Speech and Language Recognition using MFCC and DELTA-MFCC

Samiksha Sharma[1], Anupam Shukla[2] and Pankaj Mishra [3]

[1] Department of Information Technology, IIITM
Gwalior, MP,India

[2] Department of Information Technology, IIITM
Gwalior, MP,India
[3] Department of Electronics And Telecommunication,SSIPMT
Raipur, CG,India

***Abstract*— In this paper, a model is proposed to recognize speech and language by speech signal for countries like India where many languages are spoken. We have used MFCC and delta-MFCCs as acoustic features. We used a supervised learning technique to train ANN (Artificial neural network) as recognizer. To train this model resilient back propagation algorithm and radial basis function neural network used and results are compared. The ANN model tries to classify the input with respect to a set of words and languages.**

In this work four Indian languages Hindi, English, Sanskrit and Telugu are used. A multi speaker Speech recognition and language recognizer proposed for these four Indian languages

***Keywords*— ANN, Speech Recognition, language Recognition, back propagation algorithm, Radial basis function**

I. INTRODUCTION

To provide benefit of Information technology in each and every where in India, voice based interface for various computer related task is most suitable. To develop a voice based interface for countries like India is very difficult task as here approximate twenty five languages spoken. The Speech signal conveys many levels of information like what is spoken, language, speaker, gender, sentiments etc. In this paper a multilingual speech system is developed for utterances and language recognition using artificial neural network. We used neural network since HMM make very large assumption about data and number of parameters need to set in HMM are very huge. There are number of applications of multilingual speech recognition system used like conversation over phone in different languages and as teaching assistance in many languages etc.

Speech recognition basically classify as speaker dependent, speaker independent and multi speaker speech recognition system. This paper basically focuses on isolated word, multi speaker speech and language recognition system. Speech recognition has progressed very much in last decades but still there is need to research in multilingual speech and language recognition as more than half of world's population is multilingual. There are many commercial software available for monolingual speech recognition such as Dragon, natural speaking, via voice, Speak Q etc. but no commercial multilingual speech recognition system is available for Indian languages. Research and development on speech recognition and speaker recognition methods and techniques has been undertaken for over five decades and it continuous to be an active area. There are various approaches used for speech and speaker recognition such as acoustic and phonetic approach, pattern matching approach, and knowledge based approach, connectionist approach and support vector machine approaches etc [1]. In which connectionist approach is youngest field in speech recognition and due to its simplicity and uniformity its hardware implementation is also simple. The connectionist approach is inspired by human brain. Due to high degree of parallelism complex problem solve efficiently. Although many researchers have done work in speech recognition but for Hindi and other Indian languages this area is not explored much. Recognition accuracy of speech recognition system depends on various factors such as environment (signal to noise ratio), transducer, channel (Band amplitude), speakers (age, sex, physical state etc.), speech style (voice tone, pronunciation) and vocabulary size.

In paper [2] a speaker-dependent, isolated word recognizer for Hindi is described. Features are extracted using LPC and recognition is carried out using HMM. In paper [3], speech recognition is done using neural network and hidden Markov modal for Arabic isolated word/sentence recognition. Neural network is trained using Al-Aloui algorithm and results are compared to HMM recognizer. Speech recognition for sentence is done using speech segmentation and average level crossing rate identification is used for speech segmentation.

Multilingual speech recognition in mobile application is proposed in paper [8]. In this paper results from different language recognizer combined and compared the performance to a single large multilingual system and using an explicit language identification system to select the appropriate recognizer. Experiment conducted on three language English-French-Mandarin data.

In paper [7] a constructive algorithm is used for train a feed forward MLP and using this isolated word recognition system is simulated. In proposed system incremental training

procedure is used and experiment performed on 10 isolated words. MFCC techniques are used for feature extraction phase.

Many researches have been done for multilingual speaker recognition system using ANN, and there are using different model like statistical methods Hidden Markov Model (HMMs), Harmonic Product Spectrum (HPS) [4, 5].

Classification or recognition phase, there are various methods In identification of the speakers in single/different language ANNs have been used. Pattern used as vector quantization technique from signal processing to store features in codebooks [5, 6].

The system presented in this paper is uses neural network with back propagation algorithm for classification as it overcomes various assumptions related to data in HMM. The paper is organized as follows.. Section 2 explains methodology of developing multilingual speech and language recognition system. Section 3 describes experiment and result. Section 4 concluded this research and discuss about future scope.

II. .METHODOLOGY

Speech recognition is the process of converting speech signal to text so that it can be further utilized in various applications. The outputs of speech recognition can be words that used in data entry, command and controls etc or it can be a sentences. Here basically two phases occur in recognition system training phase and testing phase. In training phase a recognition model is trained using training data and then prepared modal is tested over in testing phase.

In this work we have used MFCCs and delta –MFCCs to transform speech signal into sequence of acoustic vectors as they realize human auditory system. Research on human auditory system shown that it does not follow a linear scale in listening. Thus for a tone have actual frequency say f, is mapped on Mel-scale. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. The Mel-scale mapped linearly below 1000hz and logarithmically above 1000hz.Steps used for extracting MFCCs and delta MFCCs shown in fig(2).

After speech acquisition front end process is performed on speech signal for finding out mfcc coefficients. Front-end process consists of different sub processes. At first speech signal is re-sampled because some speech signal may be at high frequencies and some on low. So to make all speech signals at one level re-sampling is done.

Then re-sampled signal passes through pre-emphasis filter which emphasize higher frequencies. A simple digital filter used for such compensation is given-

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

Here a is positive parameter to control degree of pre-emphasis filtering and usually is chosen to be less than 1. Since speech is quasi stationary signal and for reducing error in recognition frame feature extraction should be done on stationary signal. So speech is divided into short length frame of 25ms (256 samples) and frame shift is 10ms (192 samples). Overlapped frame is used for maintaining continuity between frames. Overlapping ensure high correlation between coefficients of consecutive frames.

The segment of waveform used to determine each parameter vector is usually referred to as a window. Here we used hamming window for windowing which is multiplied by each frame. Output signal is given by equation-

$$Y(n) = W(n)\, X(n)$$

Here X (n) is input signal and W (n) is window. The equation of hamming window is given as-

$$W(n) = 0.54 - 0.46 * \cos(2\pi n/N-1)$$

Where $0 \leq n \leq N-1$

After windowing, Fast Fourier transform of each frame is calculated to convert a signal from time domain to frequency domain. Further this signal is apply to Mel - frequency filter bank to scaling signal from linear scale to mel-scale and log is calculated to separate excitation signal and vocal tract impulse response. Below equation shows relation between Mel frequency and linear frequency-

$$Mf = 2595 * \log_{10}(1+f/100)$$

Then discrete cosine transform is calculated to convert signal into time domain. The result is MFCCs.

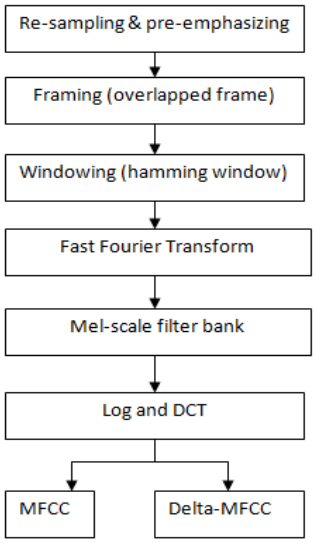


Fig. 1 MFCC and Delta-MFCC Feature Extraction

Cepstral coefficients provides better local spectral properties for analyzing frame but it doesn't provide dynamic spectral information i.e. what are the trajectories of the MFCC coefficients over time. So for this purpose delta of MFCCs are calculated using equation-

$$D_t = \frac{\sum_{n=1}^{N} n\,(C_{t+n} + C_{t-n})}{2\sum_{n=1}^{N} n^2}$$

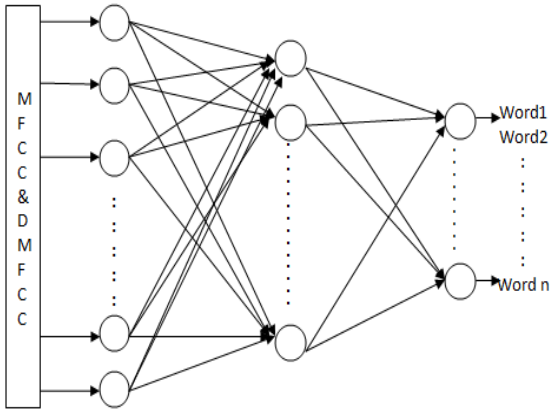
Usually N=2 is taken. In this experiment we have calculated 11 MFCCs and 11 delta-MFCCs so that features can represent local and transient properties for frames.

2.1 Classification

Classifier recognizes utterances on the basis of acoustic properties of speech signal. Two major steps are involved to develop a speech recognition system, training and testing phase. In training classifier are trained using feature vectors. Feed forward back propagation neural network and radial basis function neural network are used as classifier. Both models are described in following subsections.

2.1.1)Back propagation network :

Feed forward NN used as classifier and trained using resilient back propagation algorithm. ANN has capability to classified unknown pattern because it learns patterns behavior. Extracted feature set with target matrix is used to train NN since supervised learning is used to train. Architecture of classifier is shown in fig. (2)



fig(2) ANN classifier for speech recognition

Single hidden layer used in ANN and neurons in hidden layer depends on various factors i.e. number of neurons in input, output layer and amount of training samples [10]. Here n number of neurons used in output layer which represent number of words or languages to be classified. The input layer is composed by number of neurons MFCCs and delta MFCC coefficients extracted from each frame.

2.1.2) Radial Basis Function Network:

Radian basis function neural network is static feed forward neural network that has two layer single hidden layer and one output layer. In hidden unit of RBF network Gaussian or other kernel basis activation functions. RBF neural network much faster than multi layer perceptron trained using back propagation algorithm. They are less susceptible with non stationary input since speech is also non stationary type of signal so RBF is more appropriate classifier for speech recognition.

III. EXPERIMENT & RESULT

Database consists of speech signals from 10 speakers. Sentence “AB ISS BAAR TUM JAO” is uttered by all the speakers in four Indian languages Hindi, English, Sanskrit and Telugu. The total number of words is 18, 5 for each of Hindi and English and 4 for each of Telugu and Sanskrit. So for this experiment total number of utterances is 180. MFCCs and delta-MFCCs are extracted using DSP toolbox of Mat lab and total 22 features per frame are calculated.

Performance of back propagation algorithm and radial basis function has been computed. In Classifier that used back propagation algorithm, log sigmoid function used in hidden layer and linear transfer function used in output layer. Two separate neural networks are created for speech recognition and language recognition. Here Feature set is same as input for each ANN but target matrix differ for ANNs. Hence there are total 18 words so output layer of speech recognizer consist 18 neurons and output layer of language recognizer consists of 4 neurons.Here overall speech recognition performance is obtained 83.89% and language recognizer’s performance is 83.3%.

In second experiment, radial basis function network used as classifier. Overall word recognition rate obtained using this modal is 91.7 and language recognition rate is 91.1%.

IV. CONCLUSION

In this paper, we focused on multilingual multi speaker speech recognition and language recognition system using ANN. Two separate ANNs are develop for language and Speech recognition system and trained using back propagation algorithm and radial basis function network. Here we have used MFCCs and delta-MFCCs as acoustic features. The present work is limited to small vocabulary and four languages. In future we will develop system for large vocabulary and more languages.

REFERENCES

[1] Wiqas Ghai and Navdeep Singh. Article: Literature Review on Automatic Speech Recognition. *International Journal of Computer Applications* 41(8):42-50, March 2012. Published by Foundation of Computer Science, New York, USA.

[2] Pruthi T, Saksena, S and Das, P K *Swaranjali “Isolated Word Recognition for Hind Language using VQ and HMM”* International Conference on Multimedia Processing and Systems (ICMPS), IIT Madras.

[3] Mohamad Adnan Al-Alaoui, Lina Al-Kanj, Jimmy Azar, Elias Yaacoub,”*Speech Recognition using Artifical Neural Network and*

Hidden arkov Modal" IEEE Multidisciplinary Engineering Education ol3, No. 3, 2008 .

[4] Roopa A.Thorat , Ruchira.A.Jadhav , "Speech Recognition System ", *International Conference on Advances in Computing, Communication and Control*, 2009.

[5] Sandipan Chakroborty and Goutam Saha, "mproved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter", International Journal of Signal Processing 5, Winter 2009.

[6] Akram M. Othman, May H. Riadh , "Speech Recognition Using Scaly Neural Networks", *proceeding of world academy of science, engineering & technology*, 2008.

[7] Masmoudi, S.; Chtourou, M.; Ben Hamida, A., "Isolated word recognition system using MLP neural network constructive training algorithm," *Systems, Signals and Devices, 2009. SSD '09. 6th International Multi-Conference on* , vol., no., pp.1,6, 23-26 March 2009

[8] Hui Lin and Jui-Ting Huang and Francoise Beaufays and Brian Strope and Yun-hsuan Sung,"Recognition of Multilingual Speech in Mobile Applications," ICASSP 2012.

[9] Zissman, M.A.; Singer, E., "Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling," *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol.i, no., pp.I/305,I/308 vol.1, 19-22 Apr 1994.

[10] Leena, M.; Srinivasa Rao, K.; Yegnanarayana, B., "Neural network classifiers for language identification using phonotactic and prosodic features," *Intelligent Sensing and Information Processing, 2005. Proceedings of 2005 International Conference on* , vol., no., pp.404,408, 4-7 Jan. 2005

[11] Syama, R " *Speech Recognition System for Malayalam*" Department of Computer Science .Cochin University of Science & Technology, Cochin 2008.

[12] Deivapalan, P G and Murthy, H A, *" A syllable-based isolated word recognizer for Tamil handling OOV words,"* The National Conference on Communications, pp. 267-271 2008.

[13] Saurabh Karsoliya, "Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture," International Journal of Engineering Trends and Technology, Volume 3, Issue6, 2012

[14] Rahul Kala, Harsh Vazirani, Anupam Shukl, Ritu Tiwari, "Fusion of Speech and Face by Enhanced Modular Neural Network," ICISTM 2010.

[15] Akram M. Othman, and May H. Riadh, "Speech Recognition Using Scaly Neural Networks", World Academy of Science, Engineering and Technology, vol. 38, 2008.

[16] Chee Peng Lim, Siew Chan Woo, Aun Sim Loh, Rohaizan Osman, "Speech Recognition Using Artificial Neural Networks," wise, vol. 1, pp.0419, First International Conference on Web Information Systems Engineering (WISE'00)-Volume 1, 2000.

[17] J.H. McClellan, R.W. Schafer, M.A. Yoder, "Signal Processing First", Prentice Hall, 2003, pp. 415-426.

[18] S. Haykin, "Neural Networks: a comprehensive foundation", 2nd Edition, Prentice Hall, 1999.