ASIAN JOURNAL OF PHARMACEUTICAL AND CLINICAL RESEARCH

INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

Vol 9, Suppl. 3, 2016

Online - 2455-3891
Print - 0974-2441

Review Article

A REVIEW ON VOICE ACTIVITY DETECTION AND MEL-FREQUENCY CEPSTRAL COEFFICIENTS FOR SPEAKER RECOGNITION (TREND ANALYSIS)

P. MAHALAKSHMI*

School of Electrical Engineering, VIT University, Vellore - 632 014, Tamil Nadu, India. Email: maha_50@yahoo.com

Received: 27 July 2016, Revised and Accepted: 03 October 2016

ABSTRACT

Objective: The objective of this review article is to give a complete review of various techniques that are used for speech recognition purposes over two decades.

Methods: VAD-Voice Activity Detection, SAD-Speech Activity Detection techniques are discussed that are used to distinguish voiced from unvoiced signals and MFCC- Mel Frequency Cepstral Coefficient technique is discussed which detects specific features.

Results: The review results show that research in MFCC has been dominant in signal processing in comparison to VAD and other existing techniques.

Conclusion: A comparison of different speaker recognition techniques that were used previously were discussed and those in current research were also discussed and a clear idea of the better technique was identified through the review of multiple literature for over two decades.

Keywords: Cepstral analysis, Mel-frequency cepstral coefficients, signal processing, speaker recognition, voice activity detection.

 DOI: http://dx.doi.org/10.22159/ajpcr.2016.v9s3.14352

INTRODUCTION

Speaker recognition is very important for various voice based applications in security and monitoring systems, also these days such techniques are used in home appliances for user controlled switching of devices. Voice activity detection (VAD) is used to detect voice from silence and Mel-frequency cepstral coefficients (MFCC) is used to extract features from the voice signals. Usage of VAD has been application centric, for speech recognition precise end points for zero crossing rating are required [1]. Further, new cepstral based algorithm for VAD proved to be more efficient [1]. In few findings, we see that vector quantization and VAD have been used along with MFCC to increase the efficiency of speaker recognition [2]. Speaker recognition rates are also controlled with the help of Mel-frequency delta phase (MFDP) along with MFCC and it is found that error probability is less in MFCC, but when both MFCC and MFDP are used together, it proves to be more efficient [3]. A review of the use of phase information in speech processing, however, indicates that broadly effective phase domain features remain difficult to extract [4]. To detect nonspeech recordings, a technique called speech activity detection is used, where speech and voice are different entities. SAD uses statistical properties of speech parameters such as: Energy, pitch, and entropy. Therefore, the performance of different SAD is different and varying according to the level and type of signal-to-noise ratio (SNR). As a result, the performances of different speech based systems are significantly sensitive to the SAD technique. Therefore, SAD should be carefully chosen while designing a speech based system. SAD is rigorously studied for speech recognition, speech coding, etc. However, it is not so far thoroughly studied for speaker recognition applications. Recently, it has drawn the attention of the researchers in this area [5]. Some have used both MFCC and VAD to detect speaker, using a novel approach by changing VAD algorithm and proved better results in their technique [6].

METHODS

VAD usage for speaker recognition

Traditional methods for VAD

A typical VAD algorithm involves following steps [7].

Parameterize the input signal

Time domain or spectral domain based features such as energy, zero crossing rate, spectral shape, and cepstral coefficients are extracted from the audio signal.

Make the first VAD decision

The decision is made whether a given segment is speech or silent which is generally done frame wise. Decision rules, statistical models, or adaptive thresholds are some important methods useful in such decisions. Estimating the current SNR or determining the noise type is also involved in this step.

Smooth the final VAD decision

Since speech is highly correlated, if the current frame is speech, the next frame is also likely to be speech. Typical VAD algorithms filter and purify the first VAD decision to reduce frequently occurring transitions from speech to silence. Estimates of SNR and other running averages, etc., are illustrated in Fig. 1.

Statistical methods for VAD

Other methods for VAD use statistical models to differentiate between speech and silence. One such approach assumes that the statistics of nonspeech (silence) is stationary over a longer period than that of speech [8]. To make a VAD decision, the statistics of the current frame are compared with the estimated noise statistics. In another approach, a ratio test is applied, assuming a complex Gaussian distribution is preceded by each spectral component of speech and silence [9,10]. The two competing hypotheses in this case are as follows: Given the spectral component X(k) of a frame of noisy speech, we have,

H_0: X(k)=N(k)

H_1: X(k)=N(k)+S(k)

Where, k is the index into the discrete Fourier transform coefficient; N, S, and X represent noise, speech, and noisy speech, respectively. Another method models speech and nonspeech as Laplacians instead

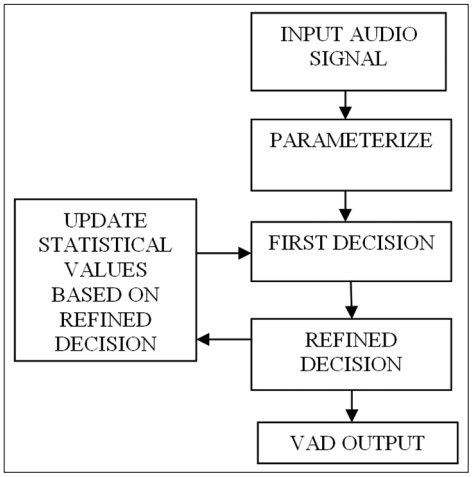

Fig. 1: Flowchart of typical voice activity detection algorithm

of Gaussians [11]. Some statistical methods for VAD make use of the observation that the higher order statistics (HOS) of speech are different from the silence. The silent part of speech is assumed to be Gaussian, and the HOS is used to distinguish speech from silence [11].

Improved VAD using thresholds

It compares the extracted features from the input speech signal with some predefined threshold. Voice activity exists if the measured feature values exceed the threshold limit. Otherwise, silence is assumed. The performance of the VAD depends heavily on the preset values of the threshold for detection of voice activity. The VAD proposed using thresholds works well when the energy of the speech signal is higher than the background noise, and the background noise is relatively stationary which is common in many applications. The amplitude of the speech signal samples is compared with the threshold value which is being decided by analyzing the performance of the system under different noisy environments.

Cepstral analysis in speaker recognition

MFCC is probably the best known and most widely used technique for both speech and speaker recognition [12,13]. The Mel scale was developed as a result of auditory perceptual experiments. A Mel is a unit of measure based on human ear's perceived frequency. The Mel scale has approximately linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The information carried by low-frequency components of the speech signal is more important compared to the high-frequency components. To place more emphasis on the low-frequency components, Mel scaling is performed. Mel filterbanks are nonuniformly spaced on the frequency axis, so we have more filters in the low-frequency regions and less number of filters in high-frequency regions. Following are the methods used for speaker detection using MFCC and its combination with other techniques.

MFCC implementation

A Hamming window is applied at a frame size of 25 ms and a frame shift of 10 ms on the speech signal. The windowed speech frame is preemphasized and converted into the frequency domain. The frequency scale is warped using the bilinear transformation.

$$\omega \text{ warped} = \omega + 2 \arctan F\omega \sin \omega \, 1 - F\omega \cos \omega$$

where, the constant Fω, varying from 0 to 1, controls the amount of warping. A bank of filters whose center frequencies are distributed uniformly between (minf, maxf) along the warped frequency axis forms the filterbank. The filter shape varies from rectangular to triangular, and this is controlled by the constant Fshape that varies from 0 to 1, with 0 corresponding to triangular and 1 to rectangular. The filterbank energies are computing by integrating the energies in each filter, and a DCT is applied to the logarithm of the filterbank energies to convert them to cepstra. The parameters used for the MFCC extraction are as follows [14]:

- Minimum frequency in Hertz minf=0
- Maximum frequency in Hertz maxf=3500
- Warping factor Fω=0.2
- Number of filters R=40
- Filter shape Fshape=0.4
- Number of cepstral coefficients Nc=18.

Linear predictive coding (LPC) cepstra (LPCC)

The LPC technique approximates a given speech sample as a linear combination of the past P samples [15,16]. In the frequency domain, LPC fits an all-pole model to the short time spectrum. The predictor coefficients ak can be used recursively to form the linear predictive cepstral coefficients cm:

127 c(0)=ln σ 2 (A.1)

c(m)=am+Xm−1 k=1 k m ! ckam−k 1 ≤m≤P (A.2)

c(m)=Xm−1 k=1 k m ! ckam−k m>P (A.3)

Where, σ 2 is the gain of the model.

F0 prediction from LPCC

A fundamental frequency prediction method which is used primarily in the voice conversion system [17]. The experimental results show that there is a relatively stable mapping relationship between LPCCs and fundamental. We find that the tone curve could be mapped by LPCC features, and the mean F0 also could be predicted well when the number of speakers of the training dataset is enough. Subjective tests certify that the tone could be understood well. This F0 prediction method could be utilized to predict pitch in whispered speech conversion system and voice conversion system.

Acoustic and para-verbal indicators of persuasiveness

As we know p*e*rsuasive measure in speech can help us identify the speaker. Across all groups, MFCC descriptors that emphasize lower frequency regions, especially MFCC 2 and MFCC 4-stood out for predicting persuasive speakers in both positive and negative reviews [18]. This information is helpful to increase *the* efficiency of speaker recognition system also.

MFCC with vector quantization

Feature extraction involves finding MFCCs of the speech and vector quantizing them to obtain the speaker specific codebook. For this short time spectral analysis is used, FFT, reference model (speaker) similarity input speech feature extraction verification result (accept/reject) decision speaker ID threshold 22 windowing, Mel spaced filter banks, and convert the speech signal to a parametric representation, i.e., to MFCC. These MFCCs are based on the known variation of the human ear's critical bandwidths with frequency, i.e., linear at low frequencies and logarithmic at high frequencies. These are less susceptible to the variation in speaker's voice. The extraction of MFCC is shown in Fig. 2.

Vector quantization is the process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented using its centroid. Centroids of all clusters are combined to form the speaker-specific codebook. In feature matching, the input utterance of an unknown speaker is converted into MFCCs and then the total VQ distortion between these MFCCs and the codebooks stored in our database is measured. VQ distortion is the distance from a vector to the closest

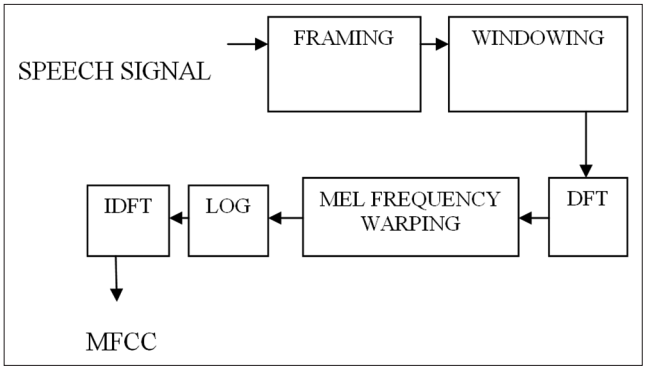

Fig. 2: Mel-frequency cepstral coefficients extraction

code word of a codebook. Based on this VQ distortion, we decide whether the speaker is a valid person or an impostor, i.e. if the VQ distortion is less than the threshold value, then the speaker is a valid person, and if it exceeds the threshold value, then he is considered as an impostor. This system is at its best roughly 80% accurate in identifying the correct speaker.

DIFFERENT IMPLEMENTATIONS OF MFCC AND OTHER TECHNIQUES

The performance of the MFCC may be affected by 1 the number of filters, 2 type of window. In this paper, several comparison experiments are brought to light to find the best implementation.

Effect of number of filters

Results of the speaker recognition performance by varying the number of filters of MFCC to 12, 22, 32, and 42 are given. The recognizer reaches the maximal performance at the filter number K=32. Too few or too many filters do not result in better accuracy.

Effect of variation in type of window using 32 filters

Considering 32 filters as a standard number of filter, the window type is changed. Experiments using two windows, viz., Hanning window and rectangular window are done. Results show that efficiency is maximum while using Hanning window [18].

Usage of different *M*el-frequency formulas

After the introduction of MFCC various corrections to the basic idea was done by people in research communities over the world. Some were based on human perception toward nonlinear pitch which significantly increased the efficiency. A clear comparison of different formulas used over time as explained in shows that computationally inexpensive representation of the Mel scale is done by Koenig scale which is linear below 1000 Hz and logarithmic above 1000 Hz, but it is not precise and deviates from the original scale very often. More precise approximation is by 1 [19]:

$$fmel=kconst.\ log\ n[1+flin/Fb] \quad (1)$$

Where, Fb =1000.

$$fmel=(1000/logn2).logn[(1+flin)/1000] \quad (2)$$

Furthermore, the formulation 2 is improved version as values of Mel-frequency remain unaltered by the value of base n in logarithm. Following representations:

$$fmel=2595.log10[1+flin/700] \quad (3)$$

$$fmel=1127.ln[1+flin/700] \quad (4)$$

are used in different MFCC implementations. As we compare the above-mentioned formulae, 3 and 4 provide better approximation of Mel scale frequencies below 1000 Hz as compared to 2 [19].

Implementation of MFCC in music

At present, extracted audio features are mainly divided into two types, namely, static and dynamic features, which are mentioned [20,21]. MFCC is a static feature widely used in music analysis and recognition techniques which are computed on a frame by frame basis (10 ms gap). MFCCs, which are sensitive to broad spectral features while remaining relatively invariant to fine spectral (pitch)structure, are extracted by most of the researchers to represent the spectral shape, which actually reflects the characteristic of the timbre in music [22]. However, as MFFC do not work well as the background music changes.

CONCLUSIONS AND FUTURE WORK

This paper has reviewed MFCC and VAD techniques broadly to provide an easy comparison on different techniques in use currently. Importance of speaker recognition is undeniable, and research on this topic is vast, to gain overview of the current trends in past few years. This paper shows MFCC in combination with different techniques such as voice and SAD and vector quantization. It was shown that Mel-frequency delta phase features extracted purely from the phase domain could be used for distinguishing speech from noise, and also for distinguishing between voices of different people. It is also shown that average use of filters result in greater efficiency and usage of Hanning window is of profit. Using the revised formula as shown in this paper, we can improve speech recognition rates. Furthermore, we can use all above-mentioned conclusion to detect speaker in case of whispered speech. Such detection can be useful for security systems.

REFERENCES

1. Haig JH, Mason JS. Robust Voice Activity Detection Using Cepstral Features IEEE TENCON'93, Beijing; 1993.
2. Nijhawan G, Soni MK. Speaker recognition using MFCC and vector quantisation. Int J Recent Trends Eng Technol 2014;11(1):7-10.
3. McCowan I, Dean D, McLaren M, Vogt O, Sridharan S. The Delta-Phase Spectrum With Application to Voice Activity Detection and Speaker Recognition, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. Vol. 19; 2011.
4. Alsteris LD, Paliwal KK. Short-time phase spectrum in speech processing: A review and some experimental results. Digit Signal Process 2007;17(3):578-616.
5. Sahidullah M, Saha G. Comparison of speech activity detection techniques for speaker recognition. arXiv preprint arXiv:1210.0297; 2012.
6. Geeta N, Soni MK. A new design approach for speaker recognition using MFCC and VAD. Int J Image Graph Signal Process (IJIGSP) 2013;5(9):43-9.
7. Ramırez J, Segura JC, Benıtez C, De La Torre A, Rubio A. Efficient voice activity detection algorithms using long-term speech information. Speech Commun 2004;42(3):271-87.
8. Srinivasan K, Gersho A. Voice Activity Detection for Cellular Networks, In: Proceedings IEEE Workshop Speech Coding for Telecommunications; 1993. p. 85-6.
9. Sohn J, Kim NS, Sung W. A statistical model-based voice activity detection. IEEE Signal Process Lett 1999;6(1):1-3.
10. ChoYD, Kondoz A. Analysis and improvement of a statistical model-based voice activity detector. IEEE Signal Process Lett 2001;8(10):276-8.
11. Gazor S, Zhang W. A soft voice activity detector based on a Laplacian-Gaussian model. IEEE Trans Speech Audio Process 2003;11(5):498-505.
12. Srinivasan A. Speaker identification and verification using vector quantization and Mel frequency cepstral coefficients. Res J Appl Sci Eng Technol 2012;4(I):33-40.
13. Tiwari V. MFCC and its applications in speaker recognition. Int J Emerg Technol 2010;1(I):19-22.
14. Enqing D, Guizhong L, Yatong Z, Xiaodi Z. Applying Support Vector Machines to Voice Activity Detection, In 6th International Conference Signal Process. Vol. 2. IEEE, 2003. p. 1124-7.
15. Cooke M, Green P, Josifovski L, Vizinho A. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun

2001;34(3):267-85.

16. Kanedera N, Arai T, Hermansky H, Pavel M. On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Commun 1999;28(1):43-55.
17. Xueqin C, Yu Y, Zhao H. F0 Prediction from Linear Predictive Cepstral Coefficient. Wireless Communications and Signal Processing (WCSP), 2014 Sixth International Conference on. IEEE, 2014.
18. Shim HK, Park S, Chatterjee M, Scherer S, Sagae K, Morency LP. Acoustic and Para-Verbal Indicators of Persuasiveness in Social Mulatimedia. Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015.
19. Ganchev T, Fakotakis N, Kokkinakis G. Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task. Proceedings of the SPECOM. Vol. 1, 2005.
20. Peeters G. Deriving musical structures from signal analysis for music audio summary generation: Sequence and state approach. Lecture Notes in Computer Science. Bologna: Springer-Verlag; 2004.
21. Peeters G, Laburthe A, Rodet X. Toward Automatic Music Audio Summary Generation from Signal Analysis, Proceeding ISMIR; 2002. p. 94-100.
22. Xianglian L, Liu R, Li M. A Review on Objective Music Structure Analysis. Information and Multimedia Technology, 2009. ICIMT'09. International Conference on. IEEE; 2009.