IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-3, Issue-7 E-ISSN: 2347-2693

Isolated Word Recognition System for Hindi Language

Suman K. Saksamudre[1*], R. R. Deshmukh[2]

Dept of computer Science & IT, Dr.Babasaheb Ambedakar Marathwada University, Auranbabad

Received: Jun/22/2015 Revised: July/01/2015 Accepted: July/17/2015 Published: July/30/ 2015

Abstract— Speech is a natural mode of communication for people. So people are so comfortable with speech recognition systems. The overall performance of any speech recognition system is highly depends on the feature extraction technique and classifier. In this paper, we presented Isolated Word Recognition System for Hindi Language using MFCC as feature extraction and KNN as pattern classification technique. The system is trained for 10 different Hindi words. The experimental result of our system is that it gives 89% accuracy rate.

Keywords— Pattern Recognition, Automatic Speech Recognition (ASR), DCT, FFT.

I. Introduction

Automatic recognition of speech by machine has been a goal of research for more than four decades. Speech recognition is the process by which computer maps an acoustic speech signal to some form of abstract meaning of the speech. This process is highly difficult [1] since sound has to be matched with stored sound bites on which further analysis has to be done because sound bites do not match with pre-existing sound pieces. Various feature extraction methods and pattern matching techniques are used to make better quality speech recognition systems. Feature extraction technique and pattern matching techniques plays important role in speech recognition system to maximize the rate of speech recognition of various persons.

There are two main phases in a speech recognition system [2]: training phase and testing phase for recognition. During the training phase, first off all features are extracted from the all speech signals using various feature extraction techniques such as MFCC, LPC, LDA and RASTA [3] etc. These features are in the form of vector. In this way a training vector is generated from the speech signal of each word spoken by the user. The training vector has the spectral features which distinguishes different words based on its class. These extracted features are the main component of whole speech recognition system. Each training vector can serve as a template for a single word or a word class. This training vector is used in the next phase of recognition. During the recognition phase, the user speaks any word for which the system was trained. A test pattern is generated for that word. That test pattern means extracted features of that word used for the testing. In this way the test pattern is tested against the training vector by using various classifier such as SVM, KNN, HMM and ANN [4] etc. This classifier classifies the pattern. If the testing word pattern matches with the training pattern class then it means that particular pattern is recognized from training phase and that corresponding pattern is displayed as the output. The system block diagram of speech recognition process is shown in Fig. 1.

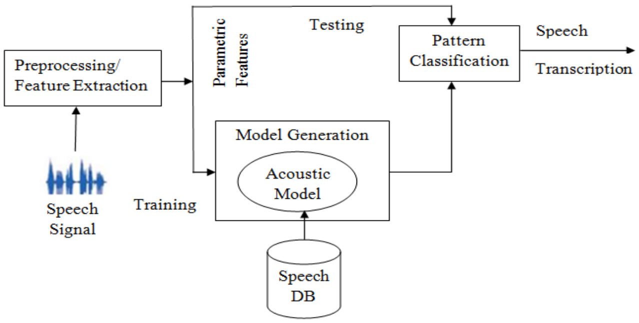

Fig 1: Working of Speech Recognition process

II. Literature Survey

Pruthi et al. [5] have developed a speaker-dependent, realtime, isolated word recognizer for Hindi. Developed system uses a standard implementation. Linear predictive cepstral coefficients are used for feature extraction and recognition is carried out using HMM. System was designed for two male speakers. The recognition vocabulary consists of Hindi digits (0, pronounced as “shoonya” to 9, pronounced as “nau”).

K. Kumar et al. [6] developed a small vocabulary, isolated Hindi speech recognition with high performance 94.63%. This system is speaker independent. For training, 5 male and 3 female speakers are used. Vocabulary size of system is 30 words. MFCC is used as a feature extraction technique and at back end HMM is used. HTK toolkit is used to develop this system.

Mishra et al. [7] in 2011 proposed a Hindi ASR system for connected digit recognition. To build this speaker independent system, 40 different speakers are used in which 23 are female and 17 are male speakers. All speakers are age group of 18-26 years. After speech recording some noises is added artificially. In this paper different feature extraction technique is used such as BFCC, RPLP, MFCC, PLP, MF-PLP in front end and HMM is used at back end. They receive high accuracy 99%, when MF-PLP is used as a feature extraction technique.

Shweta et al. [8] in 2013 proposed a speaker independent, continuous Hindi speech recognition system for with different vocabulary sizes. In this paper, Gaussian mixture HMM model with various states was used for training and recognition. In this paper, MFCC and perceptual linear prediction (PLP) with heteroscedastic discriminant analysis (HLDA) was used as a feature extraction technique. HTK and Sphinx was used to implement this system. Overall accuracy 93% was achieved with MFCC at front end and 8 states GMM (Gaussian mixture model) at beck end, when the vocabulary size was 600 words.

In 2011, R. k. Aggarwal et al. [9] proposed an ASR system, in which different state GMM was used to train the ASR system. This speaker independent system was built for 100 – 400 vocabulary size. The best performance of ASR was observed when 4 state GMM was used. This system gives 88% accuracy with 400 vocabulary size. To reduce the MFCC features, HLDA feature reduction algorithm was used in front end of the ASR system.

III. METHODOLOGY

A. *FEATURE EXTRACTION*

1) MFCC in speech recognition:

The mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency [10].

It is popular feature Extraction technique [11].Mel-frequency cepstral coefficients are the feature that collectively makes Mel-frequency cepstral (MFC) [12]. The difference between the cepstrum and the mel-frequency cepstrum is that in Mel-frequency cepstral (MFC), the frequency bands are equally spaced on the Mel scale, this mean that it approximates the human auditory system's response more firmly than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping allows better representation of sound [13].

The block diagram of MFCC as given in is shown in Fig 2:

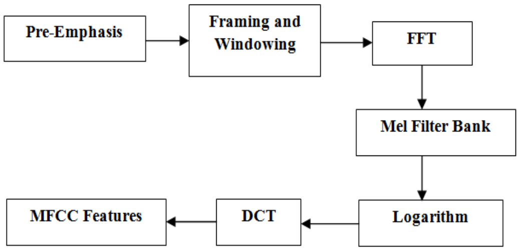

Fig 2: Block diagram of MFCC

As shown in Figure 2, MFCC consists of six computational steps. Each discussed briefly in the following:

a) Pre–emphasis

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. It can also add the importance of high-frequency formants.

The speech signal s (n) is sent to a high-pass filter [14]:

$$\mathbf{s_2\,(n) = s\,(n) - a*s\,(n-1)}$$

Where s_2 (n) is the output signal and the value of a is generally between 0.9 and 1.0 and z-transform of the filter is

$$\mathbf{H\,(z) = 1-a*z^{-1}}$$

b) Framing

An audio signal constantly changes, so to simplify this we assume that on short time scales the audio signal doesn't change much (signal doesn't change means statistically i.e. statistically stationary, samples changes constantly on even short time scales). This is why we have to frame the signal into 20-40ms frames. If the frame is much shorter we don't get enough samples to have a reliable spectral estimation and if it is longer the signal changes highly throughout the frame.

The input speech signal is segmented into frames of 20~30 ms with optional overlap of 1/3~1/2 of the frame size. Usually the frame size (in terms of sample points) is equal to power of two in order to facilitate the use of FFT [15]. If this is not the, we need to do zero padding to the nearest length of power of two. If the sample rate is 16000Hz and the frame size is 480 sample points, then the frame duration is 480/16000 = 0.03 sec = 20 ms. Additional, if the overlap is 160 points, then the frame rate is 16000/ (480-160) = 50 frames per second.

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The speech signal is divided into N samples of frames. An Adjacent frames are being separated by M (M<N).

Typical values used are M = 100 and N= 256.

c) Windowing

The important function of Windowing is to reduce the aliasing effect, when cut the long signal to a short-time signal in frequency domain [16].The most widely used window is Hamming window. Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. Each frame has to be multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame (to be detailed in the next step). If the signal of frame is denoted by s (n), n = 0…N-1, then the signal after Hamming windowing is s (n)*w (n), where w (n) is the Hamming window defined by:

$$W\,(n, a) = (1 - a) - a \cos (2pn/ (N-1))\ ,\quad 0 \leqq n \leqq N-1$$

d) Fast Fourier Transform

Spectral analysis shows that different accents in speech signals correspond to different energy distribution over frequencies. Therefore we perform FFT to obtain the magnitude frequency response of each frame.

e) Mel Filter Bank

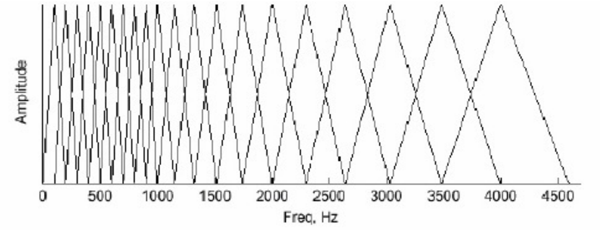

Fig 3: an example of mel-spaced filter bank

Triangular Bandpass Filters are used because the frequency range in FFT spectrum is very wide and voice signal does not follow the linear scale. We multiple the magnitude frequency response by a set of 20 triangular bandpass filters to get the log energy of each triangular bandpass filter. The position of these filters is equally spaced according to the Mel frequency, which is related to the linear frequency f by the following equation [17]:

Mel (f) =1125*ln (1+f/700)

The Mel-frequency is proportional to the logarithm of the linear frequency, reflecting similar effects in the human's subjective aural perception.

f) Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT) [18]. In this step, we apply DCT on the 20 log energy E_k obtained from the triangular bandpass filters to have L mel-scale cepstral coefficients. Formula for DCT is as shown in below:

$$C_m=S_{k=1}^{N}\cos [m*(k-0.5)*p/N]*E_k,\ m=1, 2... L$$

Where N is the number of triangular bandpass filters and L is the number of mel-scale cepstral coefficients. Generally we set N=20 and L=12.

g) MFCC Features

In this way Mel-frequency cepstral coefficients are extracted from the speech signal. These features are the main component of speech recognition process. Further classification of these features is done by the various types of Classifier.

B. KNN CLASSIFIER

KNN is Instance-based classifier [19]. Instance-based classifier performs on the premises that classification of unknown instances can be done by relating the unknown to the known according to some distance/similarity function. Opinion is that two instances far apart in the instance space defined by the appropriate distance function are less likely than two closely situated instances to belong to the same class.

1) Classification

Classification (generalization) using an instance-based classifier can be a simple matter of locating the nearest neighbor in instance space and labeling the unknown instance with the same class label as that of the located (known) neighbor. This approach is often referred to as a nearest neighbor classifier. The weakness of this simple approach is the lack of robustness that characterizes the resulting classifiers. The high amount of local sensitivity makes nearest neighbor classifiers highly susceptible to noise in the training data [20].

More robust models can be achieved by locating k, where k > 1, neighbor and letting the majority vote decide the outcome of the class labeling. A higher value of k results in a smoother, less local sensitive, function. Nearest neighbor classifiers can be considered as a special case of the more general k-nearest neighbor's classifier so referred as a KNN classifier. The disadvantage of increasing the value of k is that as k approaches n and n is the size of the instance base. The performance of the classifier can come up to the most straightforward statistical baseline, the conclusion that all unknown instances belong to the class most frequently represented in the training data.

Example of KNN:

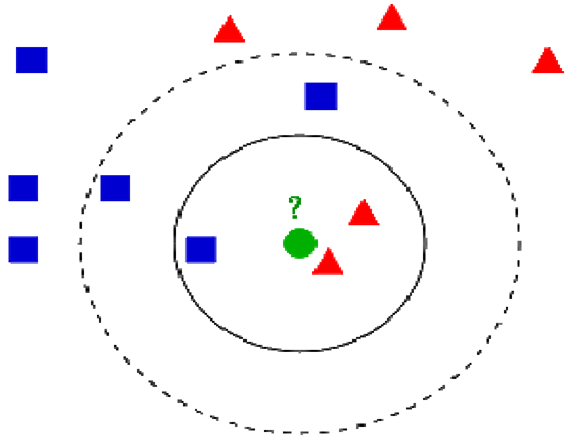

Figure 3.6: Example of KNN

In above diagram the test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If k = 3 (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If k = 5 (dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).

IV. DATABASE PREPARATION

For the development of this system, we first prepare the database .The database has ten agricultural related hindi isolated words such as chana, masur, mung, rajma, Arhar, genhu, chaval, ganna, ragi and til. These words were recorded by praat software [21] from ten Hindi speakers at 16KHZ.Recording were done in room environment. Each word is uttered 3 times. Out of these 2 utterances of every word were selected as training data and 1 utterance selected as testing data for such speaker dependent system. Hence our training data was of 200 wav files and test data of 100 wav files. All recording is done by Sennheiser PC360 Headset. The PC360 headset has noise cancellation facility and the signal to noise ratio (SNR) is less. Recorded speech files were in .wav file. Total 300 wave files of speech were there in database.

V. EXPERIMENT RESULT

MFCC features were extracted from the all 300 wav files.12th order MFCC features were used for training and test data. Both training and testing data features were applied on KNN classifier. As there were 10 words 10 classes were created. According to classes respective wave file had been classified by the KNN. After classification we got overall rate of classification as 89%. Rate of classification of each word tells how much percent that word had been matched with its respective class. Rate of classification of all ten words is shown on the following graph:

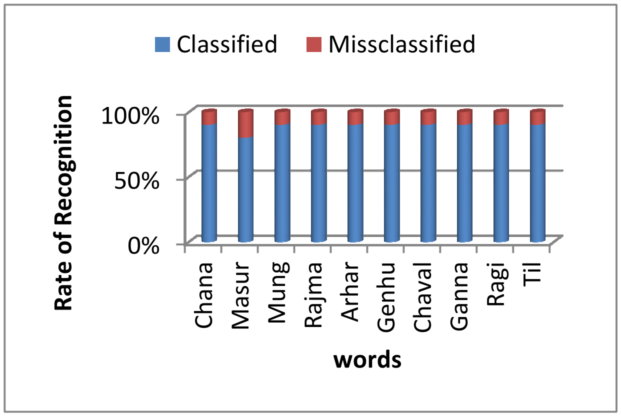

Fig 4: Graphical representation of classification of 10 Hindi words

VI. CONCLUSION AND FUTURE WORK

In this paper, we have briefly discussed MFCC as feature extraction technique. Using MFCC and KNN we developed our Isolated Word Recognition System for Hindi Language. MFCC and KNN have given us 89% of recognition rate for 300 vocabulary data. Further ANN classifier can be used.

ACKNOWLEDGMENT

This work is supported by University Grants Commission under the scheme Major Research Project entitled as "Development of Database and Automatic Recognition System for Continuous Marathi Spoken Language for agriculture purpose in Marathwada Region". The authors would also like to thank the University Authorities for providing the infrastructure to carry out the research.

REFERENCES

[1] Hemakumar, Punitha, “Speech Recognition Technology: A Survey on Indian languages”, International Journal of Information Science and Intelligent System, Vol. 2, No.4, **2013**.

[2] Santosh V. Chapaneri , “Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping”, International Journal of Computer Applications, Vol. 40– No.3, February **2012**.

[3] M. A. Anusuya, S.K. Katti, “Speech Recognition by Machine: A Review”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, **2009**.

[4] Abhishek Thakur, Naveen Kumar, “Automatic Speech Recognition System for Hindi Utterance with Regional Indian Accents: A Review”, International Journal of Electronics & Communication Technology, Vol. 4, April – June **2013**.

[5] Pruthi, T., Saksena, S. and Das, P. K. **2000**. Swaranjali:Isolated Word Recognition for Hindi Language using VQ and HMM,” Paper Presented at International Conference on Multimedia Processing and Systems(ICMPS), IIT Madras, India.

[6] Kumar, K. and Aggarwal, R. K. **2011**. Hindi Speech Recognition System using HTK, International Journal of Computing and Business Research, vol. 2, issue 2.

[7] Mishra, A. N. et al., **2012**. Robust Features for Connected Hindi digits Recognition, Int. Journal of Signal Processing, Image Processing and pattern Recognition, Vol. 4, No. 2.

[8] Sinha, S, Agrawal, S. S. and Jain, A. **2013**. Continuous density Hidden Morkov Model for context dependent Hindi speech recognition, Int. Conference on Advances in Computing, Communication and Informatics (ICACCI), pp. 1953-1958, IEEE.

[9] Aggarwal, R. K. and Dave, M. **2011**. Using Gaussian mixture for Hindi Speech Recognition System, International Journal of Signal Processing, Image Processing and pattern Recognition, SERSC Korea, vol.4, no. 4.

[10] Louis-Marie Aubert, Roger Woods, Scott Fischaber, and Richard Veitch “Optimization of Weighted Finite State

Transducer for Speech Recognition", IEEE Transactions on Computers, Vol. 62, No. 8, August **2013**.

[11] Ankit Kumar, Mohit Dua, Tripti Choudhary, "Continuous Hindi Speech Recognition Using Monophone based Acoustic Modeling"**,** International Journal of Computer Applications **2014**.

[12] S B Harisha , S Amarappa , Dr. S V Sathyanarayana, "Automatic Speech Recognition - A Literature Survey on Indian languages and Ground Work for Isolated Kannada Digit Recognition using MFCC and ANN", International Journal of Electronics and Computer Science Engineering.

[13] Borde, Prashant, Amarsinh Varpe, Ramesh Manza, and Pravin Yannawar. "Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition." International Journal of Speech Technology.**2015**.

[14] Anand Vardhan Bhalla, Shailesh Khaparkar "Performance Improvement of Speaker Recognition System", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 2, March **2012**.

[15] Munish Bhatia1, Navpreet Singh2, Amitpal Singh, "Speaker Accent Recognition by MFCC Using KNearest Neighbour Algorithm: A Different Approach", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 1, January **2015**

[16] Mel Frequency Cepstral Coefficient (MFCC) tutorial, Accessed 24 June **2015**.

[17] Rajesh Kumar Aggarwal, "Improving Hindi Speech Recognition Using Filter Bank Optimization and Acoustic Model Refinement"PHD Thesis, **2012**.

[18] M. Kalamani, Dr. S. Valarmathy, S. Anitha , "Automatic Speech Recognition using ELM and KNN Classifiers", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 4, April **2015**.

[19] Tsang-Long Pao, Wen-Yuan Liao and Yu-Te Chen, "A Weighted Discrete KNN Method for Mandarin Speech and Emotion Recognition", International Journal of Innovative Computing, Information and Control ICIC International Volume 6, February **2010**.

[20] http://www.fon.hum.uva.nl/praat/manual/kNN_classifiers_1__What_is_a_kNN_classifier_.html, accessed 25 June **2015**.

[21] Tsang-Long Pao, Wen-Yuan Liao and Yu-Te Chen, "A Weighted Discrete KNN Method for Mandarin Speech and Emotion Recognition", www.intechopen.com.