International Journal of Research in Advanced Engineering and Technology
ISSN: 2455-0876; Impact Factor: RJIF 5.44
Received: 11-09-2019; Accepted: 13-10-2019
www.newengineeringjournal.in
Volume 5; Issue 4; October 2019; Page No. 47-50

Language Identification and Speech Recognition using MFCC and DELTA-MFCC

Pardeep Sangwan
Department of ECE, Maharaja Surajmal Institute of Technology, New Delhi, India

Abstract
This research proposes a prototype to identify speech and language for multiple languages. Acoustic features such as MFCC and delta-MFCCs are used for this work. At the back-end i.e. at the classifier stage Artificial Neural Network (ANN) and radial basis function are used. ANN is based on supervised learning approach for the purpose of speech and language recognition. Back propagation algorithm is used for training ANN. ANN model attempts to classify the input speech signals on the basis of a set of words and languages. This research is based on four Indian languages Hindi, Telugu, Sanskrit and English are used.

Keywords: artificial neural network, speech recognition, language identification, radial basis function

1. Introduction

In order to deliver advantage of Information technology in every look and corner of India, interface based on speech signals for many computer related tasks is utmost appropriate. Developing a speech-based interface is a very difficult task for countries like India as multiple languages are spoken here. The Speech signal carries different types of information like language, sentiments, speaker, gender etc. In this paper, using artificial neural networks, a multiple language speech model is built for language identification. Neural network used in this paper is HMM that makes assumptions about information and the parameters that are essential to be set within HMM. On the other hand, speech recognition systems based on multiple languages are used as communication over the phone and as education assistance. Speech recognition is essentially classified as a voice-dependent as well as speaker-independent for multiple speakers. This research emphases on isolated word for speech recognition and language identification systems for multiple speakers. Recognition of speech has made considerable progress in recent years, but work is still needed in multilingual language and speech models. Recognition of speech has made considerable progress in recent decades, but work is still needed in multilingual speech and language recognition as more than half of the world's population is multilingual. For more than five decades, research and development on methods and techniques for speech recognition and speaker recognition has been pursued and it continues to be a vigorous domain [1]. Human brain is inspired by the connectionist approach. Complex problem solves efficiently due to a high degree of parallelism. Although many researchers have done speech recognition work, this field is not much explored for Hindi and other Indian languages. Precise recognition of speech recognition program. A Hindi language isolated word speech recognizer is implemented in paper [2]. Features are extracted using LPC and HMM is used to identify them. Neural network and HMM for Arabic isolated word recognition are used for speech recognition in paper [3]. Training of neural network is done with algorithm Al-Aloui and tests are related with HMM. On the other hand, recognition of sentence is achieved using speech segmentation approach Multilingual speech recognition is proposed in paper [8] in the mobile application. The results of different recognizers were combined and performance comparison is done with a single multilingual system. A proactive procedure is utilized in paper [7] to train a feed forward MLP and to simulate the use of this isolated word recognition method. Using incremental learning method in the proposed system and experiment with 10 isolated terms. MFCC techniques are used in the extraction process of functionality. Multilingual speaker recognition system research has been conducted using ANN and Hidden Markov Model (HMMs), Harmonic Product Spectrum (HPS) [4, 5]. Stage of identification different methods are used to classify the speakers. Pattern used from signal processing to storing features in codebooks as a vector quantization technique [5, 6]. Classification is done using neural network with back propagation algorithm. The paper is structured as follows. Section 2 describes how multilingual speech and language recognition systems can be created. Section 3 explains the findings and the experiment. This work was summarized in Section 4 and explored future context.

2. Methodology

Speech recognition inputs are words or they can be sentences. There are two essential stages i.e. training and testing in the recognition system. A recognition system will be trained using training data in the training phase and then prepared modal will be evaluated in the test phase. This work utilizes MFCCs and delta-MFCCs to convert speech signals into sequences of acoustic vectors. So, the frequency is mapped on Mel-scale. Two filters namely linear filters and logarithmic filters, are used in the MFCC technique. The Mel-scale that has been mapped below 1000hz frequency utilizes linearl filters and above 1000hz utilizes logarithmic filters. Steps used to remove MFCCs and delta MFCCs as shown in Figure 1. In order to find out MFCC coefficients, the front-end phase is performed on speech signal after speech acquisition. The speech signal is re-sampled because at high frequencies and at low frequencies

some speech signal may be present. So, it is done to re-sample all speech signals at one stage. Then the re-sampled signal utilizes pre-emphasis filter which emphasizes complex frequencies.

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

Here 'a' is typically selected to be less than 1. Since speech is a quasi-stationary signal, extraction should be performed on stationary signal to minimize error in the recognition frame function. Therefore, speech is separated into 25ms (256 samples) short frames and 10ms frame shift. To maintain continuity between frames, the overlapped frame is used. Overlapping guarantees a high correlation of consecutive frame coefficients. Usually referred to as a window is the waveform segment used to evaluate that vector parameter. Here we used window hamming, multiplied by each frame.

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

Here the input signal is X (n) and the window is W (n). The hamming window formula is given as

$$w(n) = 0.54 - 0.46\cos(2n\pi\,|N-1)\,, \quad 0 \le n \le N-1$$

The relation between Mel and linear frequency is shown below,

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

Then it performs DCT to translate signal into time domain. The delta of MFCCs is determined using formula:

$$\Delta C_i(n) = \frac{\sum_{-N}^{N} kC_i(n+k)}{\sum_{-N}^{N} k^2}$$

N=2 is usually taken.

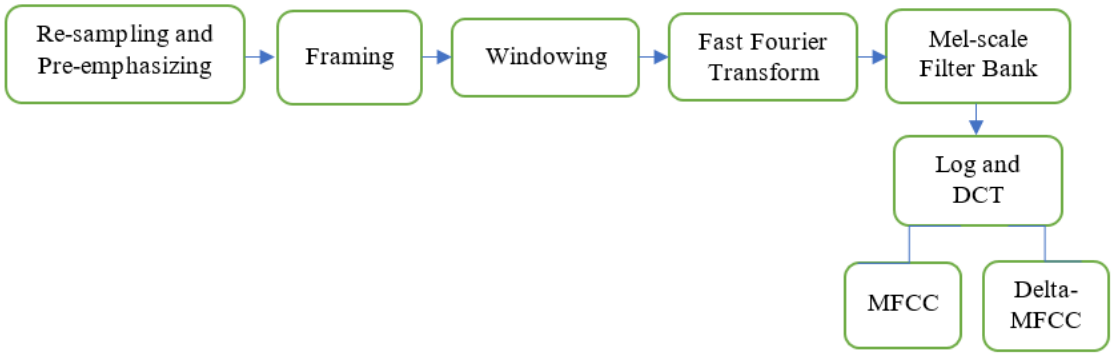

Fig 1: MFCC and Delta-MFCC Feature Extraction

2.1 Classification

Classifier identifies utterances on the basis of the speech signal's acoustic properties. The design of a speech recognition program, training and testing process includes two major steps. Only feature vectors are trained in the classifier of learning. Feed forward back propagation neural network and radial-based neural network function are used as a classifier. The following subsections describe both models.

2.1.1 Back propagation network:

Feed forward NN is trained using the robust back propagation algorithm at the classifier stage. ANN can identify unknown patterns because it studies behaviour patterns. Extracted goal matrix from the feature set is used to train NN as supervised learning is used for training. Single hidden layer is used in ANN and hidden layer neurons depends on different factors, i.e. number of input neurons, output layer and quantity of training samples [10]. Here n number of output layer neurons representing the number of words or languages to be identified. Number of MFCC neurons and delta MFCC coefficients derived from each frame make up the input layer.

2.1.2 Radial Basis Function Network

Radian base neural network feature is a static neural network feed forward with two single layers of hidden layer and one output layer. Gaussian or other kernel-based activation functions in the hidden unit of the RBF network. RBF neural network has been trained using back propagation algorithm much faster than multi-layer perceptron. With non-stationary input, they are less susceptible because speech is also a non-stationary type of signal, so that RBF is more suitable for speech recognition.

3. Experiment & Results

The database consists of 10 speakers speaking signals. All speakers speak the sentence "AB ISS BAAR TUM JAO" in four Hindi, English, Sanskrit and Telugu languages. For each of Hindi and English, the maximum number of words is 18, 5 and 4 for each of Telugu and Sanskrit. So, the total number of utterances is 180 for this test. MFCCs and delta-MFCCs are extracted using MATLAB's DSP toolbox and a total of 22 functions are measured per frame. Back propagation algorithm and radial base function performance is determined. The log sigmoid function is used in the hidden layer and the linear transfer function is used in the output layer of the Classifier based on the back-propagation algorithm. For speech recognition and language recognition, two different neural networks are developed. Here the feature set for each ANN is the same as the input, but the target matrix for ANNs is different. Hence, there are 18 words in total so that the output layer of speech recognizer consists of 18 neurons and the output layer of language recognizer consists of 4 neurons. Here the overall performance of speech recognition is 83.89 percent and the

performance of language recognizer is 83.3 percent. Radial basis function network is used as a classifier in the second experiment. The overall rate of word recognition obtained using this model is 91.7 and the rate of language recognition is 91.1%.

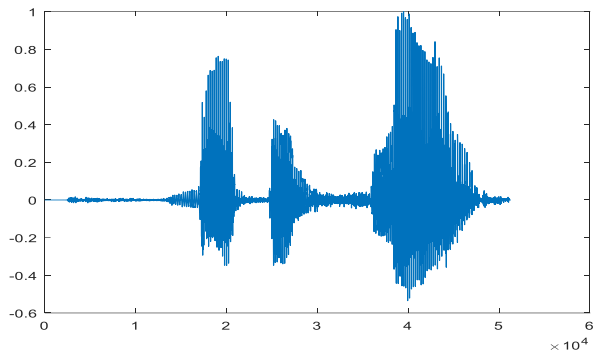

Fig 2: Input speech signal (English)

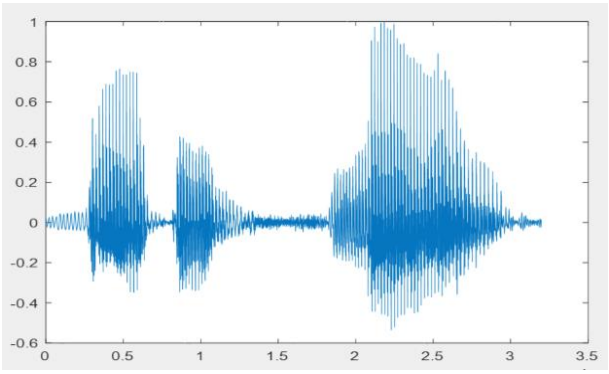

Fig 3: Input speech signal (silence removed)

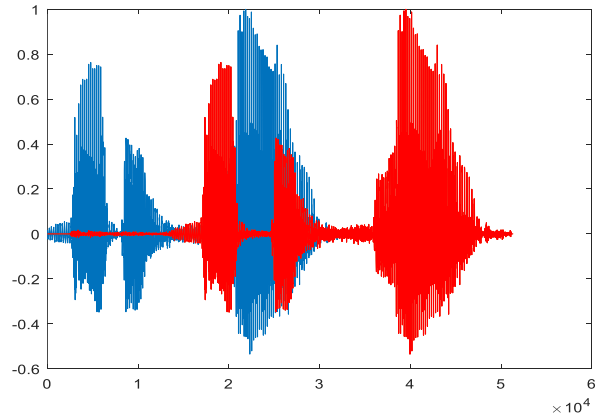

Fig 4: Input Waveform with normal data and silence removed signal

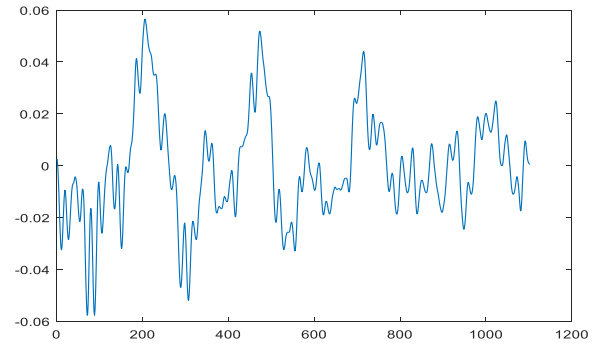

Fig 5: Framing of input speech signal

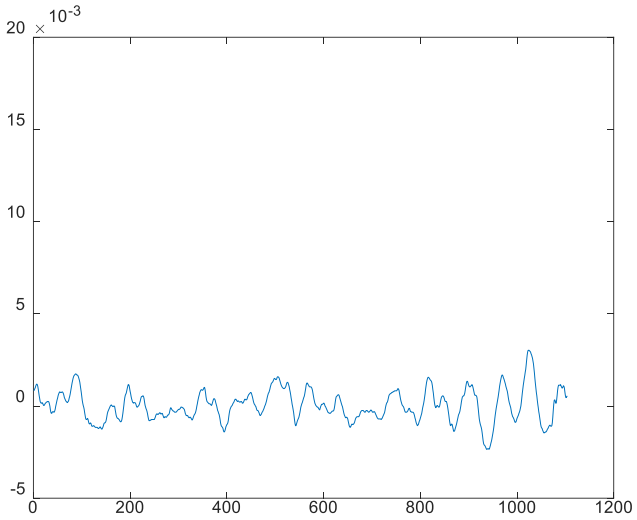

Fig 6: Frames after windowing

Table 1: Accuracy obtained for Speech and Language Identification

Speech Domain	Feature Extraction Approach	Classifier	Accuracy
Speech Recognition	MFCC	ANN	83.89%
Language Identification	MFCC	ANN	83.3%
Speech Recognition	MFCC	Radial Basis Function	91.7%
Language Identification	MFCC	Radial Basis Function	91.1%

4. Conclusion

This work is based on multilingual voice recognition and language recognition system using ANN. Two separate ANNs are developed and trained using back propagation algorithm and radial-basis function network for language and speech recognition system. Here we used the acoustic characteristics of MFCCs and delta-MFCCs. The present work is limited to four languages and a minimal vocabulary. We will be developing a system for broad vocabulary and more languages in the future.

5. References

1. Besacier L, Barnard E, Karpov A, Schultz T. Automatic Speech Recognition for Under Resourced Languages: A Survey, Speech Communication (Elsevier). 2014; 56:85-100.
2. Echeverry-Correa J, Ferreiros-Lopez J, Coucheiro-Limeres A, Cordoba R, Montero J. Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition, Expert system with applications. 2015; 42:101-112.
3. Weiqiang Z, Jia H Tao, Liu. Discriminative Score Fusion for Language Identification, Chinese Journal of Electronics. 2010; 19:124-128.
4. Li Haizhou, Bin M, Chin-Hui. A vector space modeling approach to spoken language identification, IEEE Transactions on Audio, Speech, and Language Processing. 2007; 15(1):271-84.
5. Segbroeck V, Travadi MR, Narayanan SS. Rapid language identification, IEEE Transactions on Audio, Speech, and Language Processing. 2015; 7:1118-1129.
6. Lopez-Moreno Ignacio J, Gonzalez-Dominguez D, Martinez O Plchot, Gonzalez-Rodriguez J, Moreno PJ. On the use of deep feed forward neural networks for automatic language identification, Computer Speech &

Language, 2016, 40.
7. Sim Chai K, H Li. On acoustic diversification front-end for spoken language identification, IEEE transactions on audio, speech, and language processing. 2008; 16(5):1029-1037.
8. Sun Y, Wen G, Wang J. Weighted spectral features based on local Hu moments for speech emotion recognition," Journal of Biomedical Signal Processing and Control. 2015; 18:80-90.
9. Yapanel U, Hansen J. A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," Journal of Speech Communication. 2008; 50:142-152.
10. Olvera MM, Sánchez A, Escobar LH. Web-Based Automatic Language Identification System, International Journal of Information and Electronics Engineering. 2016; 6(5):304-307.