There are many approaches among which statistical based and template based approaches are widely used modeling techniques in speech recognition technology. Most of the existing speech recognition systems are designed based on Hidden Markov models (HMMs) which is a statistical framework that supports both acoustic and temporal modeling [4]. Artificial Neural Networks (ANNs) are highly interconnected networks of relatively simple processing elements that operate in parallel. Neural nets offer many advantages over existing classifier approaches [3]. In this work, two different Artificial Neural Networks Learning Vector Quantization Artificial neural networks (LVQ-ANN) and Radial basis function artificial neural networks (RBF-ANN) are chosen for experiments along with the conventional Hidden Markov Models.

A. Hidden Markov Models

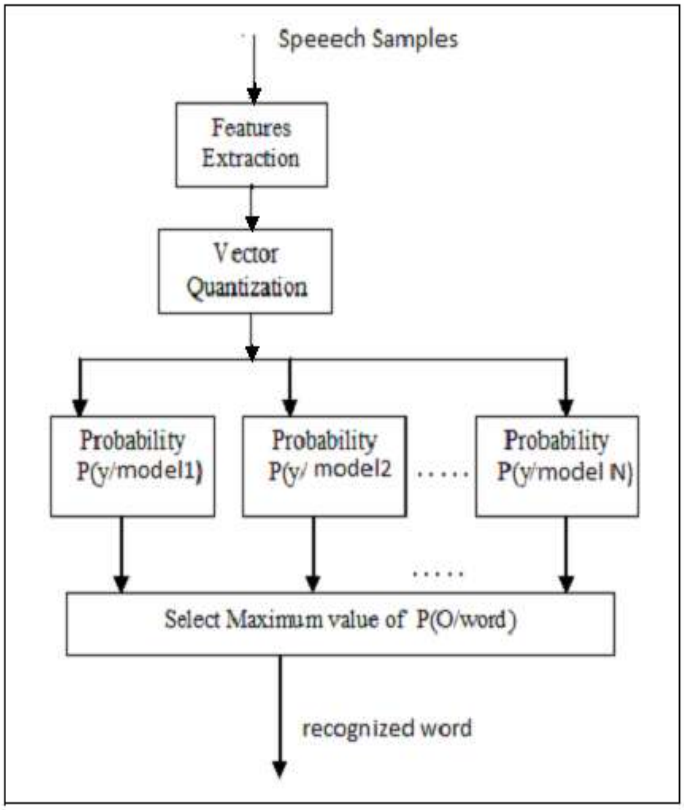

Figure 1. Recognition of spoken word using HMMs

HMMs are very popular statistical stochastic approach which is used as back-end task since many years in speech recognition systems. There are three basic problems arise in applying HMM models to Speech recognition task. Problem1 can be treated as speech recognition problem: How efficiently the Probability P (O/ λ) be computed for the given Observation sequence O= (O_1, O_2, O_3…...O_T) and the HMM model (λ = {A, B, π}). Problem2 is treated as hidden part of model: to find the Optimal state sequence for the given Observation sequence O = (O_1, O_2, O_3…… O_T) and the HMM model (λ = {A, B, π}). And problem 3 can be treated as the training problem: How the Model (λ = {A, B, π}) be adjusted to maximize the probability P (O/λ).

Codebook will be generated using the feature vectors of all speech wave forms which are collected for training purpose. HMMs are built to each spoken word using quantized feature vectors. In Recognition task the unknown word is applied to all the designed HMMs and calculates the P (O/ λ). The HMM for which the maximum value is computed is chosen as the recognized word. The same process flow is applied in speaker identification where the spoken words HMMs are replaced by speaker HMMs for a particular spoken word.

B. Learning Vector Quantization Neural Networks

LVQ networks basically have two layers competitive layer followed by linear layer as shown in Fig. (2)., the competitive layer learns to classify input vectors in much the same way as the competitive layers of Self-Organizing maps. The linear layer transforms the competitive layer's classes into target classifications defined by the user. The classes learned by the

competitive layer are referred as subclasses and the classes of the linear layer are referred as target classes.

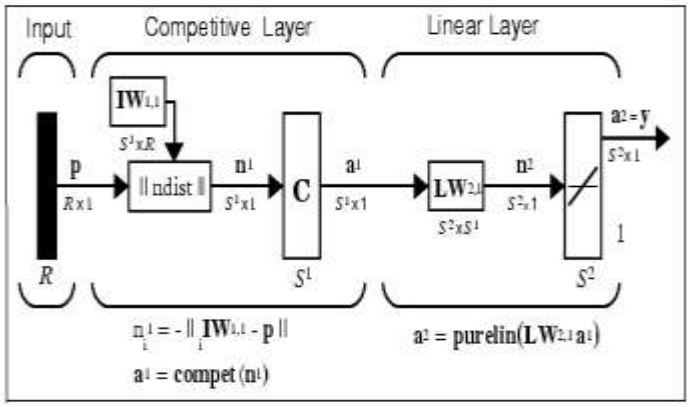

Figure 2. Learning Vector quantization networks with R inputs

Where 'R' is number of elements in input vector, 'S^{1}' is number of competitive neurons and 'S^{2}' is number of linear neurons in Fig. (2).

LVQ network is described with the MATLAB function '**newlvq'** as shown in in equation (5).

$$lvqnet = newlvq(PR, S1, PC, LR, LF) \quad (5)$$

Where:

- *PR is an R-by-2 matrix of minimum and maximum values for R inputelements.*
- *S1 is the number of first layer hidden neurons.*
- *PC is* an *S2 element vector of typical class percentages.*
- *LR is the learning rate (default 0.01).*
- *LF is the learning function (default is 'learnlv1').*

C. *Radial Basis Neural Networks*

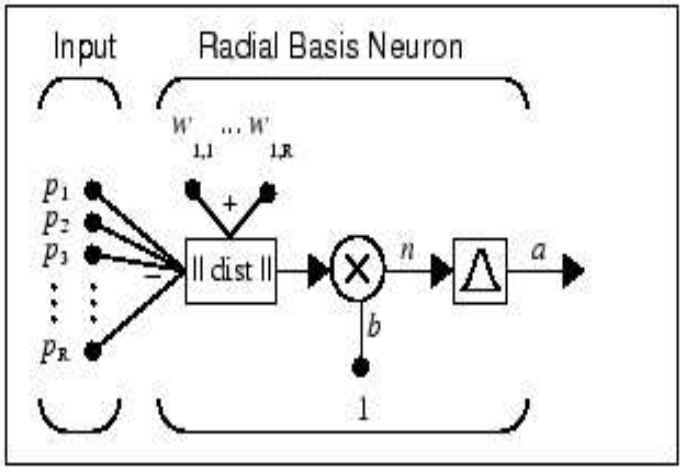

Figure 3. Neuron model of Radial basis function with R inputs

The basic Radial basis neuron is shown in Fig. (3).The final output 'a' is a Radial basis function of 'n'. The net input to the '**radbas**' transfer function is the vector distance between weight vector 'w' and input vector 'p' multiplied by the bias 'b'.

$$n = ||w-p||b \quad (6)$$

$$a = radbas(n) = exp(-n^{2}) \quad (7)$$

Radial basis function neural network is described with MATLAB function 'newrbe' as shown in equation (8)

$$rbnet = newrbe(P, T, SPREAD) \quad (8)$$

Where the function '**newrbe**' takes matrices of input vectors 'P' and target vectors 'T' and a spread constant 'SPREAD' for the radial basis layer, and returns a network with weights and biases such that the outputs are exactly 'T' when the inputs are 'P'.

The SPREAD should be chosen such that it is large enough so that the active input regions of the radbas neurons overlap enough so that several radbas neurons always have fairly large outputs at any given moment. This makes the network function smoother and results in better generalization for new input vectors occurring between input vectors used in the design. However, SPREAD should not be so large that each neuron is effectively responding in the same, large, area of the input.

V. EXPERIMENTAL SETUP

Eight different Telugu words (aagu (ఆగు)-STOP, edama (ఎడమ)-LEFT, kadulu (కదులు) -START, kudi (కుడి) - RIGHT, kinda (కింద) -DOWNWARDS, paina (పైన) - UPWARDS, venakki (వెనక్కి) - BACKWARDS, munduki (ముందుకి)- FORWARD) from 10 different Telugu speakers (5 Male and 5 Female Speakers from age group of 20-50) have been recorded with 16 KHz sampling frequency, each word recorded 10 times from each speaker in clean environment for training purpose and the same words recorded 5 times from each speaker in noisy environment (Home / Office environment where noise levels are at 15-40 dB) for testing. Total 800 samples (10 speakers*8 words*10 times) for training and 400 samples (10 speakers*8 words*5 times) for testing are recorded in '.wav' form using MATLAB functions.

Then speech wave is segmented to overlapped frames and then computed the feature vectors (MFCC /PLP /GFCC coefficients) with 10ms overlapped frames. By using End point detection algorithm the unvoiced samples are removed at both the ends of speech waveforms (Gold wave tool is used for End point detection). As the recorded speech samples may not be having same duration though the same words are recorded from the same speaker, K-means algorithm with 'k' centroids is applied to all the speech waveforms and extracted feature vectors are processed to a fixed set of k-feature vectors to each of the speech wave form. Then models are designed with below three approaches.

- Codebook has been generated using feature vectors and the HMMs of each spoken word are trained using the codebook. HMMs will be designed by Iterating the process such that the Model ($\lambda = \{A, B, \pi\}$) be adjusted to maximize the probability P (O/λ).
- LVQ Neural networks are designed by using the feature vectors as input ,user defined target vector 'T' as output , learning function ('learnlv1' is chosen in this project) and by proper learning rate(default 0.01).

- Radial Basis function networks are designed by using the feature vectors as input, user defined target vector 'T' as output and by choosing the proper SPREAD value.

VI. RESULTS

The following three different test categories have been experimented: Spoken word recognition (to recognize the word among 8 different Telugu words), Speaker identification (to recognize the speaker for a particular spoken word), and the combination of both spoken word recognition and speaker identification. As all the designing models are stochastic and so may not be giving the same results in every instance of same experiment, so each combination of features extraction and classifier technique of ASR is tested many times (10 times) and the average results are tabulated for all the 9 combinations (3 different feature extraction techniques* 3 different modelling/classifier techniques) for the above mentioned 3 test categories.The performance is analyzed by varying the number of cepstral coefficients to 12, 14, 16 and 18 from each frame of speech signal and by varying the no. of feature vectors in k-means algorithm, the rest of the other parameters are set to their best typical values.

The results are consolidated by averaging the values acquired from the experiments conducted for 10 times for each combination of feature extraction and classifier technique. The total success count of spoken word recognition, speaker identification, and combination of both spoken word recognition and speaker identification rate for 400 test samples and the total response time taken for recognizing the spoken word and identification of speaker, and the time taken for designing the model are tabulated. The important parameters for complete 400 test samples are shown in Tables I, II and III.

TABLE I. HIDDEN MARKOV MODELS

Number of cepstral coefficients	MFCC		PLP		GFCC	
	Combined Success count	Response time (sec)	Combined Success count	Response time (sec)	Combined Success count	Response time (sec)
12	192	458	178	198	175	291
14	224	456	182	191	204	293
16	234	453	208	200	238	294
18	235	458	209	200	239	296

TABLE II. LEARNING VECTOR QUANTIZATION NETWORKS

Number of cepstral coefficients	MFCC		PLP		GFCC	
	Combined Success count	Response time (sec)	Combined Success count	Response time (sec)	Combined Success count	Response time (sec)
12	221	360	179	237	231	276
14	245	439	178	322	230	351
16	241	551	182	425	237	461
18	244	712	195	568	241	630

TABLE III. RADIAL BASIS FUNCTION NETWORKS

Number of cepstral coefficients	MFCC		PLP		GFCC	
	Combined Success count	Response time(sec)	Combined Success count	Response time(sec)	Combined Success count	Response time(sec)
12	193	247	161	118	284	176
14	226	248	196	120	290	176
16	228	245	206	121	**311**	**175**
18	234	264	208	145	287	303

VII. CONCLUSIONS

The performance of ASRs are analyzed in terms of spoken word recognition success rate, speaker identification success rate, response time taken for spoken word recognition/Speaker identification, and designing time taken for training/modeling the ASRs for Limited Vocabulary of Telugu words. The ASR system is designed using clean data (speech samples recorded in clean environment) and tested in general noisy environment (Home/Office). HMM models with GFCC features showing slightly better results over PLP and MFCC for proper selection of number of cepstral coefficients per speech frame. ASR systems are further designed using Learning Vector Quantization neural networks and the performance is analyzed for different number of cepstral coefficients per frame ,the recognition/identification success rate is slightly improved compare to the HMM models, but the response time taken also increased compare to HMMs. Further ASR systems are designed using Radial Basis Function neural networks, spoken word recognition and speaker identification success rates are significantly improved with GFCC feature extraction over MFCC and PLP, and the response time also reduced to less than 1 second in all the 3 feature extraction techniques. Form the experiments it is understood that ASR system with a combination of RBF networks and GFCC feature extraction technique is outperformed, it is also observed that the time taken for designing the ASR systems using RBF is significantly less compare to HMM and LVQ classifiers.

REFERENCES

[1] K. H. Davis, R. Biddulph, and S.Balashek, "Automatic Recognition of spoken Digits", The Journal of the Acoustical Society of AMERICA, Volume 24, Number 6, pp. 637-642, November 1952.

[2] Lawrence. R. Rabiner, Stephen .E .Levinson, Aaron. E. Rosenberg, and Jay .G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques", IEEE Transactions. Acoustics, Speech, Signal Processing. ASSP-27, No.4, pp: 336-349, August 1979.

[3] RICHARD P. LIPPMANN, "Neural Network Classifiers for Speech Recognition", The Lincoln Laboratory Journal, Volume 1, Number 1 ,1988.

[4] B. H. Juang; L. R. Rabiner, "Hidden Markov Models for Speech Recognition", Technometrics, Vol. 33, No. 3,pp. 251-272, Aug., 1991.

[5] Douglas A. Reynolds, "Automatic Speaker Recognition Using Gaussian Mixture Speaker Models", VOLUME 8, NUMBER 2, THE LINCOLN LABORATORY JOURNAL, 1995.

[6] Roy D. Patterson and John Holdsworth," A functional model of neural activity patterns and auditory images", Advances in Speech, Hearing and Language Processing,Volume 3, Part B, pp: 547-563,1996 .

[7] Sahar E. Bou-Ghazale, Member, IEEE, and John H. L. Hansen, Senior Member, IEEE, " Comparative Study of Traditional and

Newly Proposed Features for Recognition of Speech Under Stress", IEEE transactions on speech and audio processing, VOL. 8, NO. 4, JULY 2000.

[8] Qifeng Zhu, Abeer Alwan, "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise" , Elsevier, Computer Speech and Language 17 ,pp:381–402, 2003.

[9] Rafik Djemili, Mouldi Bedda, and Hocine Bourouba, "Recognition of Spoken Arabic Digits Using Neural Predictive Hidden Markov Models", The International Arab Journal of Information Technology, Vol. 1, No. 2, July 2004

[10] Florian Honig, Georg Stemmer, Christian Hacker, Fabio Brugnara, "Revising Perceptual Linear Prediction (PLP)" September 4-8, Lisbon, Portugal , INTERSPEECH 2005

[11] Manal El-Obaid, ,Amer Al-nassiri, Iman Abuel Maaly, "Arabic Phoneme Recognition Using Neural Networks", Proceedings of the 5th WSEAS International Conference on Signal Processing, Istanbul, Turkey, pp99-104,May, 2006.

[12] Iosif Mporas, Todor Ganchev, Mihalis Siafarikas, Nikos Fakotakis, "Comparison of Speech Features on the Speech Recognition Task", Journal of Computer Science 3 (8): 608-616, 2007.

[13] R. Schluter, I. Bezrukov, H. Wagner, H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition", ICASSP, 2007.

[14] Khalid Saeed, Member, IEEE, and Mohammad Kheir Nammous, "A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image", IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, VOL. 54, NO. 2, APRIL 2007

[15] Ji Ming, Member, IEEE, Timothy J. Hazen, Member, IEEE, James R. Glass, Senior Member, IEEE, and Douglas A. Reynolds, Senior Member, IEEE, "Robust Speaker Recognition in Noisy Conditions", IEEE transactions on audio, speech, and language processing, VOL. 15, NO. 5, JULY 2007.

[16] Meysam Mohamad pour, Fardad Farokhi, "An Advanced Method for Speech Recognition", International Scholarly and Scientific Research & Innovation World Academy of Science, Engineering and Technology,Vol:3 2009.

[17] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, Senior Member,IEEE, "Neural Networks used for Speech Recognition", JOURNAL OF AUTOMATIC CONTROL,UNIVERSITY OF BELGRADE,VOL.20:1-7,2010.

[18] Fu Guojiang, "A Novel Isolated Speech Recognition Method based on Neural Network", 2nd International Conference on Networking and Information Technology IPCSIT vol.17, IACSIT Press, Singapore, 2011.

[19] Fatma zohra Chelali, Amar.Djeradi, Rachida.Djeradi , "Speaker Identification System based on PLP Coefficients and Artificial Neural Network" , Proceedings of the World Congress on Engineering , London, U.K, Vol II, WCE 2011, July 6 - 8, 2011,

[20] Mondher Frikha, Ahmed Ben Hamida, "A Comparative Survey of ANN and Hybrid HMM/ANN Architectures for Robust Speech Recognition", American Journal of Intelligent Systems, 2(1): 1-8, 2012.

[21] Djellali Hayet, Laskri Mohamed Tayeb, "Using Vector Quantization for Universal Background Model in Automatic Speaker Verification", Proceedings ICWIT 2012.

[22] Addou Djamel, Selouani Sid Ahmed, Malika Boudraa, and Bachir Boudraa, "An Efficient Front-End for Distributed Speech Recognition over Mobile", International Journal of Computer and Communication Engineering, Vol. 1, No. 3, September 2012.

[23] Hamdy K. Elminir , Mohamed Abu ElSoud, L. M. Abou El-Maged, "Evaluation of Different Feature Extraction Techniques for Continuous Speech Recognition" , International Journal of Information and Communication Technology Research Volume 2 No. 12, December 2012

[24] Mahmoud I. Abdalla, Haitham M. Abobakr, Tamer S. Gaafar, "DWT and MFCCs based Feature Extraction Methods for Isolated Word Recognition", International Journal of Computer Applications (0975 – 8887) Volume 69– No.20, May 2013.

[25] Mark Hudson Beale, Martin T. Hagan, Howard B. Demuth, Neural Network Toolbox-User's Guide, R2014a, The Math Works, Inc.,2014