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 REVIEW ON MULTIMODAL SPEAKER RECOGNITION

KHADAR NAWAS K*

School of Computing Science and Engineering, VIT University, Chennai Campus, Tamil Nadu, India. Email: khadarnawas.k@vit.ac.in

Received: 29 March 2017, Revised and Accepted: 03 April 2017

ABSTRACT

A review on multimodal speaker recognition (SR) is being presented. For many decades, the SR has been studied, and still, it has grabbed the interest of many researchers. SR includes two levels - system training and system testing. The robustness of the SR system depends on the training environment and testing environment as well as the quality of speech. Air conducted (AC) speech is a source, from which speaker is recognized by extracting the features. The performance of the SR system depends on AC speech. Further to improve the robustness and accuracy of the SR system various other sources (modals) like throat microphone, bone conduction microphone, array of microphones, non-audible murmur, non-auditory information like video are used in complementary with standard AC microphone. This paper is purely a review on SR and various complimentary modals.

Keywords: Speaker recognition, Multimodal speaker recognition, Throat microphone, Bone microphone, Vector quantization, Gaussian mixture model.

© 2017 The Authors. Published by Innovare Academic Sciences Pvt Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4. 0/) DOI: http://dx.doi.org/10.22159/ajpcr.2017.v10s1.19761

INTRODUCTION

Speaker recognition (SR) is one of the particular methods which help in identifying or verifying a speaker from his vocal speech. SR is used in different strategies such as transactions authentication, remote access control through telephones, speech forensic, and personalization of electronic gadgets. SR system is two typed as speaker verification system and speaker identification system. Speaker verification is aimed to claim speaker's identity from a set of speakers using speech signal. The speaker identification method involves identification of the speaker is being enrolled within a group of persons. SR may be classified as text dependent and text independent. According to text-dependent SR, the speaker has to speak a defined text during training and testing. In the format of text independent SR, the speaker could make random speak [1]. SR can be divided into different forms they are digitized speech data acquisition, feature extraction, speaker modeling, and classification. The air conducted (AC) speech signal is influenced more by the background noise, TC speech and bone conducted (BC) speech are having less influence of the background noise. This review paper is structured based on multimodal SR as sections 2,3 and 4. In 2nd section, we discuss about features and feature extraction. The 3rd section discusses about speech enhancement in 4th speaker models section.

FEATURES AND FEATURE EXTRACTION

Speech signal is a specification which features high-level information and low-level information. The information contained in speech signal identified as tone, silence rate and speech patterns, peculiar terms usage, and peculiar pronunciations of a speaker. Spectral characteristics of the speech are the low-level information. Human beings are good at extracting such higher level features. SR system mainly centered in extracting low-level spectral features from spectral from speech signals. From literature, it is proven that the spectral features cultivated from speech signals are the most efficient features used by many automatic SR systems [2].

MULTIMODAL SR SYSTEM

A complementary source with speech is used in the multimodal SR system. It has been studied that by adding complementary source increases the robustness of the SR system [3] following are the complimentary source used in literature, throat microphone (TM), bone conduction microphone, array of microphones, and video. Various feature extraction techniques used in where discussed.

TM

TM is a transducer which senses the vibration of the skin place near in contact with the larynx. The speech recorded is intelligible, void of noise due to the air vibrations [3]. The TM is a machine which not only captures vibrations but also jointly works with the voice reeds while during lung air evictions. The correct speech is resulted due to the resonating vocal cavity and by proper placement of TM at the larynx closure near the region of pharynx. The TM captures speech signals with low-frequency resulted due to voice reed vibrations that are filtered through the muscles of the throat [3]. It is found that TM speech is similar with normal microphone (NM) speech, TM and NM speech features are extracted. In Kinnunen and Li's study [4], an overview of features extracted from speech is presented. The most popular feature extractor based on literature is described here. Linear predictive coefficient (LPC) is the model which depicts the voice reeds using all-pole model. LPC coefficients are a model which indulges resonance property of vocal tract. Mel frequency Cepstral coefficients (MFCCs) are also relevant to cochlea. Delta MFCC is the model which accepts the speech transition property from the speech sound. Perceptual LPC coefficients are subjected over a short term spectrum from speech [5].

Real cepstral coefficients (RCC)

The frequency components from the speech signal are obtained by transforming it to frequency field from time as given in in equation 1. The resultant feature is the RCC [5].

$$\text{Real cepstral}=\text{IFFT}(\log(\text{FFT}(s(n)))) \quad (1)$$

FFT - Fast Fourier transform

IFFT - Inverse Fourier transform.

MFCC

MFCC is formulated on the Mel scale that is linear between 0-10 KHz and above 1 KHz it is log scale. In this method for each frame after Fourier transform is fed through a filter bank that is not in uniform space in frequency. The log of the signal is calculated as in (2).

$$\text{Mel}(f) = 1000/\ln(1+10/7)*\ln(1+f/700) \quad (2)$$

Fig. 1 depicts how MFCC coefficients are calculated from various steps. The first and the foremost step is how Fourier transform of

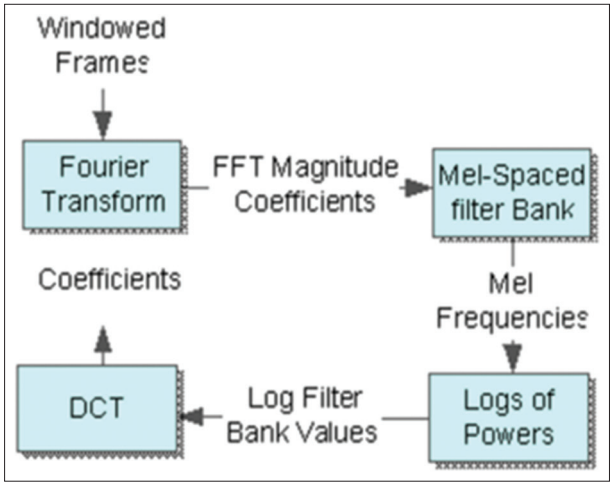

Fig. 1: Mel frequency Cepstral coefficients feature extraction

each frame of signal is computed. Second step computes, the power of the spectrum which is mapped with Mel scale. Third step depicts, each Mel frequencies as the logs of the power is taken and calculated accordingly. Finally, discrete cosine transform is computed from the Mel frequencies [5].

Delta-MFCC

In Delta-MFCC, it has certain subjections of MFCC which reflects the static characteristics of the signal, which subjects that human hear is more responsive to the static and dynamic characteristics of a signal. The delta MFCC is appended to MFCC to reflect the dynamic information [5].

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (3)$$

Where in equation 3, dt is delta coefficient, t is frame computed in terms of the static coefficients c_{t+n} to c_{t-n} and N is 2.

Linear predicition coefficients

The voice tract resonance property is represented in the LPC features, LPC models the voice reed by all-pole model. By a linear combination of the past P samples linear predictive coding estimates the signal, s (n).

$$s(n) = \sum_{k=1}^{P} a_k s(n-k) + e(n) \quad (4)$$

Where e (n) is the prediction error and ak values are linear prediction coefficients [6].

Linear prediction cepstral coefficients

The cepstral coefficients can be calculated from the LPC parameters through recursive procedure as given below.

C1 = a1

$$C_n = a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right)_k c[k] a_{n-k} \quad \text{for } 1 \le n \le p$$

$$C_n = \sum_{k=1}^{n-1} \left(\frac{n-k}{n}\right)_k c[n-k] a_k \quad \text{for } n > p \quad (5)$$

As by equation 5, it is subjected that as a1, ap which specifies p-order LPC feature vector, cn, n=1, p as the coefficients, and p as the first p values of the cepstrum.

Weighted dynamic MFCC

In Yamasaki and Shimamura's study [12], a new feature is propose, weighted dynamic MFCC is series of coefficients obtained by combining traditional MFCC and dynamic MFCC. Weighted dynamic MFCC is computed as shown in the equation.

New MFCC = MFCC + a.ΔMFCC + b.Δ2MFCC

Where new MFCC is the weighted dynamic MFCC, ΔMFCC is the first order delta MFCC, Δ2MFCC is the second order delta MFCC, a and b are their weights respectively.

SPEECH ENHANCEMENT

In this speech enhancement, the voice signal is transmitted through different formats they are (i) AC in the normal path, (ii) as vibration along voice reeds, and (iii) skull bone through cochlea [7]. BC speech does not affected by the noise due environment. In BC speech, the high frequency components are filtered by the speaker's body due to lack in high frequency component the intelligibility of speech is low. It can be used as supplement with the AC speech to enhance the accuracy of the SR system by enhancing the speech. Placement of the BCM influences intelligibility of the speech, literature studies show the various location to place the BCM [8-10]. Locations near the larynx result in higher intensity, while locations near the temple result in higher intelligibility [9]. In Tsuge *et al.*'s study [11], it specifies proposed speaker verification using AC speech and BC speech which indulges in reducing the ERR of AC speech by 16% and ERR of BC speech about 71.7%. Through literature study it is found that no special features or feature extraction techniques were proposed for the BC speech. By similarity, the BC speech with the AC speech features an extracted as like AC speech [11]. When BC speech enhances with spectral subtraction technique it helps in improving the SR system accuracy in noisy environment which is subjected relative to the AC speech signal enhanced by spectral subtraction method [12,13].

SPEAKER MODELS

SR involves feature extraction, Speaker modeling (Training), speaker testing and decision making using scoring techniques as depicted below figure. There are different speaker modeling techniques in practice, they are vector quantization (VQ), Gaussian mixture model (GMM) and support vector machine (SVM).

VQ

VQ is the form which is subjected as the method of lossy compression which works on block coding [14]. VQ is applied for classifying the speakers in SR. It is the process of converting a larger set of feature vector in to a smaller set of the feature vector as the centroids of the distribution. The feature vectors are clustered into set of code books. The clustering of vector is subjected by famous old form of algorithm that is K-means algorithm; it is used to create the speaker code book. The procedure of K-means algorithm works is that it creates M centroids from T feature vectors. Every new features are allocated to the nearest centroid, the modifies centroids are used to make new clusters and it continues until the mean square error between the feature vectors and the cluster centroid is below the threshold [15,16]. In VQ SR, the test speaker's (Unknown) feature vector is compared with the generated code books. The distortion measure is computed by matching the features with the code book. The minimum distortion measure results in the identified speaker from the set of trained speakers. The distortion measure is specified by Euclidean distance measured in between the points such as P = (P1,P2,...Pn) and Q(Q1,Q2,...Qn) which specifies that sum of the squared distance between the points P and Q.

$$\sqrt[2]{\left((P1-Q1)^2\right)+\left(P2-Q1)^2\right)+\ldots+\left(Pn-Qn)^2\right)}$$

GMM

GMM is a statistical model which specifies the density function with probability parameters and they are given as a weighted sum of

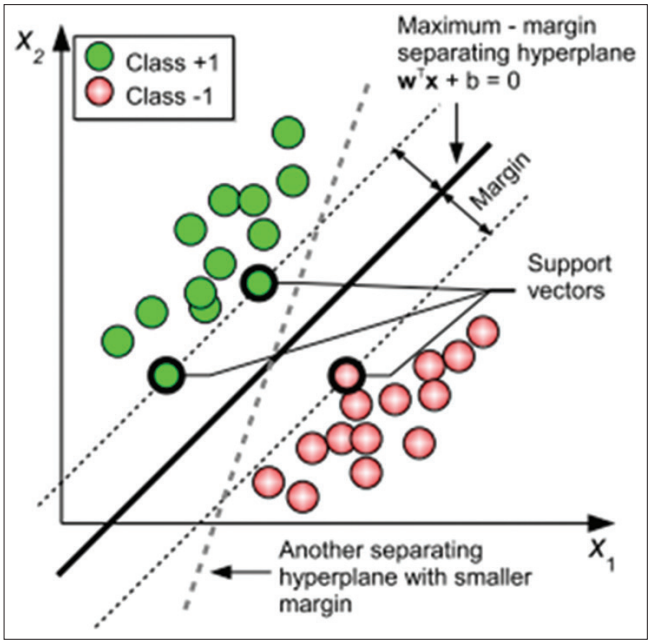

Fig. 2: Support vector machine classification

Gaussian component densities [17]. The feature vectors from the speakers are modeled using Gaussian mixture densities. The mixture density for a speaker by specifying D as dimensional feature.

∑ vector is defined as

$$p(x|\lambda)=\sum_{i=1}^{M} p_i^s\, b_i^s(x) \quad \text{where} \quad \sum_{i=1}^{M} p_i^s = 1 \tag{6}$$

The density for the weighted linear combination with M component Gaussian densities.

$$b_i^s(x)=\frac{1}{(2\pi)^{D/2}|\sum|^{1/2}} X \exp\{-1/2(x-\mu_i^s)'(\sum_i^s)^{-1}(x-\mu_i^s)\}$$

The mixture weights p_i^s, satisfies the condition ∑i = 1 m p_i^s = 1. Using expectation-maximization algorithm used iteratively to calculate the maximum likelihood as measure to identify the speaker. Single-state hidden markov model can be viewed with GMM density probabilities.

SVM

SVM is classifier adopted in SR. It can be applicable for spectral, prosodic and high-level features. SVM combined with GMM resulted in high accuracy. SVM is a factor which is subjected as binary classifier that models a decision boundary formulates a hyper-plane between two classes as depicted in Fig. 2. The hyper-plane with maximum-margin that separates the positive and negative.

The SVM function is given in equation 7.

$$f(x)=\sum_{i=1}^{N}\alpha_i t_i K(x,x_i)+d \tag{7}$$

Here $t_i \epsilon \{+1, -1\}$ specified as the ideal output values, $\sum_{i=1}^{N}\alpha_i t_i = 0$ and $\alpha_i > 0$. The support vectors such as x_i, and their corresponding weights αi and the bi is specified in terms of d. K is given as kernel function which maps the kernel feature to higher dimensionality.

CONCLUSION

The accuracy of the SR system depends on the quality of the speech. The speech signals are influenced by the environmental noise. Hence, different mode of collected speech (i.e.,) throat speech and BC speech along with the AC speech are used together to increase the accuracy of the SR. In this paper, we have discussed about the features, feature extraction methods and speaker modeling methods applicable for all kind of speech signals.

REFERENCES

1. Campbell JP Jr. Speaker recognition: A tutorial. Proceedings of the IEEE. Vol. 85. No. 9. September; 1997.
2. Zanuy MF, Moreno EM. State-of-the-art in speaker recognition. IEEE Abre Systems Magazine, May; 2005.
3. Mubeen N, Shahina A, Khan AN, Vinoth G. Combining spectral features of standard and throat microphones for speaker identification. International Conference on Recent Trends in Information Technology, ICRTIT; 2012. p. 119-22.
4. Kinnunen T, Li H. An overview of text-independent speaker recognition: From features to supervectors. Speech Commun 2010;52(1):12-40.
5. Sapijaszko GI, Mikhael WB. An overview of recent window based feature extraction algorithms for speaker recognition. Midwest Symposium on Circuits and Systems. 2012. p. 880-3.
6. Ramachandran RP, Farrell KR, Ramachandran R, Mammone RJ. Speaker recognition—general classifier approaches and data fusion methods. Pattern Recognit 2002;35:2801-21.
7. Rahman MS, Shimamura T. A Study on Amplitude Variation of Bone Conducted Speech Compared to Air Conducted Speech; 2013.
8. McBride M, Tran P, Letowski T, Patrick R. The effect of bone conduction microphone locations on speech intelligibility and sound quality. Appl Ergon 2011;42(3):495-502.
9. Srinivasan S, Kechichian P, Sriram I. Robustness Analysis of Speech Enhancement using a Bone Conduction Microphone - Preliminary Results. September, 2012. p. 4-6.
10. Tran P, Letowski T, McBride M. Bone conduction microphone: Head sensitivity mapping for speech intelligibility and sound quality. ICALIP 2008 - 2008 International Conference on Audio, Language and Image Processing, Proceedings. 2008. p. 107-11.
11. Tsuge S, Koizumi D, Fukumi M, Kuroiwa S. Speaker verification method using bone-conduction and air-conduction speech. ISPACS 2009 - 2009 International Symposium on Intelligent Signal Processing and Communication Systems, Proceedings, (Ispacs). 2009. p. 449-52.
12. Yamasaki N, Shimamura T. Accuracy Improvement of Speaker Authentication in Noisy Environments Using Bone-Conducted Speech; 2010. p. 197-200.
13. Weng Z, Li L, Guo D. Speaker recognition using weighted dynamic MFCC based on GMM. Proceedings - 2010 International Conference on Anti-Counterfeiting, Security and Identification. ASID; 2010. p. 285-288. DOI: 10.1109/ICASID.2010.5551341.
14. Gray RM. Vector quantization. IEEE ASSP Magazine. 1984. p. 4-29.
15. Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. Pattern Recognit 2003;36(2):451-61.
16. Khan SS, Ahmed A. Cluster center initialization for K-means algorithm. Pattern Recognit Lett 2004;25(11):1293-302.
17. Cherifa S, Messaoud R. New technique to use the GMM in speaker recognition system (SRS). International Conference on Computer Applications Technology. 2013. p. 1-5.