International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 7, July 2017)

Automatic Recognition of Spoken Words of Marathi and Hindi Language through Machine using Ensemble Feature Extraction and Neural Network

Agnihotri P. P.[1], Khanale P. B.[2]

[1]*School of Technology, S.R.T.M. University, Nanded Sub-Centre, Latur, India*
[2]*Dnyanopasak College, Parbhani, India*

Abstract: **Speech recognition is a process of converting speech into a sequence of words through a computer system. This process enables the people to use speech as one more input mode to interact with computer system effectively. Speech recognition interface in native language will enable the people to use the technology effectively without knowing the standard input methods. Only a few researchers have worked on Marathi and other Indian languages.**

In this paper, Speech recognition system for spoken Marathi words which are based on ensemble feature extraction techniques and neural network are proposed. It was observed that ensemble feature extraction techniques are more accurate than individual one and neural network is effective tool to classify the pattern.

Keywords: **Automatic Speech Recognition, Filtering Techniques, MFCC, PLP, RASTA, PCA, Neural Network**

I. INTRODUCTION

Automatic Recognition of Voice commands by the machine is a need over a long period of time. In India the language of the people changes over every 300 Kms. People often need to use machines for controlling applications through commands, data entry and its storage, document preparation, data analysis, information retrieval, entertainment etc.

In India there are 122 major languages and 1599 other language spoken by the people [1]. In Indian census 2001, 73 millions people reported Marathi to be their native language. One common feature is that all Indian languages are phonetic in nature[2]. The script used for writing Marathi is called Devanagari derived from the Brahmi script of Ashoka. Devanagari was originally developed to write Sanskrit but was later adapted to write many other languages, including Marathi. Devanagari has 65 consonants, 18 full vowel letters, 17 vowels symbols and 2 symbols for nasal sounds. In Marathi language only 36 consonantal syllables, 13 vowels and 2 symbols for nasal sounds are used. The main feature of the script is left-to-right and words are spelled phonetically [3].

This paper is divided into six phases: Database preparation and collection, Feature extraction, pattern recognition, Proposed Algorithm, Result and Conclusion.

1. *Database Preparation and Collection:* In this study primary and secondary both sources were used. In primary data source Speech is recorded using PRAAT which is freeware software giving the facility to record, store and study the various properties of Speech samples. Speech samples (52 words per speaker for both languages) were recorded at 16kHz sampling frequency using laptop and unidirectional condenser microphone. These three speakers belong to different age groups and having different socio linguistic background. So total 156 samples were prepared for the study.

In secondary source fourty speech samples were collected from CIIL (Central Institute of Indian Languages) in different forms like most common words, most common names and sentences .

2. *Feature extraction:* In this study, four feature extraction techniques were used, that are Mel Frequency Cepstral Coefficient (MFCC), Linear predictive coding LPC), Perceptual Linear Prediction (PLP) and Relative Spectra Processing (RASTA).

i. *Mel Frequency Cepstral Coefficient (MFCC):* These features are commonly used in automatic speech and speaker recognition.

Mel scale frequency

The Mel scale is related with perceived frequency, or pitch, of a pure tone to its actual measured frequency. Mel scale frequencies are used to match more closely what humans hear[4].

In human speech production system, sounds generated by a human are filtered by the shape of vocal tract including tongue, teeth etc. The shape of the vocal tract is used for determination of envelope of the short time power spectrum. The Mel Frequency Cepstral Coefficient (MFCC) is to accurately represent such kind of envelope [5].

The equation for converting linear frequency to Mel scale is [6]:

$$M(f) = 1125\ln\left(1+\frac{f}{700}\right) \qquad (1)$$

Following equation is used to convert mel scale frequency to linear scale frequency:

$$M^{-1}(m) = 700(exp\left(\frac{m}{1125}\right) - 1 \quad (2)$$

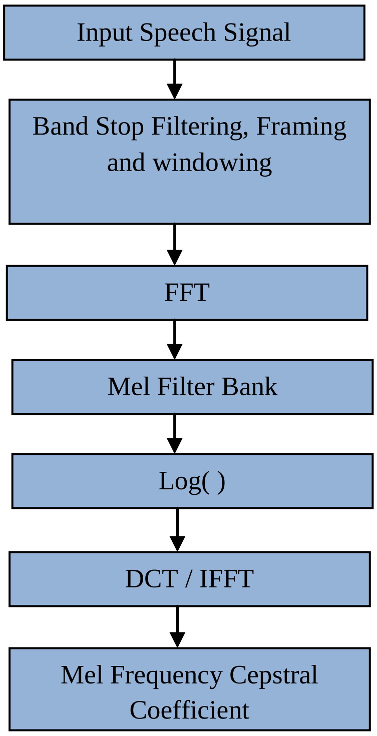
Input Speech Signal
Band Stop Filtering, Framing and windowing
FFT
Mel Filter Bank
Log( )
DCT / IFFT
Mel Frequency Cepstral Coefficient

Figure 1: MFCC Computation

MFCC procedure

Steps of determination of Mel Frequency Cepstral Coefficient (MFCC)

Step 1: Band Stop Filtering: As speech is recorded using microphone and laptop, the noise is added into the speech signal. This noise is removed using band stop filter which is the best filter as compared to other filters [6].

Step 2: Frame blocking: The speech signal is recorded at 16 kHz sampling frequency. The frame length for a 16kHz signal is 0.025*16000 = 400 samples, Single frame length is usually something like 10ms (160 samples), which allows some overlapping to the frames.

Step 3: Windowing: once the signal is split up into frames, window function is multiplied with each frame. The window function used in speech processing is the Hamming window which is as follows

$$w(n) = \begin{cases} 0.54 - 0.46.\cos\left(\frac{2\pi n}{N}\right), 0 \le n \le N-1 \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases} \qquad (3)$$

Step 4: Discrete Fourier Transform: it has been applied on each frame of the speech signal to convert the signal from time domain to frequency domain. It also generates the power spectrum of the signal. The Discrete Fourier Transform equation is as follows

$$S_i(k) = \sum_{n=1}^{N} S_i(n)h(n)e^{-2\pi kn/N} \quad 1 \le k \le K \quad (4)$$

where h(n) is an N sample long analysis window (e.g. hamming window), and K is the length of the DFT. The periodogram-based power spectral estimate for the speech frame $S_i(n)$ is given by:

$$P_i(k) = \frac{1}{N}\,|S_i(k)|^2 \qquad (5)$$

This is called the Periodogram estimate of the power spectrum. It takes the absolute value of the complex fourier transform, and square the result.

Step 5: Compute the Mel-spaced filter bank: The filter bank is prepared using following steps

i. Convert the linear frequencies into mel scale frequencies using equation 1

ii. Find out the central frequencies of all adjacent mel frequencies

iii. The first filter bank will start at the first point; reach its peak at the second point, then return to zero at the 3rd point. The second filter bank will start at the 2nd point, reach its max at the 3rd, then be zero at the 4th etc. Repeat this step up to last mel scale frequency to form a triangular shape filter bank as shown in the figure.

Step 6: Take the log of each of the 26 energies.

Step 7: Take the Discrete Cosine Transform (DCT) of the 26 log filterbank energies to give 26 cepstral coefficents.

For ASR, only the lower 12-13 of the 26 coefficients are kept. The resulting features (12 numbers for each frame) are called Mel Frequency Cepstral Coefficients.

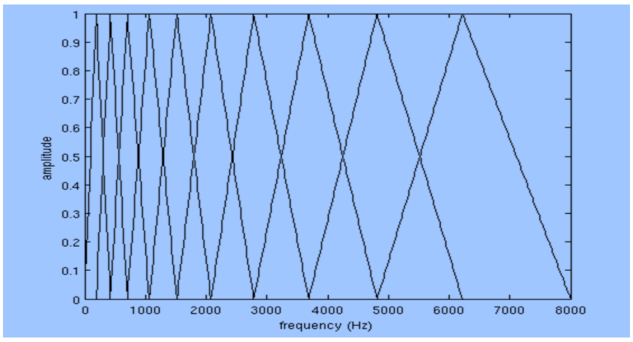

Figure 2: Triangular shaped filter bank

ii. *Perceptual Linear Prediction:* In PLP models the speech is based on psychophysics hearing. PLP and LPC is similar except that its spectral characteristics have been transformed to match human auditory system. PLP parameters have robust features when they are used for speech recognition [7].

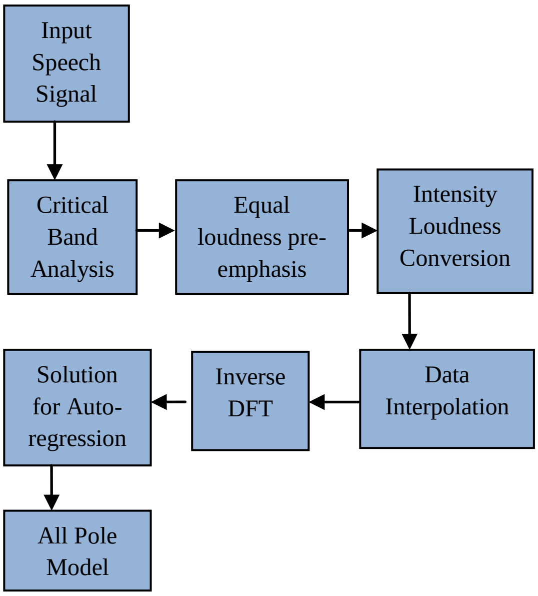
Input Speech Signal
Critical Band Analysis
Equal loudness pre-emphasis
Intensity Loudness Conversion
Data Interpolation
Inverse DFT
Solution for Auto-regression
All Pole Model

Figure 2: PLP computation [8]

First of all Power Spectrum Density (PSD) is computed for each frame of the signal. A measure of power intensity of input signal in frequency domain is called as Power Spectrum Density. FFT is used to determine PSD[8].

Perceptual Linear Prediction procedure:

*Step 1:*Input the speech signal
*Step 2:*Do the framing of the signal and apply hamming window function on it.
*Step 3:*Take Discrete Fourier Transform of signal to generate the power spectrum
*Step 4:*Take logarithm of the signal
*Step 5:*Pre-emphasize the signal by simulating the equal loudness curve to match the frequency magnitude response of the human auditory system.
*Step 6:*Convert the intensity loudness or set the intensity loudness as required.
*Step 7:*Do the inverse of Discrete Fourier Transform
*Step 8:*Apply regression using Durbin's method
*Step 9:*Perform cepstral recursion to determine the PLP Coefficients

Relative Spectra (RASTA):

RASTA is a feature extraction method which is designed to reduce or suppress the convolutional or additive noise.

The Frequency and Temporal masking as a psychoacoustic phenomena has been used in RASTA. Human ear trying to perceive the speech has some intensity called as maskee.

A speech component, which is adjacent to the frequency of another spectral component, drawing out the presence of maskee is called as masker. If determined intensity level of maskee is not available in presence of masker then that intensity level is called as masking threshold of maskee.

In masking, spectral component of one frame is masked by a stronger spectral component of another nearest frame. The similar kind of masking is done for all adjacent or closest frames. In Frequency Domain the masking is done on the basis of critical bands [7,8, 9].

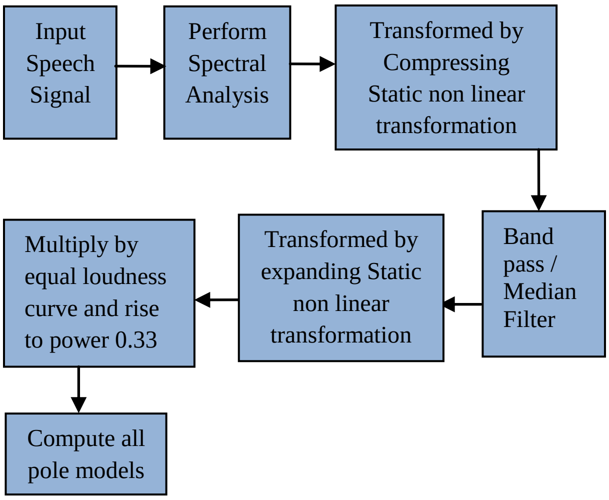
Input Speech Signal
Perform Spectral Analysis
Transformed by Compressing Static non linear transformation
Band pass / Median Filter
Transformed by expanding Static non linear transformation
Multiply by equal loudness curve and rise to power 0.33
Compute all pole models

Figure 4: RASTA computation

RASTA Procedure

Step 1: Input the speech signal
Step 2: Do the framing of the signal and apply hamming window function on it.
Step 3: Take Discrete Fourier Transform of signal to generate the power spectrum
Step 4: Take logarithm of the signal
Step 5: Pre-emphasize the signal by simulating the equal loudness curve to match the frequency magnitude response of the human auditory system.
Step 6: Multiply power spectrum with 0.33 to simulate power law of hearing
Step 7: Do the inverse of Discrete Fourier Transform
Step 8: Apply regression using Durbin's method
Step 9: Perform cepstral recursion to determine the RASTA Coefficients

II. Proposed Algorithm

The algorithm is designed for recognizing Marathi and Hindi spoken words using ensemble feature extraction techniques. The proposed algorithm is as follows:

Step 1: Input the speech signal stored in wav files

Step 2: Filter the signal using band stop filtering technique.

Step 3: Determine the Mel frequency cepstral coefficient (MFCC) with Mel frequency using MFCC procedure

Step 4: Determine the Perceptual Linear Prediction (PLP) coefficient using PLP procedure.

Step 5: Determine the RASTA Coefficient using RASTA procedure

Step 6: Principal Component Analysis is used to reduce the dimensions of feature vector.

Step 7: Combine the first final input vector obtained using step 3 and step 4 is: F1 = [M1 P1]

Step8: Combine the second final input vector obtained using step 3 and step 5 is: F2 = [M1 R1]

Step 9: Combine the second final input vector obtained using step 3, 4 and step 5 is: F3 = [M1 R1]

Step 10: Create the target vector i.e. target which represents the number of classes

Step 11: Create Multi layer Feed-Forward Network net1, net2 and net3, Train all networks using final input vectors F1, F2, F3 and a target vector

Step 12: Simulate the network using test samples and target vector.

Step 13: Determine the accuracy of recognition of spoken words using following formula

Accuracy of recognition = No. of words correctly classified/ total number of words

III. Pattern recognition

Neural Networks is an effective tool to classify the acoustic patterns in speech recognition system. This tool is used in many applications such as isolated word speech recognition, speaker adaptation and speaker verification.. In MATLAB, Neural Network Toolbox provides various functions, algorithm and Apps in order to create, visualize, train and simulate neural network [10].

In this study, multi layer feed forward network is used as a classifier. Network has been established by setting the network parameters as follows

Inputs=13x156 feature vector
Target=52x156 identity matrix
Learning rate= 0.2
No. Of epochs=1000
No. Of hidden layers=2
Number of neurons=33 and 52
Initial weight = 0.1
Initial bias=0.1
Minimum gradient=1e^{-6}
Performance goal=0.3

Network is two-layered feed forward network with hidden and output neurons, having enough neurons in hidden layer to classify vectors well. It requires a target value for each input value and then output which are compared with the targets. Therefore, it is referred as supervised learning method.

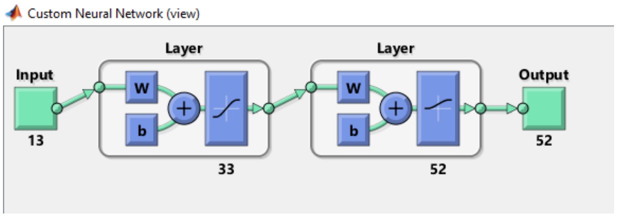

Figure 4: Multi layer Feed Forward Network

IV. Results

Ensemble feature extraction techniques are used in automatic speech recognition system in order to improve performance and accuracy over basic approach. The developed Multi layer Feed Forward Neural Network with combination of features of test sample is simulated and result is stored into output vector i.e. Ytest, Then correct classification of samples are determined on the basis of diagonal positions values of output vector (ytest).

Table I:
Comparison of various feature extraction techniques

S. No.	Feature Extraction technique	Average Accuracy (%)
1	Mel Frequency Cepstral Coefficient (MFCC)	93.58
2.	Perceptual Linear Prediction (PLP)	86.57
3.	Relative Spectra (RASTA)	89.10
4.	MFCC+PLP	94.87
5.	MFCC+RASTA	95.51
6.	MFCC+PLP+RASTA	**98.08%**

It is found that combination of MFCC+PLP+RASTA technique gives highest recognition accuracy of spoken Marathi and Hindi language words as compared to other individual as well as other ensemble techniques.

V. Conclusion

In this paper a method for Marathi and Hindi spoken words recognition based on combination of MFCC, PLP and RASTA features is proposed. The proposed method is implemented and patterns are classified using multi layer feed forward network and it has been observed that the Neural Network is effective tool to classify the acoustic patterns which provide better accuracies as compared to other conventional methods.

REFERENCES

[1] www.censusindia.gov.in/census_data_2001/census_data_online/ language /statement6.aspx accessed on Dec. 2013

[2] O.N. Koul, 2008, "Modern Hindi grammar", Dunwoody Press, USA

[3] Akshar Bharati, Vinit Chaitanya, Rajiv sangal, 2010, " Natural Language Processing : A Paninian perspective", PHI New Delhi

[4] Aldebaro Klautau,2011, " article on :The MFCC" accessed on Sep. 2015

[5] Sanjay A. Valaki, Harikrishna B. Jethva ,2016, "A Survey on Feature Extraction and Classification Techniques for Speech Recognition", IJARIIE-ISSN(O)-2395-4396, Vol-2 Issue-6, pp830-837

[6] Agnhotri P.P., Shinde A.R., Khanale P.B., 2016, "Development of Noise Free Marathi Speech database using various filtering techniques" Paripex-Indian Journal of Research" ISSN: 2250-1991, vol 5, issue 2, pp 23-25.

[7] Namrata Dave, 2013, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition" International Journal For Advance Research In Engineering And Technology, Volume 1, Issue 6, pp 1-5

[8] E.Chandra, K.Manikandan and M. Sivasankar, 2014, "A Proportional Study on Feature Extraction Method in Automatic Speech Recognition System", International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering Vol. 2, Issue 1,pp 772-775

[9] Shreya Narang, Ms. Divya Gupta,2015, "Speech Feature Extraction Techniques: A Review" , International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, Vol. 4, Issue. 3, pg.107 – 114

[10] Poonam Sharma, Anjali Garg, 2016, "Feature Extraction and Recognition of Hindi Spoken Words using Neural Networks", International Journal of Computer Applications (0975 – 8887), Volume 142 – No.7, pp 12-17