A Review on User Identification using Speech

Mandeep Kaur[1], Simrat Kaur[2]
[1]Student, Dept. of CSE, BBSBEC, Fatehgarh sahib, (Punjab) India
[2]Assistant Professor, Dept. of CSE, BBSBEC, Fatehgarh Sahib, (Punjab) India

Abstract -The most critical task in security is User identification. User identification will be the techniques of automatically recognizing who's going to be communicating on the basis of particular personal info contained in speech waves. In recent year, there has been great development in security techniques. One of the emerging new biometric for identification, is speech. This paper presents a review on leading techniques used for user identification using speech signals and studied their unique features.

Keywords: User Identification, Speaker Identification, GMM, Vector Quantization

I. INTRODUCTION

The speech conveys several levels of information. Primarily, the speech signal conveys the words or message being spoken, but on a secondary level, the signal also conveys information about identity of the talker. While the area of speech recognition is concerned with extracting the underlying linguistic message in an utterance, the area of speaker recognition is concerned with extracting the identity of the person speaking the utterance [1]. In automatic speaker recognition methods, the speaker to be recognized is usually required to speak the same utterance which was used to obtain the reference pattern for that speaker. However, such a restriction is not generally necessary for speaker recognition by humans [2].

In speaker recognition, the basic requirements are extraction of a few features from the speech and then cluster the speech features of same user in one group and the different users in n different groups. An estimation maximization algorithm is then applied on a new test speech to identify the users. Several problems exist in a class based speech model. Since there are N numbers of classes, an unknown speech is likely to match to one of them if its probability is maxima of all the probability, however small it may be. Simple techniques like probability thresholding can be applied to make sure unknown speech is not matched to any existing speech in the database.

Several techniques exist for modeling the speech of a user. One of the well known methods is Gaussian mixture model (GMM). The use of Gaussian mixture models for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities [1].

Another very important method for speech modeling is VQ or Vector Quantization. Vector Quantization is a lossy data compression method based on the principle of block coding. VQ can be thought of as a process of redundancy removal that makes the effective use of nonlinear dependency and dimensionality by compression of speech spectral parameters [3].

This paper describes: Architecture of speaker identification under section II. The speaker modeling and recognition techniques, GMM based model and VQ based model. In both techniques, the core features are MFCCs features also known as Mel frequency cepstral coefficients under section III. Finally, concludes the paper under section IV.

II. ARCHITECTURE OF SPEAKER IDENTIFICATION

Speaker identification machine has two levels: Training segment and Testing Segment. In training segment all the speech samples to be had within the speech database are pre-processed and feature vectors are received. Then these features are modeled. In testing segment, the test speech sample is pre-processed and its features are extracted, when Log likelihood ratio is taken between the test speakers and all to be had speaker items in the database [21]. Then, the speaker is recognized as the one who has maximum likelihood ratio as shown in Fig.1.

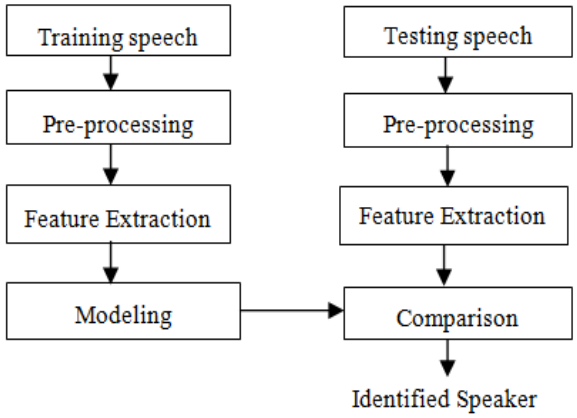

Fig. 1Implementation of proposed approach (flow of events) [21].

III. SPEAKER MODELING AND RECOGNITION TECHNIQUES

The goal of modeling technique and Recognition Techniques would be to extract feature vectors from available speech samples and create speech models by making use of speaker particular feature vector. Speaker recognition can be classified into two categories: speaker dependent and speaker independent modes.

In speaker independent mode recognize system can extract the intended message but can't extract the speaker characteristics of speech signal [5] [6].On the other hand in speaker dependent mode of recognition would be extract the characteristics of speech signal [8] [16].In order to developing speech models there are a lot of techniques such as: Mel Frequency Cepstral Coefficients (MFCCs), Gaussian Mixture Model (GMM), Vector quantization (VQ) etc [16].

A. Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in speaker recognition. They were introduced by Davis and Mermelstein in 1980's, and have been start-of-the-art ever since. Prior to the introduction of MFCCs. Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) were the main feature types for automatic speaker recognition systems [13].

The most popular spectral based parameter used in recognition approach is the Mel Frequency Cepstral Coefficients called MFCC. MFCCs are coefficients, which represent audio, based on perception of human auditory systems. MFCC is used to extract those features which are used by human ears to listen. Firstly, in MFCC, the signal is divided into frames for which the feature vectors are calculated individually. Then the Hamming window is done to each frame. Further, after applying Fast Fourier Transformation (FFT), a Mel Filter Bank is generated. After Mel Frequency wrapping is done to obtain the coefficients, Inverse Discrete Fourier Transformation (IDFT) is calculated for cepstral coefficient generation.

The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely to what humans hear [2].

The formula for converting Hz frequency scale to Mel frequency scale is: [20]

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right)$$

The inverse of it is:

$$M^{-1}(m) = 700(e^{\left(\frac{m}{1125}\right)-1}$$

The functions used for feature extraction [x_cep, x_E, x_delta, x_acc]. MFCC are chosen for the following reasons:-

1. MFCC is the most important features, which are required among various kinds of speech applications.

2. It gives high accuracy results for clean speech.

3. MFCC can be regarded as the "standard" features in speaker as well as speech recognition.

B. A Gaussian Mixture Model (GMM)

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system [18]. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum A-Posteriori (MAP) estimation from a well-trained prior model [4].

A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation, [7]

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i)$$

Where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), $wi,\ i = 1 \ldots M$ are the mixture weights; and $g(x|\mu_i, \Sigma_i)$, i = 1… M, are the component Gaussian densities [7].

GMMs are often used in biometric systems, most notably in speaker recognition systems, due to their capability of representing a large class of sample distributions. One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrarily shaped densities [13]. The classical unimodal Gaussian model represents feature distributions by a position (mean vector) and an elliptic shape (covariance matrix) and a vector quantizes (VQ) or nearest neighbor model represents a distribution by a discrete set of characteristic templates. A GMM acts as a hybrid between these two models by using a discrete set of

Gaussian functions, each with their own mean and covariance matrix, to allow a better modeling capability [12].

The GMM not only provides a smooth overall distribution fit, its components also clearly detail the multi-modal nature of the density [10].

A GMM can also be viewed as a single-state HMM with a Gaussian mixture observation density, or an ergodic Gaussian observation HMM with fixed, equal transition probabilities [16]. Assuming independent feature vectors, the observation density of feature vectors drawn from these hidden acoustic classes is a Gaussian mixture.

C. Vector Quantization (VQ)

Vector quantization (VQ) is a classical quantization technique from signal processing that allows the modeling of probability density functions by the distribution of prototype vectors. It was originally used for data compression. It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. Each group is represented by its centroid point, as in k-means and some other clustering algorithms [3].

The density matching property of vector quantization is powerful, especially for identifying the density of large and high-dimensioned data. Since data points are represented by the index of their closest centroid, commonly occurring data have low error, and rare data high error. This is why VQ is suitable for lossy data compression. It can also be used for lossy data correction and density estimation [11].

Vector quantization is based on the competitive learning paradigm, so it is closely related to the self-organizing map model and to sparse coding models used in deep learning algorithms such as autoencoder [19].

VQ was also used in the eighties for speech and speaker recognition. Recently it has also been used for efficient nearest neighbor search and on-line signature recognition. In pattern recognition applications, one codebook is constructed for each class (each class being a user in biometric applications) using acoustic vectors of this user. In the testing phase the quantization distortion of a testing signal is worked out with the whole set of codebooks obtained in the training phase. The codebook that provides the smallest vector quantization distortion indicates the identified user [15].

The main advantage of VQ in pattern recognition is its low computational burden when compared with other techniques such as dynamic time warping (DTW) and hidden Markov model (HMM). The main drawback when compared to DTW and HMM is that it does not take into account the temporal evolution of the signals (speech, signature, etc.) because all the vectors are mixed up [18].

IV. CONCLUSION

This paper presents two of the very important techniques used in speaker modeling. These two techniques are Gaussian Mixture Models and Vector Quantization. Both these techniques have their own pros and cons. While, GMMs are used on shorter speech segments to generate better results as for a shorter segment, the highly accurate modeling is done for speech signal. VQ is a lossy data compression method which is used as pattern recognition system by basically reducing the feature space and thus reducing the computing cost for attaining higher level of accuracy.

REFERENCES

[1] Reynolds, Douglas, and Richard C. Rose. "Robust text-independent speaker identification using Gaussian mixture speaker models." Speech and Audio Processing, IEEE Transactions on 3.1 (1995): 72-83.

[2] Atal, B. S. "Text-Independent Speaker Recognition." The Journal of the Acoustical Society of America 52.1A (1972): 181-181.

[3] Gill, Manjot Kaur, Reet kamal Kaur, and Jagdev Kaur. "Vector quantization based speaker identification." International Journal of Computer Applications 4.2 (2010): 0975-8887.

[4] Matsui, Tomoko, and Sadaoki Furui. "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's." Speech and Audio Processing, IEEE Transactions on 2.3 (1994): 456-459

[5] Reynolds, Douglas A., and William M. Campbell. "Text-independent speaker recognition." Springer Handbook of Speech Processing. Springer Berlin Heidelberg, 2008. 763-782.

[6] Gish, Herbert, and Michael Schmidt. "Text-independent speaker identification." Signal Processing Magazine, IEEE 11.4 (1994): 18-32.

[7] Zong, Feng. "Speaker Recognition Techniques." Applied Mechanics and Materials. Vol. 599. 2014.

[8] Furui, Sadaoki. "Speaker-dependent-feature extraction, recognition and processing techniques." Speech Communication 10.5 (1991): 505-520.

[9] Rose, Richard C., and Douglas Reynolds. "Text independent speaker identification using automatic acoustic segmentation." Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on. IEEE, 1990:293-296.

[10] Park, Alex, and Timothy J. Hazen. "ASR dependent techniques for speaker identification." INTERSPEECH. 2002: 1337-1340

[11] Kinnunen, Tomi, Evgeny Karpov, and Pasi Franti. "Real-time speaker identification and verification." Audio, Speech, and Language Processing, IEEE Transactions on 14.1 (2006): 277-288.

[12] David, Petr. "Experiments with speaker recognition using GMM." Proc. Radioelek-tronika (2002): 353-357.

[13] Kinnunen, Tomi, et al. "Comparing maximum a posteriori vector quantization and Gaussian mixture models in speaker verification." Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009:88-92

[14] M. Nishida, Masafumi, and Tatsuya Kawahara. "Speaker indexing and adaptation using speaker clustering based on statistical model selection." Acoustics, Speech, and Signal Processing, 2004. Proceedings (ICASSP'04). IEEE International Conference on. Vol. 1. IEEE, 2004:353-356.

[15] Nishida, Masanori, and Tatsuya Kawahara. "Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion." Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP'03). 2003 IEEE International Conference on. Vol. 1. IEEE, 2003:172-175

[16] Hemakumar G, Punitha P “Speech Recognition Technology: A Survey on Indian Languages” International Journal of Information Science and Intelligent System, Vol. 2, No.4, 2013:1-38

[17] Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." Digital signal processing 10.1 (2000): 19-41.

[18] Gray, Robert M. "Vector quantization." ASSP Magazine, IEEE 1.2 (1984): 4-29

[19] Soong, Frank K., et al. "Report: A vector quantization approach to speaker recognition." AT&T technical journal 66.2 (1987): 14-26.

[20] Zivkovic, Zoran. "Improved adaptive Gaussian mixture model for background subtraction." Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Vol. 2. IEEE, 2004:28-31.

[21] S. Dey, S.Barman, R.Bhukya, R.Das, Haris, S.Prasanna and R. Sinha,“ Speech Biometric Based Attendance System”, IEEE Twentieth National Conference on Communications (NCC), 2014, Page(s) 1 – 6.