Delta-MFCC based text-independent speaker recognition system

[1] Shivangi Chaudhary, [2]Deepali Jain
[1]Student, [2]Student
[1]Communication Engineering,
[1]Galgotias University, Greater Noida, India

***Abstract* - Speaker Recognition is a technique that uses the acoustic features of the speech of the individual for his/her identification among the group of known/unknown speakers. Recognition of the particular individual is based on the information present in his/her voice. In this paper, MFCC (Mel Frequency Cepstral Coefficients) features and Delta MFCC features are extracted from the speech samples of the speaker. Paper includes comparative studies of the speaker identification results by Delta-MFCC features and MFCC features using Gaussian Mixture Model. In the present work, the advantage of using Delta-MFCC features over MFCC is proposed, as the MFCC only gives the power spectral envelope of the particular frame, to include the trajectories of MFCC coefficients over time delta features has been calculated.**

***Index Terms* - MFCC, Delta-MFCC, GMM**

I. INTRODUCTION

Speaker Recognition systems finds their application mainly in biometrics these days [1][2]. Their use is increasing day by day for accessing the personal information passwords in computer, control access to services such as voice dialing ,tele-banking services, services related to database access, security control and so many. Further Speaker recognition can be divided into two categories, speaker identification and speaker verification [3]. Speaker identification does not claim the identity of the person, although it gives the closest match of the individual based on matching their train samples with the test samples. Speaker verification deals with the procedure adapted to fulfill the claim of the speaker's identity.

Speaker recognition systems can be further classified as text independent and text Dependent.[4]Text dependent systems are the systems in which speaker has to speak the same code during identification that is already present in the database during training phase. On the other hand, text-independent systems are independent of the text of the speaker during the identification. They are unaware of the speaker's spoken words.

Text – independent and dependent systems can be further classified into closed set and open set. In text, independent speaker identification where result is the most likely speaker of the test speech can be further classified into closed set and open set, where closed set deals with the task of identifying the person who is already a member of the set of registered speakers.[4] The limitation of this method is there is a risk of false identification, to correct these false recognition results open set is preferred where system is able to identify the speaker whether he/her belongs to or outside the set of the enrolled speakers. However, closed set systems find their applications only where it has to be surely used among the set of registered speakers.

Feature extraction is the process of extracting relevant information from the speaker's voice sample by reducing its dimension. For-extracting feature, different techniques already exist in literature [5]. As shown in [6] LPC (Linear Prediction Coding), LPCC (Linear Prediction Coding – Derived Cepstral Coefficient) [7] and MFCC (Mel Frequency Cepstral Coefficients) [8]. MFCC is used for feature extraction purpose due to its far superior performance from the other methods. However, a MFCC feature suffers the problem that they are static in nature. They alone are unable to represent the information related to trajectories. To include the information related to trajectories Delta-MFCC features are calculated. Delta-MFCC features are the time derivative of the MFCC features also called as Dynamic features [9].

In this paper, the speaker identification results are evaluated using Delta-MFCCfeatures and MFCC features. Section 2 contains brief description about methodology used for feature extraction. A description of classifier used for identification process is briefly explained in section 3. Further, section 4 covers the result and discussion. Finally, conclusion is included in the section 5.

II. PROPOSED METHOD

Speech is a non-stationary signal. Human voice continuously varies with time. Speech waveform contains individual speaker information. For recognition, a system that solely depends upon human voice for speaker identification extracts relevant information from their voice samples. Several feature extraction methods used in recognition systems, but the most efficient method for feature extraction is MFCC. Mel Frequency Cepstral Coefficients is used in recognition systems due to its far superior performance from the other methods.

III. FEATURE EXTRACTION

Mel- Frequency Cepstral Coefficients (MFCC) features are the most common method used for feature extraction. It gives the static coefficients. MFCC features alone do not convey any trajectory over time. To include the temporal information the difference of the MFCC of the adjacent frames are computed, calculated coefficients are known as Delta-MFCC features [9].

MFCC almost mimics the human auditory system by the use of Mel scale. To obtain the acoustic features the speech is segmented into frames. Hamming window of 20-30 ms is applied to nullify the distortion present at the beginning and end of the frame with an overlapping of 10ms. For the conversion of the time domain to frequency domain Fast Fourier transform is applied to each windowed frame. The obtained linear frequency spectrum is then converted into the Mel frequency spectrum by using Mel filter bank. After this the logarithm of all filter bank energies is taken and then by taking the DCT of the log energies, 13 MFCC coefficients are obtained and rest are discarded.13Delta-MFCC features are obtained by taking the time derivative of the MFCC coefficients. To use Delta-MFCCfeatures for Speaker Identification, Gaussian Mixture Model classifier has been used as a classifier in the present work. Due to the problem suffered by MFCC coefficients, which are static in nature, Delta-MFCC features are calculated which are dynamic in nature.

IV. GAUSSIAN MIXTURE MODEL

Gaussian mixture model is nothing but single state Hidden Markov Model (HMM). HMM springs forth from Markov Processes or Markov Chains. It is a canonical probabilistic model for the sequential or temporal data it depends upon the fundamental fact of real world, "Future is independent of the past but driven by the present". The HMM is a doubly embedded stochastic process, where final output of the system at a particular instant of time depends upon the state of the system and the output generated by that state[10].

There are two types of HMMs: Discrete HMMs and Continuous Density HMMs. The type of data that they operate upon distinguishes these. Discrete HMMs (DHMMs) operate on quantized data or symbols, on the other hand, the continuous density HMMs (HMMs) operate on continuous data and their emission matrices are the distribution functions. The basic notations of HMM are as shown in Table I[10].

Table I Basic Notations

Variable	Notations
Number of States	N
Number of observation symbols per state	M
Observation symbols	$V\{v_1, v_2, \ldots, v_M\}$
Observation Sequence	$O\{O_1, O_2 \ldots, O_T\} \in X$ {discrete value, real value}
State Sequence	$Q\ \{q_1, q_2, \ldots, q_T\}$
Transition Matrix	$A = \{a_{i,j}\} = P(q_{t+1} = S_j \mid q_t = S_i) 1 \leq i, j \leq N$
Emission Matrix	$b_i(k) = P(v_k \text{ at } t \mid q_t = S_i) 1 \leq i \leq N;\ 1 \leq k \leq M$
Initialization Matrix	$\pi(i) = P(q_1 = S_i) 1 \leq i \leq N$
Space of all state sequence of length T	ϱ
Mixture component for each state at each time	$m\{m_{q1}, m_{q2}, \ldots m_{qT}\}$
Mixture component (i state and *c*omponent)	$c_{il}, \mu_{il}, \sum_{il}$
Model of the System	$\lambda(A, B, \pi)$
Parameter for Maximum Likelihood estimation	λ_{ML}
EM Auxiliary Function	$Q(\lambda, \lambda^{(i-1)})$, where superscript for iteration (i-1)

There are three major design problems associated with an HMM outlined here

1. Given the Observation Sequence {O1, O2, O3,OT} and Model λ(A, B, π), the first problem is the computation of probability of the observation sequence P (O|λ).
2. The second problem finds the most probable state sequence Q {q1, q2,… ,qT}
3. The third problem is related to the choice of the model parameters λ (A, B,π), such that the Probability of the Observation sequence, P (O|λ) is the maximum.
   The solution to the above problems emerges from three algorithms: Forward, Viterbi and Baum-Welch.

V. EXPERIMENTAL RESULTS

Database

In this paper, TIMIT database is taken as reference for the extraction of the features. We performed experiment on the speech database of 30 speakers for speaker recognition. TIMIT acoustic speech corpus contains16 kHz recordings of 630 speakers of 8 dialects of American English, each reading phonetically 10 rich sentences, out of which 8 sentences are used for the training

phase and rest 2 sentences for the validation of the method. As per experiment requirement, only 30 speaker's voice samples are taken, each reading 10 sentences. Table 2 Compare the results of proposed feature extraction method with the MFCC.

Table II. Speaker Identification Results

Feature Extraction Technique	Identification Results
Liner Predictive Model	73
Auto Regressive Model	76

VI. Conclusion

Speaker recognition systems are the most interesting area of research. As the MFCC is one of the efficient methods used for Feature extraction, but it suffers the problem of not conveying temporal information alone, Delta-MFCC features are used for speaker Identification for better accuracy results in the paper by using Gaussian Mixture Model as a classifier. Delta MFCC features could give better identification results than the conventional MFCC if used for identification.

References

[1] Reynolds, D. (1994). Experimental evaluation of features for robust speaker identification. IEEE Transactions, (pp. 639-643)

[2] Reynolds, D., & Rose, R. (1995). Robust text-independent speaker identification using Gaussian mixture Speaker models. IEEE Transactions, (pp. 72-83).

[3] Campbell, J. (1997). Speaker recognition: a tutorial. IEEE, (pp. 1437- 1462).

[4] Doddington, G. (1985). Speaker recognition—Identifying people by their voices. IEEE, (pp. 1651,1664)

[5] Wang., X., & k.paliwal. (2007). Feature extraction and dimensionality reduction algorithms and their Application in vowel recognition. International Conference, (pp. 254-260).

[6] Paul, A., Das, D., & Kamal, M. (2009). Bangla Speech Recognition System using LPC and ANN. ICAPR (pp. 171,174).

[7] Hai Yan Yang, X.-X. J. (2012). Performance test of parameters for speaker recognition system based on svm- Vq. International Conference, (pp. 321-325).

[8] Hossan, M., Memon, S., & Gregory, M. (2010). A novel approach for MFCC features extraction. International Conference, (pp. 13-15)

[9] Tokuda, K., Kobayashi, T., & Imai, S.(1995).Speech parameter generation from HMM using dynamic Features. (pp. 660-663).

[10] Bhardwaj, S., Srivastava, S., Gupta, J., Bhandari, A., Gupta, k., & Bahl, H. (2012). Wavelet Packet Based Mel Frequency Cepstral Features for Text Independent Speaker Identification. Springer, (pp. 237-247).