ASIAN JOURNAL OF PHARMACEUTICAL AND CLINICAL RESEARCH

INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

Special Issue (April)

Online - 2455-3891
Print - 0974-2441

Full Proceeding Paper

A GAUSSIAN MIXTURE MODEL-BASED SPEAKER RECOGNITION SYSTEM

KUMARI PIU GORAI*, THOMAS ABRAHAM JV*

Department of Computer Science, School of Computing Science and Engineering, VIT University, Chennai, Tamil Nadu, India.
Email: kumaripiu.gorai2016@vitstudent.ac.in,thomasabraham.jv@vit.ac.in

Received: 23 January 2017, Revised and Accepted: 03 March 2017

ABSTRACT

A human being has lot of unique features and one of them is voice. Speaker recognition is the use of a system to distinguish and identify a person from his/her vocal sound. A speaker recognition system (SRS) can be used as one of the authentication technique, in addition to the conventional authentication methods. This paper represents the overview of voice signal characteristics and speaker recognition techniques. It also discusses the advantages and problem of current SRS. The only biometric system that allows users to authenticate remotely is voice-based SRS, we are in the need of a robust SRS.

Keywords: Speaker recognition, Mel-frequency cepstral coefficients, Gaussian mixture model, Support vector machine, Robust speaker recognition system.

 DOI: http://dx.doi.org/10.22159/ajpcr.2017.v10s1.19596

INTRODUCTION

Speaking for the exchange of information is the most natural human way. The speech signals are identical to a person. Every speaker has his/her own rhythm, intonation, acoustical, and utterance because we all have unique glottal, vocal tract, lexicon, and accent.

Robust speaker recognition is a system to maintain delicate recognition system to provide a correct decision, according to the sound signals/waves which do not depend on the surrounding of speaker who spoke. Basically the task of the speaker recognition classified into two categories, one is speaker identification and another one is speaker verification [1,2]. Speaker identification is a task to determine whether the word spoken from known or unknown set of registered speaker voices (Fig. 1). The speakers can be from either closed/fixed set or open set. Closed set is a set of fixed numbers of registered speaker in a training system. Open set includes any number of registered speakers, and there is a possibility that unknown speaker also present, known as imposter. Imposter means the voice of the person is not belonging from the specific speaker [3].

Speaker verification is a task to check whether or not a voice token belongs to a specific speaker. The result of the speaker verification is either accepted or rejected (Fig. 1).

The speaker recognition system (SRS) can be either one of the following category:

1. Text-dependent - The utterance which is spoken is predefined; speaker speaks the same phrases, i.e., speaker speaks the same sentence during both enrollment and testing.
2. Text-independent - The utterance which is spoken may vary during the enrollment and testing.

Challenges for an SRS

Some of the real-time applications of speaker recognition are transaction authentication, access control, information retrieval, and forensics. However, a good SRS has to address the following challenges:

- The main disadvantage is that the sound/speech varies from time to time, or it would be different when we speak right before waking up from sleep, or we get aged or we sound a bit different voice due to illness. All these may confuse the SRS.
- The performance of the SRS depends on the quality of speech signal.
- Less accurate because the speech signals are influenced by environmental noise.

METHODS

Doddington [2] describes the application of the SRS. It is used for personal identification, security software, physically access control, etc. This paper discusses about the limits of speaker by their characteristics of the oral, jaw, tongue, lips, and velum. In this system, Texas Instruments check the user voice for verification to control physical entry. In the entry booth, the user gives the user ID into keypad if it matches the system allows him. This voice record is stored in bytes stream in the table, and the voice stream is matched with the table according to the user ID. Every word is compositions of six frames template of size 20-30 ms.

The performance of the system is good for most of the users but for some cases its performance is not good.

Kinnunen and Li [4] mainly discuss about the features extracted from the speech signals. Features can be categorized as short-term spectral features, voice source features, spectro-temporal features, prosodic features, and high-level features.

- Short-term spectral features - speech signals are fragmented into frames of 20-30 ms duration and features are extracted when the speech signals are in stationary condition. Mel-frequency cepstral coefficients (MFCCs) feature is one of them.
- Voice source features - it is characterize from the glottal excitation, vocal tract, and physical structure of the mouth
- Spectro-temporal features is based on the signals formant transitions and energy modulation
- Prosodic features - depend on the non-segmentable characteristics
- High-level features - depend on the lexicon of the voice.

This paper also discussed about vector quantization (VQ) speaker model. The VQ model is an example of text-independent model. It is used for computational speedup techniques. Another approach was Gaussian mixture model (GMM) model. The GMM is a composition of fixed mixture of multivariate Gaussians components. We can also say that it is extension of VQ. Support vector machine (SVM) is a popular verification system. Generalization performance of this system is also good to determine unseen data, which is one reason for the popularity of SVM. This paper gives the result of NIST 2008 speaker recognition

evaluation used by short-term spectral features. It is also work with GMM model for more accurate quality of the speaker recognition.

DISCUSSION

Automatic speaker recognition is mainly used as a biometric authentication system in business application. These systems are used in industries, national laboratories, and universities. Two main phases of this system is (a) enrollment system (training phase) and (b) testing phase. In enrollment phase, [5] a specific target model is determined using the features of a voice signal as shown in Fig. 2. Moreover, in testing phase (also called identification/verification phase), unknown speaker's voice sample is matched with the target speaker model shown in Fig. 2.

Features of speaker recognition

The features of a voice signal are extracted using some feature extraction techniques. The operation of a feature extraction typically consists of two level of information: (a) Low-level information and (b) high-level information. The low-level information carries the physical structure of the speech. The high-level information depends on the speaker's speaking behavior such as intonation, accent, and excitation. The two popular low-level features are as follows:

a. Linear predictive cepstral coefficients (LPCC)
b. MFCCs.

LPCC

LPC is a signal analysis technique for evaluating the parameters of voice signal. In LPCC, the voice samples define the linear combination of past voice samples. LPC coefficients are fixed the sum of squared differences between the actual voice samples and predicted samples. LPC represents an exact estimate of speech feature vectors. This technique derived from the psychoacoustic properties of the human ear and modeled by filter bank. The steps involving in LPC feature extraction are frame blocking, windowing, and autocorrelation analysis, after that LPC analysis gives the LPCC.

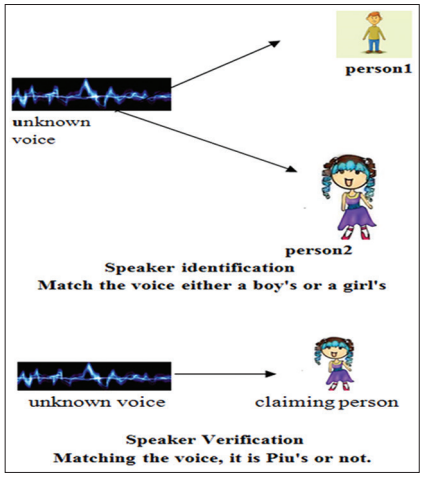

Fig. 1: Speaker identification and verification

MFCC

MFCC is a feature which is used widely in SRS. Davis and Mermelstein introduced MFCC in 1980. The sequence of the steps of extracting MFCC [6,7] are shown in Fig. 3:

- Pre-emphasis – When human voice is recorded, the frequencies of the voice waveform are dissimilar, so increase the small frequency and decrease the high frequency of the voice wave.
- Framing – The voice wave form is chunked into a fixed time duration fragments called frames. The length of each frame is between 20 ms and 30 ms and overlapped by every 10 ms.
- Windowing – To smoothen the framed signal, each frame is multiplied by a windowing function (such as hanning, hamming, and rectangular window). For SRSs, hamming window is the widely used windowing function.
- Mel-filter bank – Applying a mel scale makes our features match more closely to what humans hear. It is splitting the frequency and takes the central frequency.
- Discrete cosine transform – The filter bank energies are correlated with each other because filter banks are overlapped. The discrete cosine transform decorrelates the energies so that diagonal covariance matrices can be used to model the features.

Speaker models

The aim of the speaker modeling is to build a model that can take the variations in a set of features extracted from a given speech and represent the speaker. Speaker models can be categorized as: (1) Generative model or (2) discriminative model. Generative models are based only on the targeted speaker and capturing the speech of the spoken speaker not others, and discriminative model based on both two things, one is targeted speaker and the imposter model.

GMM is one of the popular generative speaker models and SVM is a discriminative speaker models.

a. GMM – GMM is a generative speaker model because GMM model typically involves capturing data from targeted speaker. It is unique model which has *de facto* reference method and generally used for robust SRS using short-speech statement. It has better advantage than other models because the training is quite fast and can be scaled and update the system to add new speakers with relative ease. A GMM model is composed of limited mixture of multivariate Gaussian components; it is a collection of several spectral features that are valid for deigning a speaker model for a targeted speaker. Suppose a speaker has 2 or 3 utterance, and from each utterance, we extract D-dimensional features then how we connect all the features in one model? The MFCC features of each speaker are represented by Gaussian Mixture Model. MFCC coefficients are used for extracting features and minimum processing time in GMM is 10 ms for speech utterance. The parameters for GMM model is mean

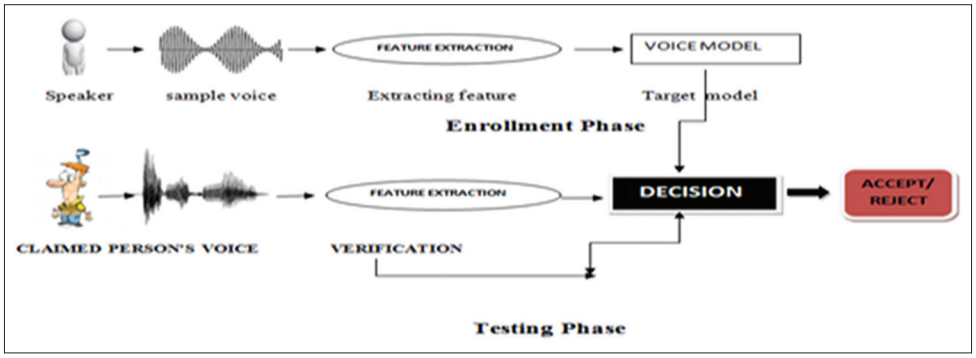

Fig. 2: Speaker recognition system

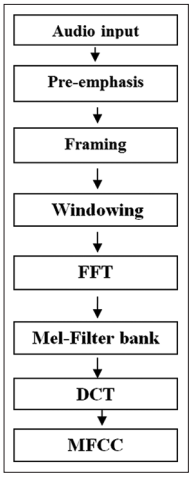

Fig. 3: Mel-frequency cepstral coefficients feature extraction procedure

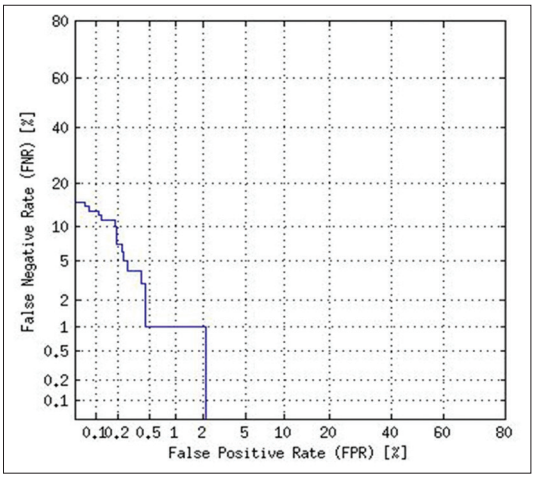

Fig. 4: Detection error trade-off curve of universal background model Gaussian mixture model-based speaker recognition system

vectors, densities (is a sum of M numbers component density), and covariance matrices [3].

b. SVM – SVM is a discriminative speaker model which based on targeted speaker as well as imposter speaker. It is powerful speaker technique model which gives high-standard performance even with very less utterance or data. It maps inputs in high-dimensional space [8]. Both low-level features and high-level features are used with the SVM model. Moreover, it is also working with the GMM model to find out more superior results of the recognition. The principal of the SVM is to fix a boundary line to distinguish or separate between two classes one of them is target speaker model and other is imposter model and for the boundary SVM uses hyperplane. Each point around the hyperplane is called support vectors [9].
c. VQ – VQ is a technique to produce vector region from a large number of vector space in that space. Every region is a cluster and its center is called centroid. A group of all such centroids forms a codebook. The codebook is a collection of code word in VQ.

Even though the codebook is smaller than the original sample, it still accurately represents a person's voice characteristics. Since the amount of data is significantly less, it reduces the amount of computations needed for comparison in later stages.

EXPERIMENT

The experiment was conducted with a clean speech dataset, TIMIT. For speaker recognition, the MSR identity toolkit was used. The MSR identity toolkit is user-friendly and convenient to use with little modifications. The TIMIT dataset consists of 630 speakers of which 192 are female and 438 are male. For background model training, we took 530 speakers and for testing took 100 speakers (30 female and 70 male). For each speaker, we have ten short sentences in TIMIT. To train the background model, all ten sentences from 530 speakers (i.e., 5300 speech recordings in total) were used in this research. To create speaker-specific model training, we used 9 out of 10 sentences per speaker and keep the remaining 1 sentence for tests. Verification trials consist of all possible model-test combinations, making a total of 10,000 trials (100 target vs 9900 impostor trials). The MFCC features of each speaker are represented by GMM (Fig. 4).

PROPOSED SYSTEM

The GMM approach gives better result if the speech signal is clean, and the performance is highly degraded for a noisy speech data [6,10]. May *et al.* [8] showed that the usage of a universal background model (UBM) in combination with missing data recognition yields substantial improvements in recognition performance, especially in the presence of highly non-stationary background noise at low signal-to-noise ratio. Alam *et al.* [11] suggested a robust feature extraction technique based on filter bank. The future work is to find a robust feature extraction technique using Hilbert transform for a noisy data so that the performance of an automatic speech recognition would be improved.

CONCLUSION

In this paper, an overview of an SRS was discussed. The main and common strength of speaker recognition is very easy to use, and the weakness is the voice signals of a person changes time to time. The experiment result of a baseline UBM-GMM system was presented and the results are promising. However, it may not be case for speech signal in a noisy environment, and hence, we need a better robust SRS.

REFERENCES

1. Campbell JP. Speaker recognition: A tutorial. Proceedings of the IEEE. Vol. 85. No. 9. September; 1997.
2. Doddington GR. Speaker recognition—Identifying people by their voices. Proc IEEE 1985;73;1651-64.
3. Kenny P, Boulianne G, Ouellet P, Dumouchel P. Speaker and session variability in GMM-based speaker verification. IEEE Trans Audio Speech Lang Process 2007;15(4):1448-60.
4. Kinnunen T, Li H. An overview of text-independent speaker recognition: From features to supervectors. Speech Commun 2010;52:12-40.
5. Togneri R, Pullella D. An overview of speaker identification: Accuracy and robustness issues. IEEE Circuits and Systems Magazine Second Quarter. 2011. p. 23-60.
6. Liu G, Lei Y, Hansen JH. Robust feature front-end for speaker identification. In: Proceeding ICASSP. Kyoto, Japan: March; 2012. p. 4233-6.
7. Campbell W, Campbell J, Reynolds D, Singer E, Carrasquillo PT. Support vector machines for speaker and language recognition. Comput Speech Lang 2006;20(2-3):210-29.
8. May T, van de Par S, Kohlrausch A. Noise-robust speaker recognition combining missing data techniques and universal background modeling. IEEE Trans Audio Speech Lang Process 2012;20(1):108-21.
9. Solomonoff A, Campbell WM, Boardman I. Advances in channel compensation for SVM speaker recognition. In: Proceeding ICASSP. 2005. p. 629-32.
10. Reynolds D, Rose R. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans Speech Audio Process 1995;3(1):72-83.
11. Alam MJ, Kenny P, Shaughnessy DO. Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique. Digit Signal Process 2014;29:147-57.