Speech Recognition System with Speaker Verification using HMM, LPC & MFCC

Rupinder Kaur, Pushpalata S

Abstract— Speech recognition system has been under research since a long time. Many commercial grade speech recognition system are now available across devices. Some of the popular SDKs includes Microsoft Speech SDK, Dragon Core Speech Assistant, SIRI, Google Speech recognition and so on. Speaker verification on the other hand is associated with Biometric system. In the past these two closely related domains are studied differently due to different applications and different methodology. In this work we propose a novel framework for speech recognition system by combining ANN and HMM classifiers through Wavelet energy and LPC features. We analyze the independent accuracy of the systems and combined accuracy. Then we extend the system for real time digit recognition system. This is further used for speaker verification system. Our results shows that same set of features and classifier can be used for both speech recognition and speaker verification system.

***Index Terms*— ANN, HMM, MFCC, Speech, Voice, Wavelets.**

I. INTRODUCTION

Speech processing is one of most important branches in digital signal processing. Speech signals can be used for speech recognition, speaker recognition or voice command recognition systems. The task of speaker identification is to determine the identity of a speaker by machine. To recognize voice, the voices must be familiar in case of human beings as well as machines. The second component of speaker identification is testing, namely the task of comparing an unidentified utterance to the training data and making the identification. Depending upon the application the area of speaker recognition is divided into two parts. One is identification and other is verification. In speaker identification there are two types, one is text dependent and another is text independent. Speaker identification is divided into two components: feature extraction and feature classification. In speaker identification the speaker can be identified by his voice, where in case of speaker verification the speaker is verified using database.

II. EXISITING SYSTEM

As all speech recognition techniques including commercially available ones like Nuance, Siri, Microsoft Speech SDK depends upon the clarity of spoken words and phrases for recognition accuracy, such systems cannot be adopted for cases or words which are not much clear. the major themes and advances made in the past fifty years of research so as to provide a technological perspective and an appreciation of the fundamental progress that has been accomplished in this important area of speech communication. Although many techniques have been developed, many challenges have yet to be overcome before we can achieve the ultimate goal of creating machines that can communicate naturally with people. Such a machine needs to be able to deliver a satisfactory performance under a broad range of operating conditions. A much greater understanding of the human speech process is required before automatic speech and speaker recognition systems can approach human performance. . Most of voice recognition systems contain two main modules as follow “feature extraction” and “feature matching”. In this project, MFCC algorithm is used to simulate feature extraction module. Using this algorithm, the cepstral coefficients are calculated on mel frequency scale. VQ (vector quantization) method will be used for reduction of amount of data to decrease computation time. Accuracy of system also increases if we increase number of repetitions for each command in training stage

III. PROPOSED SYSTEM

The aim of the work is to design and develop a roboust speech recognition technique that can recognize spoken words and digits with high accuracy in the absence of substantial training data and fluctuations in speech (for example the patients of dysarthria). As all speech recognition techniques including commercially available ones like Nuance , Siri, Microsoft Speech SDK depends upon the clarity of spoken words and phrases for recognition accuracy, such systems can not be adopted for case presented here. Therefore we develop this unique technique. We also extend the system to speaker verification system which can verify the user based on speech features.

The project is mainly divided into two stages and three modules. The first stage is the training stage where phrase is acquired from users and feature extraction is performed. Significant thing to be noted here that the system must support recognition of speech sample from shaky and changed voices. Which means that if the training samples are collected from user A, it should be applicable for user B.

Manuscript received June, 2015.

Rupinder Kaur, Computer Science and Engineering From Godutai College of Engineering For Womens, Gulbarga, Karnataka India, Mobile no: 9164582145.

Asst..prof. Pushpalata.S, *Computer Science And Engineering,* Working in Godutai Engineering College for Women, Gulbarga, Karnataka ,India .

The second stage is Recognition where the given test phrase should be classified against the training sample to recognize the spoken word.

We use Hidden Markov Model for the classification as it is based on probability and state transition calculation from one state of features to another state of features. As the training samples are low and test phrases are expected to be significantly different from that of training phrases, other classifiers like knn is not expected to produce good result as they mainly depend upon feature matching rather than prediction of probability in migration from one state to another.

For extraction of features we need speech samples. However speech samples are expected to contain environmental noise and silence period. Therefore we need to use a suitable preprocessing technique. We use low pass filter with hamming window to suppress the noise in the signal and remove the silence signal by thresholding the Fourier domain.

Features should represent the frequency model of the spoken word or phrase. As time domain analysis does not reveal too many characteristic of the signal, we aim to go with frequency domain features. The features are Mel Frequency Cepstrum. Mel frequency Cepstrum is essentially a filtered chunk of FFT signal represented with the signal power (logarithmic scale).

As HMM is based on state transition model and number of states should be fixed, we use feature normalization and windowing to obtain fixed length Mel Cepstrum from every speech sample, irrespective of the spoken duration.

IV. DESCRIPTION OF THE PROPOSED PROJECT

The work is mainly divided into two stages and three modules. The first stage is the training stage where phrase is acquired from users and feature extraction is performed. Significant thing to be noted here that the system must support recognition of speech sample from shaky and changed voices. Which means that if the training samples are collected from user A, it should be applicable for user B.

The second stage is Recognition where the given test phrase should be classified against the training sample to recognize the spoken word.

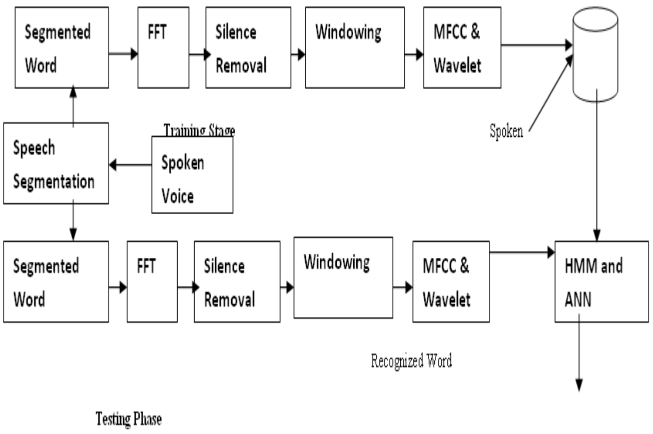
We use Hidden Markov Model for the classification as it is based on probability and state transition calculation from one state of features to another state of features. As the training samples are low and test phrases are expected to be significantly different from that of training phrases, other classifiers like knn is not expected to produce good result as they mainly depend upon feature matching rather than prediction of probability in migration from one state to another.

For extraction of features we need speech samples. However speech samples are expected to contain environmental noise and silence period. Therefore we need to use a suitable preprocessing technique. We use low pass filter with hamming window to suppress the noise in the signal and remove the silence signal by thresholding the Fourier domain.

Features should represent the frequency model of the spoken word or phrase. As time domain analysis does not reveal too many characteristic of the signal, we aim to go with frequency domain features. The features are Mel Frequency Cepstrum. Mel frequency Cepstrum is essentially a filtered chunk of FFT signal represented with the signal power (logarithmic scale).

Wavelet transform converts the signal in band of signal by separating low frequency signal from high frequency signal. We obtain the energy of each band which is a good representation of the Pitch of the bands. We then combine the features of LPC and wavelet to construct an independent feature set.

As HMM is based on state transition model and number of states should be fixed, we use feature normalization and windowing to obtain fixed length Mel Cepstrum combined with normalized Energy features from wavelet decomposition from every speech sample, irrespective of the spoken duration.



Recognized Word

Testing Phase

Figure 1: Overall block diagram of Speech Verification System

This same system can be further used for speaker verification system. In case of speaker verification process, the speaker needs a long phrase. In this case the independent words are not extracted. Rather preprocessing is applied directly on the raw speech data. The block diagram is shown in figure 2.

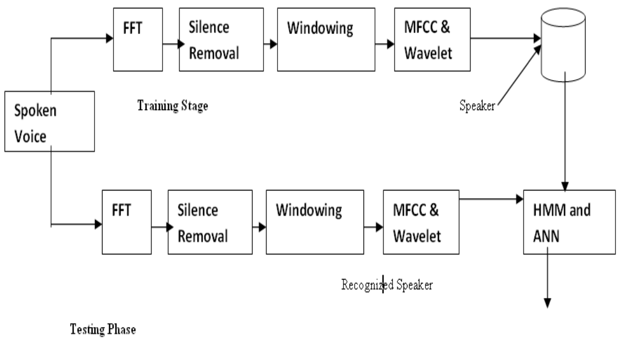


Figure 2: Speaker Verification System

A Fast Fourier transform (FFT)

It is an algorithm to compute the discrete Fourier transform (DFT) and its inverse. Fourier analysis converts time (or space) to frequency and vice versa; an FFT rapidly computes such transformations by factorizing the DFT matrix into a product of sparse (mostly zero) factors. As a result, fast Fourier transforms are widely used for many applications in engineering, science, and mathematics. The basic ideas were popularized in 1965, but some FFTs had been previously known as early as 1805. Fast Fourier transforms have been described as "the most important numerical algorithm [s] of our lifetime".

Silence removal

This is a simple method for silence removal and segmentation of audio streams that contain speech. The method is based in two simple audio features (signal energy and spectral centroid). As long as the feature sequences are extracted, as thresholding approach is applied on those sequence, in order to detect the speech segment.

Windowing

Windowing is a technique used to shape the time portion of your measurement data, to minimize edge effects that result in spectral leakage in the FFT spectrum. By using Window Functions correctly, the spectral resolution of your frequency-domain result will increase.

MFCC (Mel Frequency Cepstrum Coefficient)

In this project we are using the Mel Frequency Cepstral Coefficients (MFCC) technique to extract features from the speech signal and compare the unknown speaker with the exits speaker in the database. Figure 7 shows the complete pipeline of Mel Frequency Cepstral Coefficients

The Mel-frequency Cepstrum Coefficient (MFCC) technique is often used to create the fingerprint of the sound files. The MFCC are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech. Studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the Mel scale. The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.

The algorithm divides the speech sample into fames and then computes mfcc of each frame and stores in the matrix Mel-frequency cepstrum coefficients, mathematical coefficients for sound modeling For speech/speaker recognition, the most commonly used acoustic features are mel-scale frequency cepstral coefficient (MFCC for short). MFCC takes human perception sensitivity with respect to frequencies into consideration, and therefore are best for speech/speaker recognition. We shall explain the step-by-step computation of MFCC in this section.

HMM

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (*hidden*) states. An HMM can be presented as the simplest dynamic Bayesian network. The mathematics behind the HMM was developed by L. E. Baum and coworkers. It is closely related to an earlier work on optimal nonlinear filtering problem by Ruslan L. Stratonovich, who was the first to describe the forward-backward procedure.

In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a *hidden* Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a 'hidden' Markov model even if these parameters are known exactly.

Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics. A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other. Recently, hidden Markov models have been generalized to pairwise Markov models and triplet Markov models which allow consideration of more complex data structures and the modelling of nonstationary data.

Speech segmentation

Speech segmentation is the process of identifying the boundaries between words, syllables, or phonemes in spoken natural languages. The term applies both to the mental processes used by humans, and to artificial processes of natural language processing.

Speech segmentation is a subfield of general speech perception and an important subproblem of the technologically-focused field of speech recognition, and cannot be adequately solved in isolation. As in most natural language processing problems, one must take into account context, grammar, and semantics, and even so the result is often a probabilistic division (statistically-based on likelihood) rather than a categorical one. Though it seems that coarticulation - a phenomenon which may happen between adjacent words just as easily as within a single word - presents the main challenge in speech segmentation across languages, some other problems and strategies employed in solving those problems can be seen in the following sections.

This problem overlaps to some extent with the problem of text segmentation that occurs in some languages which are traditionally written without inter-word spaces, like Chinese and Japanese, compared to writing systems which indicate speech segmentation between words by a word divider, such as the space. However, even for those languages, text segmentation is often much easier than speech segmentation, because the written language usually has little interference between adjacent words, and often contains additional clues not present in speech (such as the use of Chinese characters for word stems in Japanese).

Digit segmentation

In this paper we propose a method to evaluate segmentation cuts for handwritten touching digits. The idea of this method is to work as a filter in segmentation-based recognition system. This kind of system usually rely on over-segmentation methods, where several segmentation hypotheses are created for each touching group of digits and then assessed by a general-purpose classifier. The novelty of the proposed methodology lies in the fact that unnecessary segmentation cuts can be identified without any attempt of classification by a general-purpose classifier, reducing the number of paths in a segmentation graph, what can consequently lead to a reduction in computational cost. An cost-based approach using ROC (receiver operating characteristics) was deployed to optimize the filter. Experimental results show that the filter can eliminate up to 83% of the unnecessary segmentation hypothesis and increase the overall performance of the system.

Digit Recognition

Recognizing natural speech is a challenging task. Human speech is parameterized over many variables such as amplitude, pitch, and phonetic emphasis that vary from speaker to speaker. The problem becomes easier, however, when we look at certain subsets of human speech. For instance, vowels and consonants in the English language are produced in different ways by the vocal tract and accordingly possess unique features that can be exploited to differentiate them from each other. The group, Digital Bubble Bath, from the class of 1996 utilized Formant analysis to isolate, characterize, and identify vowels by their resonant frequencies with great success. Digital Bubble Bath focused on the periodic steady state characteristics of speech. Our group aimes to identify speech by its transient characteristics which include recognition of consonants. We chose the spoken digits one to five as our study set since they are short monosyllabic words with a detectable amount transient behavior. The time domain representation of a spoken five is shown in Figure 2.1. A sample set of all five spoken digits may be found here.

As you can see each signal possesses both periodic or steady state behavior as well as transient behavior. The periodic sections - in general the latter portion of the signals - correspond to the pronunciation of vowels while the transient spikey sections correspond to the pronunciation of consonants. Consonants are physically generated by the stopping of air, intuitively confirming the consanant's transient behavior.

Wavelet Basic

It turns out that there do exist basis functions that fit the bill - namely wavelets. Wavelets are a cross between the impulse and the sinusoid - a wiggle that's localized in time. The wavelet dies off at negative and positive infinity giving location in time. The wavelet's wiggle gives the frequency content. For our project we chose the 32 point Daubechies wavelet generated by the Matlab command *daubcqf.m* from the Rice Wavelet Toolbox for MATLAB for two reasons.

1. Apparently it is the default wavelet for time frequency analysis.

2. It looks a lot like the transient parts of speech.

With Fourier Analysis we compared our signals to a basis consisting of sinusoids that differed in frequency. With wavelet analysis we compare our signals to a basis consisting of wiggles that differ in frequency and temporal location. Surprisingly such a set is generated by one wavelet prototype or mother wavelet. The wavelet W may be represented through the wave equation as a function of two parameters - frequency and time and thus may be expressed as:

W = g(f*t + t')

Algorithms

We use mainly two algorithms for this project: MFCC and HMM. MFCC is used for representing the features from a speech sample and HMM is used to classify the features to classes (nothing but words).

MFCC
* Obtain Speech sample
* Perform FFT (N=1024)
* Perform Fixed Size windowing (HAMMING)
* MFCC= $10\log_{10}(x(f))$ where $x(f)=fft(x(n))$

Wavelet

- Decompose signal into 16 sub bands

- Calculate Energy of each subband. Energy is given by total squared magnitude response of each band

- Normalize the energy values by dividing the values by maximum of all the values

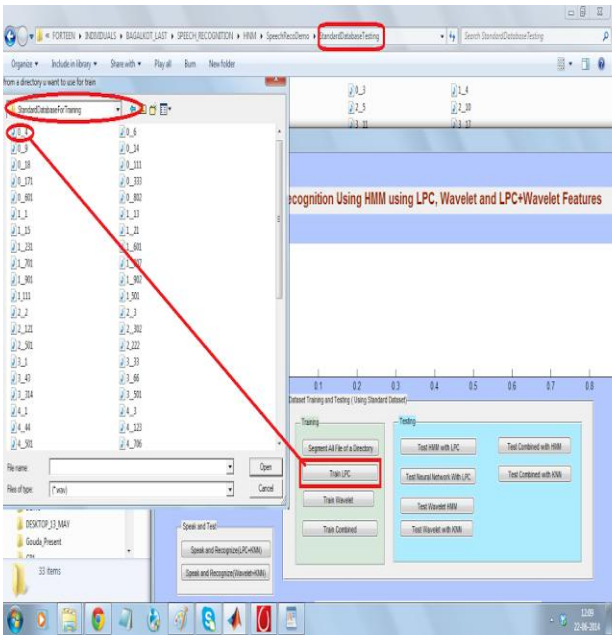
Combined Features

Combine 16 LPC and 16 wavelet features into a single feature vector.

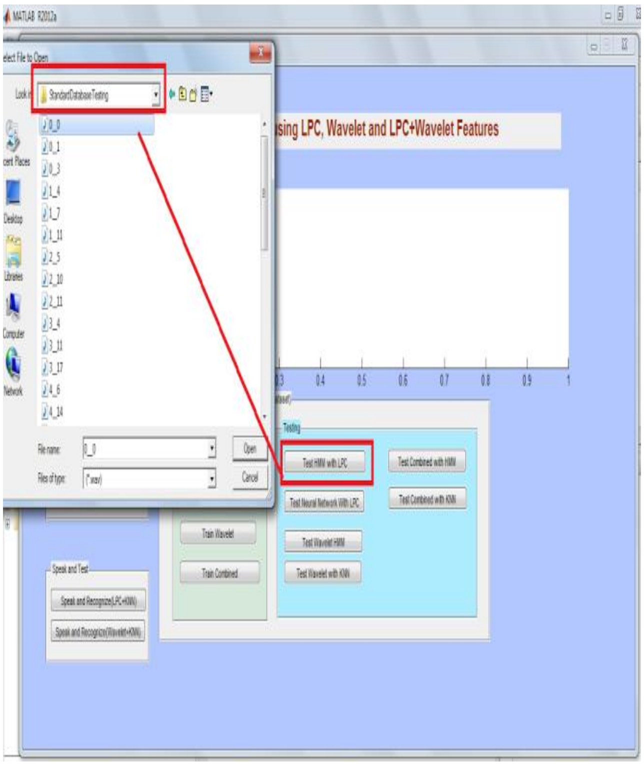
HMM

1. Obtain MFCC from spoken word.
2. Give it a class marking
3. Assume a equal priori probability of occurrence of each word
4. During testing, create a Transition Matrix
5. Update the matrix with feature data to recreate the transition and emission matrix
6. Obtain test features.
7. Check log likelyhood probability of the features to be belonging to each class and select the class with maximum probability score

V. RESULTS

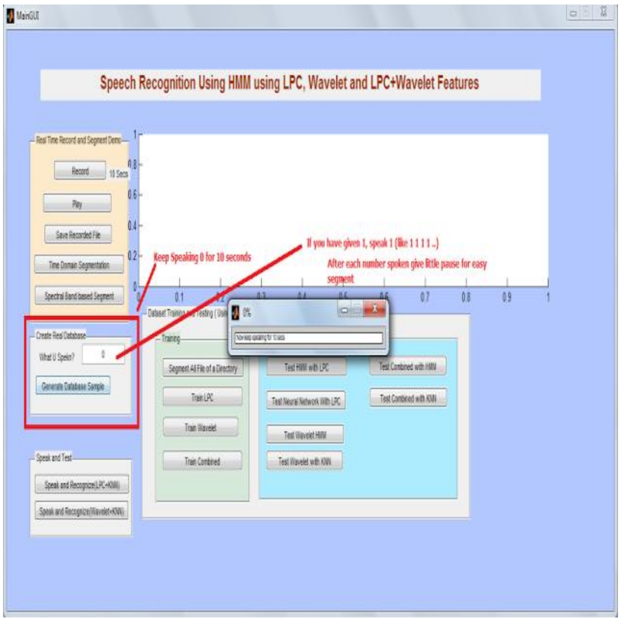

StandardDatabaseTesting
StandardDatabaseForTraining
ecognition Using HMM using LPC, Wavelet and LPC+Wavelet Features
Dataset Training and Testing (Using Standard Dataset)
Training
Segment All File of a Directory
Train LPC
Train Wavelet
Train Combined
Testing
Test HMM with LPC
Test Neural Network With LPC
Test Wavelet HMM
Test Wavelet with KNN
Test Combined with HMM
Test Combined with KNN
Speak and Test
Speak and Recognize(LPC+KNN)
Speak and Recognize(Wavelet+KNN)
File name:
Files of type:
Open
Cancel

Training LPC Method



Testing HMM with LPC

	LPC+ ANN	LPC + HMM	Wavelet+HMM	Wavelet+ANN	Wavelet+LPC +HMM	Wavelet + LPC + HPP+ ANN
Accuracy Speaker verification	91	93	93.7	90.2	96.5	98.7
Accuracy Speech Recognition	84	88	91	83	92	96

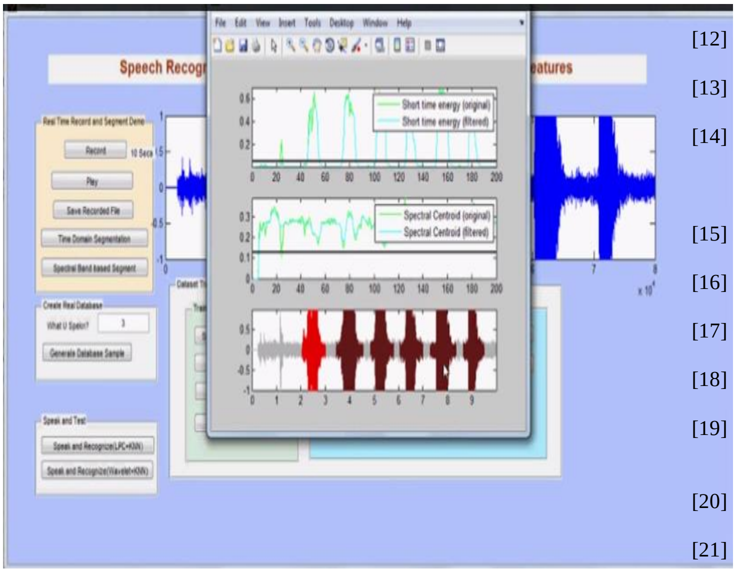


Keep Speaking 0 for 10 seconds

If you have given 1, speak 1 (like 1 1 1 1 ..)

After each number spoken give little pause for easy segment

Real Time Testing


File Edit View Insert Tools Desktop Window Help
Speech Recogn
eatures
Real Time Record and Segment Demo
Record
10 Secs
Play
Save Recorded File
Time Domain Segmentation
Spectral Band based Segment
Create Real Database
What U Speak?
3
Generate Database Sample
Speak and Test
Speak and Recognize(LPC+KNN)
Speak and Recognize(Wavelet+KNN)
Short time energy (original)
Short time energy (filtered)
Spectral Centroid (original)
Spectral Centroid (filtered)
0
20
40
60
80
100
120
140
160
180
200
0.6
0.4
0.2
0.3
0.1
0.5
-0.5
0
1
2
3
4
5
6
7
8
9
x 10^4

Speech Segmentation

Conclusion

User speaks small word or digit like '1', '2' and system must convert them to text. The work is designed to recognize speech in varying criteria like shaky and unclear voices. It was observed that the performance of the system also varies with sensitivity to microphones. Different quality microphone attributes to different noise cancellation. Low quality microphones may affect the accuracy to a great deal; Another important factor is speed of the system. HMM is found to be quite slow due to it's O(n^2) Complexity where n is number of classes. However accuracy wise HMM is far better than ANN under varying voice. Hence with more features and classes, the log likelihood estimation needs several iteration. Therefore this work helps reducing HMM processing by allowing less features and by combining with the results of ANN classifier too.

REFERENCES

[1] Sadaoki Furui, 50 years of Progress in speech and Speaker Recognition Research , ECTI Transactions on Computer and Information Technology, Vol.1. No.2 November 2005.

[2] [2]. Mahdi Shaneh, and Azizollah Taheri Voice Command Recognition System Based on MFCC and VQ Algorithms.

[3] [3]. T.B.Martin, A.L.Nelson, and H.J.Zadell, Speech Recognition b Feature Abstraction Techniques,Tech.Report AL-TDR-64-176,Air Force Avionics Lab,1964.

[4] [4]. D.R.Reddy, An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave ,Tech.Report No.C549, Computer Science Dept., Stanford Univ., September 1966.

[5] [5]. F.Itakura, Minimum Prediction Residula Applied to Speech Recognition ,IEEE Trans.Acoustics,Speech,Signal Proc., ASSP-23(1):67-72,February 1975.

[6] [6]. F.Jelinek, The Development of an Experimental Discrete Dictation Recognizer, Proc.IEEE, 73(11):1616-624, 1985.

[7] [7]. D.Klatt, Review of the ARPA Speech understanding project, J.A.S.A. 62(6), pp.1324-1366, 1977.

[8] [8]. R.K.Moore, Twenty things we still don t know about speech, Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research and Technology , 1994.

[9] [9]. Ms. Arundhati S. Mehendale and Mrs. M. R. Dixit “speaker identification” signals and image processing

[10] C.C .Tappert,N.R.Dixon, A.S.Rabinowitz, and W.D.Chapman, Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, DualClassification, Sequential Decoding and Error Recover ,Rome Air Dev.Cen, Rome, NY,Tech.Report TR-71-146,1971.

[11] K.Nagata, Y.Kato, and S.Chiba, Spoken Digit Recognizer for Japanese Language, NEC Res.Develop. No. 6, 1963.

[12] K.H.Davis, R.Biddulph, and S.Balashek, Automatic Recognition of spoken Digits, J.Acoust.Soc.Am., 24(6):637-642,1952.

[13] H.F.Olson and H.Belar, Phonetic Typewriter, J.Ac oust.Soc.Am.,28(6):1072-1081,1956.

[14] D.B.Fry, Theoritical Aspects of Mechanical speech Recognition , and P.Denes, The design and Operation of the Mechanical Speech Recognizer at Universtiy College London, J.British Inst. Radio Engr., 19:4,211-299,1959.

[15] J.W.Forgie and C.D.Forgie, Results obtained from a vowel recognition computer program , J.A.S.A.,31(11),pp.1480-1489.1959.

[16] J.Suzuki and K.Nakata, Recognition of Japanese Vowels Preliminary to the Recognition of Speech , J.Radio Res.Lab37(8):193-212,1961.

[17] T.Sakai and S.Doshita, The phonetic typewriter,information processing 1962 , Proc.IFIP Congress, 1962.

[18] K.Nagata, Y.Kato, and S.Chiba, Spoken Digit Recognizer for Japanese Language, NEC Res.Develop. No.6,1963.

[19] T.B.Martin, A.L.Nelson, and H.J.Zadell, Speech Recognition b Feature Abstraction Techniques ,Tech.Report AL-TDR-64-176,Air Force Avionics Lab,1964.

[20] T.K.Vintsyuk, Speech Discrimination by Dynamic Programming, Kibernetika, 4(2):81-88,Jan.-Feb.1968.

[21] H.Sakoe and S.Chiba, Dynamic programming algorithm optimization for spoken word recognition ,IEEE Trans.Acoustics, Speech, Signal Proc., ASSP-26(1).pp.43-49,1978.

[22] D.R.Reddy, An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave ,Tech.Report No.C549, Computer Science Dept., Stanford Univ., September 1966.

[23] V.M.Velichko and N.G.Zagoruyko, Automatic Recognition of 200 words, Int.J.Man-MachineStudies,2:223,June 1970.

[24] H.Sakoe and S.Chiba, Dynamic Programming Algorithm Optimization for Spoken Word Recognition ,IEEE Trans.Acoustics, Speech, Signal Proc.,ASSP-26(1):43-49,February 1978.

Rupinder Kaur received the B.E degree in Information science and engineering from visvesvaraya technological university, Belgaum , Karnataka. India in 2011, pursing M.Tech (4th sem) degree in computer science and engineering from visvesvaraya technological university, Belgaum, Godutai Engineering College for Women, Gulbarga, Karnataka India in 2015

.

Pushpalata S (Asst.Prof) received her Mtech degree (CSE) from Poojya Doddappa Appa Engineering College, Gulbarga, and Presently working as Asst Prof in Godutai Engineering College,Gulbarga.

.