RESEARCH ARTICLE OPEN ACCESS

Automatic Text-Independent Speaker Tracking System

Santoshi Pandiri [1], M. Bhargavi [2]
M.Tech Student [1], Assistant Professor [2]
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad
Telangana – India

ABSTRACT

Speaker tracking system aims to detect speech segments corresponding to known target speakers in a known audio resource. Three major applications of speaker tracking are broadcast news, meetings or seminars and telephone conversations. Speaker tracking system finds the conversations between several speech sources (persons) in which few are already enrolled to speech tracking system and other are unknown speech sources, and a target speaker will be chosen in a set of enrolled users. In the first step, speech is segmented in to knowledge about speakers. Then, the resulting segments are checked to belong to one of the target speakers. Speaker tracking is also analyzed by lip motion features and dynamic facial muscle model. The main disadvantage with the above approaches s the speaker has to be physically present. To overcome this disadvantage speaker tracking using voice is used. The aim of my proposed research work is modified Speaker tracking procedure, which results in higher tracking rates which is very useful technique for many clients. To provide smooth flow of tracking and detecting speakers in multi-party conversations by maintaining the performance, clarity of online speaker tracking system and to manage high number of different speakers and also to encounter unseen speakers, we are proposing hidden Markov model for segmenting and classifying speakers.

Keywords:- Speaker Identification; MFCC; HMM; Feature Extraction; Vector Quantization Forward-Backward.

I. INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking on the basis of information obtained from speech waves. This technique will make it possible to verify the identity of persons accessing systems, that is, access control by voice, in various services. These services include voice dialling, banking transaction over telephone network, telephone shopping, database access services, information and reservation system, voice mail, security control for confidential information areas, and remote access to computers. Speaker recognition is probably the only biometric which may be easily tested remotely through the telephone network, this makes it quite valuable in many real applications, and it will become more popular in the future.

Speaker recognition is divided into speaker verification and speaker identification. For speaker verification an identity is claimed by the user, and the decision required of the verification system is strictly binary; i.e., to accept or reject the claimed identity.

Speaker identification is the process of determining

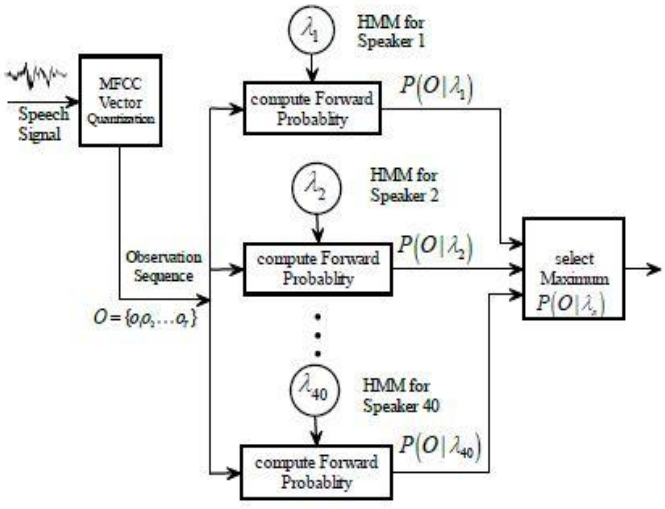

Fig1. Block diagram of HMM-Based Recognizer

which speaker in a group of known speakers most closely matches the unknown speaker. The data used in the recognition is divided into text-dependent and text-independent. In text-dependent, the speaker is required to provide utterances having the same text

for both training and recognition, whereas the text-independent systems allow the user to utter any text. To identify the unknown speaker we select HMM model whose likelihood is the highest.

II. MEL - FREQUENCY

Psychophysical studies have shown that human perception of the frequency content of sounds does not follow a liner scale. That research has led to the concept of the subjective frequency, i.e., the perceived frequency of sounds is defined as follows. For each sound with an actual frequency, f, measured in Hz, a subjective frequency is measured on a scale called the "Mel scale".

Mel-frequency can be approximated by

$$mel\ (f) = 1127 In(\frac{f}{700} + 1)$$

where *f* in Hz, is the actual frequency of the sound.

III. CEPSTRUM

Cepstrum is defined as the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform. The first application of cepstrum to speech processing was proposed by Noll, who applied the cepstrum to determine the pitch period. The cepstrum used also to distinguish underground nuclear explosions from earthquakes.

Cepstrum =ifft(log(|fft(signal)|))

IV. TRIANGULAR FILTERS BANK

The human ear acts essentially like a bank of overlapping band-pass filters and human perception is based on Mel scale. Thus, the approach to simulating the human perception is to build a filter bank with bandwidth given by the Mel scale and pass the magnitudes of the spectra, through these filters and obtain the Mel-frequency spectrum. We define a triangular filter-bank with M filters (m=1,2,…,M) and N points Discrete Fourier Transform (DFT) (k=1,2,…,N), where, [ ] m H k , is the magnitude (frequency response) of the filter.

$$H_m[k] = \begin{cases} 0, & k < f[m-1] \\ \frac{(k - f[m-1])}{(f[m] - f[m-1])}, & f[m-1] \le k \le f[m] \\ \frac{(f[m+1] - k)}{(f[m+1] - f[m])}, & f[m] \le k \le f[m+1] \\ 0, & k > f[m+1] \end{cases}$$

Such filters compute the average spectrum around each center frequency with increasing bandwidths. Let f_l and f_h be the lowest and highest frequencies of the filter-bank in Hz, f_s the sampling frequency in Hz, *M* the number of filters, and *N* the size of the Fast Fourier Transform.

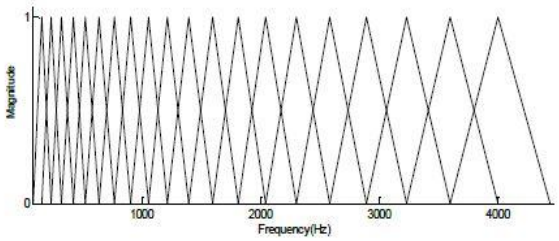

Fig2. Filter bank for generating Mel-Frequency Cepstrum Coefficients

V. HIDDEN MARKOV MODELS

Hidden Markov model (HMM) describes a two-stage stochastic process. The first stage consists of a Markov chain. In the second stage then for every point in time t an output or emission (observation symbol) is generated. This sequence of emissions is the only thing that can be observed of the behavior of the model. In contrast, the state sequence taken on during the generation of the data cannot be observed.

The input speech signal is converted into vectors of MFCC. Then the feature vectors are quantized into observation sequences. The quantization is achieved by k-mean algorithm and classification procedure.

Vector quantization is required to map each continuous observation vector (MFCC) into a discrete codebook index (or symbols). The resulting symbols are new features which were used as input to estimate the HMM parameters.

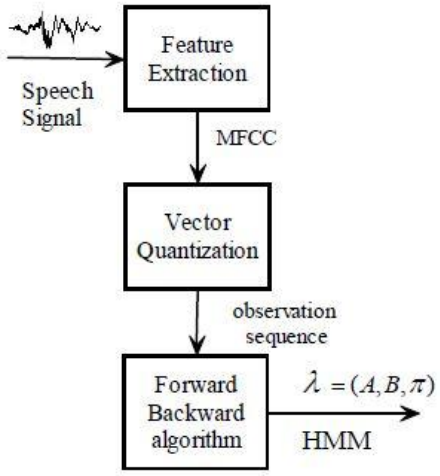

Fig3. Steps used to estimate the parameters of HMMs

The input speech signal is converted into vectors of MFCC. Then the feature vectors are quantized into observation sequences. The quantization is achieved by k-mean algorithm and classification procedure.

Vector quantization(VQ) is required to map each continuous observation vector (MFCC) into a discrete codebook index (or symbols). The resulting symbols are new features which were used as input to estimate the HMM parameters.

Finally, the models parameters are estimated from the observation sequences using the Forward-Backward Algorithm.

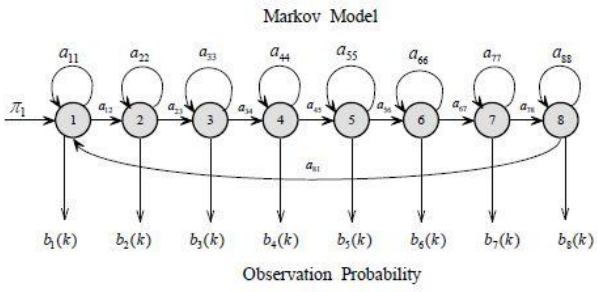

Fig5. HMM-Model

In the training phase, using the clustering algorithm described in a speaker-specific, VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. By using these training data, features are clustered to form a codebook for each speaker.

In the recognition stage, the data from the tested speaker is compared to the codebook of each speaker and measure the difference. These differences are then use to make the recognition decision. So, both training and testing vectors were quantized to generate the observation sequences to be input into the HMMs.

VI. CONCLUSION

We have attempted to describe a Automatic Text-Independent Speaker Tracking System based on Mel Frequent Cepstral Coefficient (MFCC) feature vectors and Hidden Markov Model(HMM). The Voice Recognition Using Feature Extraction employ wavelets in voice recognition for studying the dynamic properties and characteristics of the voice signal. This is carried out by estimating the formant and detecting the pitch of the voice signal by using MFCC. The voice recognition system that is developed is word dependant voice verification system used to verify the identity of an individual based on their own voice signal using the statistical computation, formant estimation and wavelet energy. A GUI is built to enable the user to have an easier approach in observing the step-by-step process that takes place in Wavelet Transform. By using the preloaded voice signals from five individuals, the verification tests have been carried and good accuracy rate of has been achieved.

VII. FUTURE SCOPE

The identification rate achieved in this paper was carried out by using a close-set database. In the future research, we will apply an open-set database since an open-set may cause the experiment to be similar to real-life situations. We will study and use Gaussian Mixture Models (GMM) to estimate the probability function of the feature vectors. Future work will be focused on other models such as Support Vector Machine (SVM) classifier or Kernel method.

REFERENCES

[1] C. H. Lee, F. K. Soong and K. K. Paliwal, Automatic Speech and Speaker Recognition:Advanced Topics, Boston: Kluwer Academic Publishers, 1996.

[2] H. Beigi, Fundamentals of Speaker Recognition, New York: Springer Science and Business Media, Inc, 2011.

[3] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, New Jersey: Prentice-Hall, 1978.

[4] R. L. Klevans and R. D. Rodman, Voice Recognition, Boston: Artech House, 1997.

[5] [5] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[6] L.R.Rabiner and B. Juang, Fundamentals of Speech Recognition, New Jersey: Prentice-Hall, Inc, 1993.

[7] X. Huang, A. Acero and H. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, New Jersy: Prentice-Hall,Inc, 2001.

[8] L. R. Rabiner and R. W. Schafer, Theory and Applications of Digital Speech Processing, New Jersey: Pearson Higher Education, 2011.

[9] S. Chakroborty, A. Roy and G. Saha, "Improved Closed Set Text-Independent Speaker Identification by combining MFCC with Evidence from Flipped Filter Banks," International Journal of Information and Communication Engineering, Vol. 4, No. 2, pp.114-121, 2008.