Speaker Identification techniques in Overlapping Speech Analysis: A Review

Mrunal Bhogte
Department of Computer Engineering
Shree L. R. Tiwari College of Engineering, Thane , India.
mrunalbhogte92@gmail.com

Prof. Shanthi Therese
Department of Information Technology
Thodomal Shahani Engineering College, Mumbai, India.
shanthitherese123@gmail.com

Prof. Madhuri Gedam
Department of Information Technology
Shree L. R. Tiwari College of Engineering, Thane, India.
madhuri.gedam@gmail.com

Abstract--This paper addresses the issue of speaker identification in overlapping speech where two speakers are speaking simultaneously over a single communication channel where objective is to find individual speaker identities. This task can be accomplished using feature extraction techniquesbased on which classification model can be developed.This paper studies different feature extraction techniques such as MFCC and GFCC. Also various modeling techniques such as GMM, HMM, SVM, UBM, ANN and DNN are studied.

Keywords—Speaker, Identification, supervised, feature extraction.

*****

I.INTRODUCTION

Speech is an audio signal produced through several transformations at linguistic, articulatory, semantic, and acoustic levels. Differences between these levels result in the difference of properties of the speech signal. Also, based on the individual speaker, the differences between inherent characteristics in the vocal tract and the way of speaking will differ. All these factors are taken into account while identifying speakers from a speech signal.Speaker identification or speaker recognition is the process of identifying the individual speaker from a speech signal either in overlapping or cochannel or isolated signal domain. Speaker identification in overlapping speech where multiple speakers are talking simultaneously over a single communication channel is a problem where the task is to separate the speech based on voice characteristics and then identifying or labeling the speakers separately. Out of these two procedures separating the speech from mixture of signal is animportant task. Singer identification has applications in the area of security systems, access control systems, criminal investigations in digital forensics etc.

II. CLASSIFICATION OF SPEECH ANALYSIS

Speech analysis can be done in various ways according to the problem at hand. The various possible ways are depicted in the following diagram.

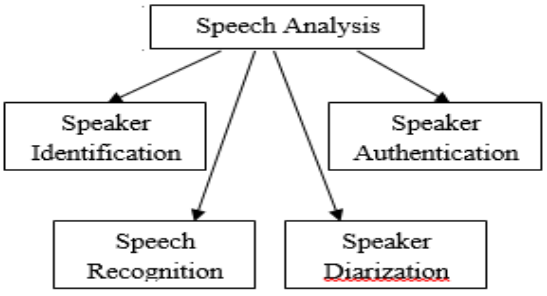

Fig.1 Classification of Speech analysis

1.Speaker identification: It is the recognition of a speaker or a person from features of voice biometrics.

2.Speech recognition: It is the process of identifying what is being said and speaker recognition is identifying who is speaking.

*3. Speaker verification:*It is the process of authenticating the speaker.

*4. Speaker diarisation:*It is the process of identifying when the same speaker is speaking.

III. CLASSIFICATION OF SPEAKER IDENTIFICATION TECHNIQUES

The speaker identification problem is broadly classified as depicted in the following figure.

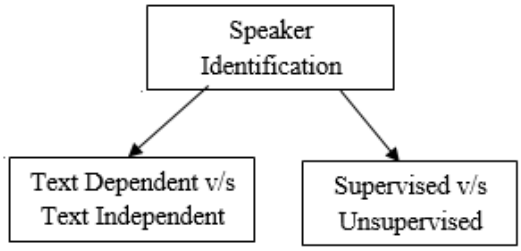

Fig. 2 Different approaches of Speaker Identification

A.Text Dependent v/s Text Independent technique.

This type of approach is dependent on the text or speech used for identification during testing and training.

1. *Text dependent:*If the text must be the same for enrolment and verification this is called text-dependent identification.
2. *Text Independent:*Text-independent systems are mostly used for speaker identification as they require very little if any cooperation by the speaker. In this case the text during enrollment and test is different.

B. Supervised v/s unsupervised approach

1. Supervised Approach: If during identification, speaker identities or models are available, then such an approach is known as supervised approach.

2. *Unsupervised Approach*: If the speaker models or identities are not available during identification of the speaker, then such an approach is known as unsupervised approach.

IV.FEATURE EXTRACTION TECHNIQUES IN OVERLAPPING SPEECH

Speaker identification is a pattern recognition problem. Preliminary stage of speaker identification is feature extraction. The various techniques which can be used for extracting features from a given signal are addressed in the following section including MFCC and GFCC.

A. Mel-Frequency Cepstral Coefficient (MFCC)

The most commonly used technique for feature extraction for speaker identification is MFCC which can be efficiently used in speech analysis. MFCC works based on the characteristics of human auditory system (i.e. pitch and loudness). It tries to remove frequency harmonics to reduce speaker dependent features. Also the feature vector changes over time. MFCC is a preferred technique due to its high accuracy and low complexity level. Major drawback of MFCC technique is, it does not perform well in presence of noise.

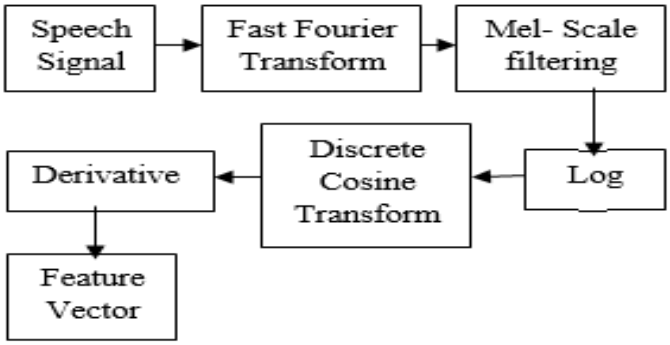

Fig. 3 Block diagram for MFCC feature extraction technique

B. Gammatone Frequency Cepstral Coefficient (GFCC)

Major drawback of MFCC can be reduced by using more efficient GFCC technique. GFCCs use gammatone filters with rectangular bandwidth bands. Gammatone filters model the filter response of human auditory system. From experiments, it is shown that higher classification accuracies are achieved using GFCCs. Also, GFCC is more robust in noisy environment.

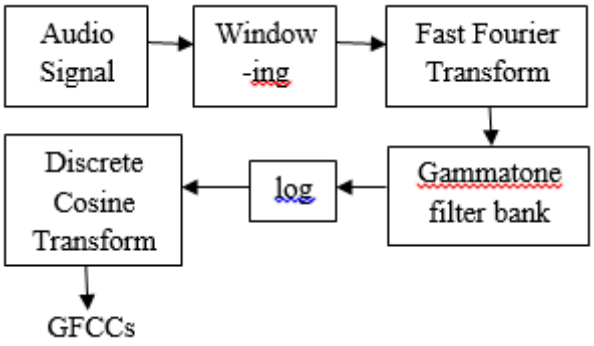

Fig. 4. Block diagram for GFCC feature extraction technique

V. SPEAKER MODELING TECHNIQUES IN OVERLAPPING SPEECH

A. Gaussian Mixture Model (GMM)

GMM is a supervised learning classification algorithm. GMM is widely used technique in speaker recognition. In GMM, first the data is splittedinto k no. of clusters. Each cluster is represented by multivariate Gaussian distribution.The distribution is then fitted to the data using Expectation-Maximization (EM) technique.It is useful for recognition of static and non-temporal patterns. It can provide bad results if problem domain has too many dimensions (i.e. greater than 6 dimensions). This problem can be reduced by using SVM. Another disadvantage of the GMM algorithm is that the user must set the number of mixture models that the algorithm will try and fit to the training dataset. In most of the cases, this is not feasible. GMM is best suited when problem domain

comprises of some hidden non observable patterns since this algorithm is assigning a probability to each point to belong to certain cluster, instead of assigning a flag that the point belongs to certain cluster.

B. Support Vector Machine (SVM)

SVM is a supervised learning method where classification and regression techniques are used. In SVM, the data is classified based on the trained model that assigns either of the two values. Hence, it is a binary classifier. There are two types of these classifiers i.e. linear and non-linear. Non-linear classification can also be implemented using kernel trick. An unsupervised version of SVM can also be implemented when labeled data is not available. It is widely used in the industry. In recent research, new SVM based technique is proposed which is text independent system of identification. It is observed that the training data, time and storage are reduced compared with traditional SVM resulting in robustness. [4]

C. Universal Background Model (UBM)

The universal background model (UBM) is a
Gaussian mixture of many signals that covers entire speech signal. It can be used in speaker identification problem by using a MAP scheme i.e.maximum a posteriori scheme where each speaker is given a value based on MAP results. In one of the researches, UBM is used along with GMM [5] where Bayesian model is used for deriving speaker models from UBM resulting in computational inexpensiveness and high accuracy.

D. Hidden Markov Models (HMM)

HMM assumes the system to be Markov process with hidden states. Hiddenspecifies the sequence of states through which the model passes. It is a Bayesian network.Each state has a probability distribution over the possible output tokens. Hence, the sequence of tokens generated by an HMM gives some information about the sequence of states.A hidden Markov model can be considered a generalization of a mixture model where the hidden variables which control the mixture component to be selected for each observation, are related through a Markov process. They are not independent of each other.Hidden Markov has applications in speech, handwriting, gesture recognition, bioinformatics etc. In recent studies, HMM are categorized as pairwise Markov models and triplet Markov models. This categorization allows consideration of more complex data structures.For a given set of seed sequences, there are many possible HMMs, and choosing one can be difficult. It is expensive, both in terms of memory and compute time.

E. Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is network of artificial neurons, inspired by the principles of biological neural networks. This network can be used in pattern recognition as an efficient tool.Hence, it can be used in the area of speaker identification since speaker identification is a pattern recognition problem. Neural networks can be implemented either as a generative algorithm or discriminative algorithm. Recently, Greedy based algorithm is proposed using neural networks [6] which has shown better result than standard discriminative algorithms.

F. Deep Neural Network (DNN)

A deep neural network (DNN) is a special category of an artificial neural network (ANN) with many hidden layers between the input and output layers. These extra layers extract features from lower layers. This enables to model the complex data efficiently. DNNs are classified as Feed forward network and recurrent neural network. In recent years, DNNs have been used in the field of speech analysis specifically in speaker identification domain which is proven to be very effective compared to traditional techniques. [1][2][3]. In recent studies, it is observed that DNN shows better efficiency than MFCC [1]. Other research studied DNN in noisy and reverberant environment and robust results were observed. [3]

V. CONCLUSION

In this paper, the problem of speaker identification in overlapping speech is studied. The speaker identification consists of two parts, feature extraction and classification for modeling. MFCC and GFCC feature extraction techniques are studied. MCC provides high accuracy and GFCC provides robustness in presence of noise. For classification purpose, GMM, HMM, SVM, UBM, ANN and DNN techniques are studied. Out of these techniques, GMM, SVM, UBM and HMM are traditional techniques. GMM proves to be efficient but suffers a problem in case of higher dimensions. To solve this problem SVM can be used. UBM is mostly used alongwith GMM using a MAP scheme. HMM assumes the system to be Markov process with hidden states. It is an expensive algorithm in case of both memory and computing time.

VI. REFERENCES

[1] K. Chen and A. Salman, “Learning speaker-specific characteristics with a deep neural architecture,” IEEE Trans. Neural Netw., vol. 22, no. 11, pp. 1744–1756, Nov. 2011.

[2] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme forspeaker recognition using a phonetically-aware deep neural network,”in Proc. ICASSP, 2014, pp. 1695–1699.

[3] X. Zhao, Y. Wang, and D. L. Wang, "Robust speaker identification innoisy and reverberant conditions," IEEE/ACM Trans. Audio, Speech,Lang. Process., vol. 22, no. 4, pp. 836–845, Apr. 2014.

[4] Yan Wang, Xuevan Liu, Yuiuan Xing and Ming Li, "A Novel Reduction Method for Text-Independent Speaker Identification'', Natural Computation, 2008. ICNC '08. Fourth International Conference ,Vol.4, pp-66-70, Oct. 2008.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Process., vol.10, pp. 19–41, 2000.

[6] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," Neural Comput., vol. 18, pp. 1527–1554, 2006.

[7] Vimala.C, Radha.V, "Suitable Feature Extraction And Speech RecognitionTechnique for Isolated Tamil Spoken Words", International Journal of Computer Science And Information Technologies, Vol. 5 (1), 2014, 378-383.

[8] Xiaojia Zhaoand Deliang Wang, "Analyzing Noise Robustness of MFCC and GFCfeatures in SpeakerIdentification", IEEE International Conference on Acoustics, Speech and Signal Processing 2013.

[9] Om Prakash Prabhakar, Navneet Kumar Sahu, "Performance improvement of Human VoiceRecognition System Using Gaussian MixtureModel", International Journal of Advanced Research in Computer and Communication Engineering. Vol. 3, Issue 1, January 2014.

[10] Suma Swamy, K.V Ramakrishnan, "An Efficient Speech RecognitionSystem", Computer Science & Engineering: An International Journal (CSEIJ), Vol. 3, No. 4, August 2013.

[11] Jayanth M, B Roja Reddy, " Speaker Identification Based on GFCC Using GMM-UBM", International Journal Of Engineering Science Invention ,Volume 5 Issue 5, Pp.62-65, May 2016.

[12] Daniel Povey, Stephen M. Chu, BalakrishnanVaradarajan, " Universal Background Model Based Speech Recognition" IEEE International Conference on Acoustics, Speech and Signal Processing 2008.