IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-3, Issue-8 E-ISSN: 2347-2693

Speaker Recognition System Techniques and Applications

Sukhandeep Kaur[1*] and Kanwalvir Singh Dhindsa[2]

[1*,2] *Dept.of, CSE, BBSBEC FatehgarhSahib, Punjab Technical University, INDIA*

Received: Jul /09/2015 Revised: Jul/22/2015 Accepted: Aug/20/2015 Published: Aug/30/ 2015

Abstract- Speaker verification is feasible method of controlling access to computer and communication network. It is an automatic process that uses human voice characteristics obtained from a recorded speech signal, as the biometric measurements to verify claimed identity of speaker. It can be classified into two categories, text–dependent and text-independent system. This paper introduces the fundamental concepts of speaker verification for security system. It focuses on techniques and their unique features.

Keywords- Speaker Identification, Gamma Tone Frequency Cepstral Coefficient, Mel Frequency Cepstral Coefficient

1. INTRODUCTION

Speech is the primary way of communication between humans. Speaker recognition is the process of automatically recognizing an individual on the basis of characteristics of words spoken. Speaker recognition has always target on security system for managing the access to protected information from being used by anyone. Speaker verification is the branch of biometric authentication. This paper runs over comparison of voice recognition techniques. The parameter should be easily extracted, not be easily imitated, not to change with space and time as far as signal contains LCP, LPCC, MFCC, GFCCetc[4]. The current commonly used methods for speaker recognition are GMM (Gaussian
Mixture Model) , HMM (Hidden Markov Model), ANN (Artificial Neural Network)
etc. GMM extends of Gaussian probability density function working well in speaker recognition systems because of its capability to approximate the probability density distribution of arbitrary shape perfectly. HMM performs well in speaker recognition has a high accuracy. The three different methods based on HMM are DHMM, CHMM, and SCHMM [1]. ANN is a computational model based on the structure and functions of biological neural networks.ANN have three layers that are interconnected. The first layer consists of input neurons. Those neurons send data on to the second layer, which in turn sends the output neurons to the third layer.

2. RELATED WORK

Mukherjee et al. [2]discussed voice is one of the most assure and develop biometric modalities for access control. This paper presents a new method to recognize speakers by involve a new set of characters and using Gaussian mixture models (GMMs). In this research, the method of shifted MFCC was introduced so as to incorporate accent information in the recognition algorithm. The algorithm is evaluated using TIDIGIT dataset and the results showed improvements.

Wang and Ching[7]focussed on the features estimation method leads to robust recognition performance, specially at low signal-to-noise ratios. In the context of Gaussian mixture model-based speaker recognition with the presence of additive white Gaussian noise, the new approach produces logicalreduction of both recognition error rate and equal error rateat signal-to-noise ratios ranging from 0 to 15 db.

Faraj and Bigun [8]presented the first extended study investigationthe added value of lip motion features for speaker and speech-recognition applications. Digit identification and person-recognition andconfirmation experiments were conducted on the publicly available XM2VTS database showing goodresults (speaker verification was 98 percent, speaker recognition was 100 percent, and digit identification was 83 percent to 100 percent).

Sinith et al. [9]detailed the lay accent on text-Independent speaker recognition system where we adopted Mel-Frequency Cepstral Coefficients (MFCC) as the speaker speech feature argument in the system and the concept of Gaussian Mixture Modeling (GMM) for modeling the extracted speech feature. The Maximum likelihood ratio detector algorithm is used for the decision making process. The experimental study has been performed for various speeches time duration and several languages and was conducted around MATLAB 7 language environment. Gaussian mixture speaker model achieve high recognition rate for various speech durations.

3. SPEECH RECOGNITION ARCHITECTURE

Various architectures have been proposed for speech recognition. Voice is one of the most promising and mature biometric modalities for secured access control [3]. Speech recognition system mainly consist of mainly two modules: The recognition engine module and speech feature extraction module [4]. There are two phases in recognition model training and recognition. In feature extraction part, the voice record should be preprocessed to eliminate noise and emphasize individual differences caused by physiological structure of natural sound system and pronunciation habits.

The preprocessing has very important effects results of entire system [1]. It is real-time applications for access control, authentication and identification. It improves recognition rates for wide variety of noise scenarios.

In the training phase, each registered speaker has to provide samples of their speech, so that the system can train a reference model of that speaker.

In testing phase, the input speech is matched with stored reference model and a recognition decision is made.

Matching of pattern is the actual comparison of extracted frames with known speaker models, these results in the matching score which qualifies the similarity in between the voice recording and known speaker model, e.g. HMM. Speaker recognition is the identification of person who is speaking by the characteristics of their voices. Speaker recognition has two categories, text dependent - if the text is same for registration and verification, text-independent - if the text is different for registration and verification.

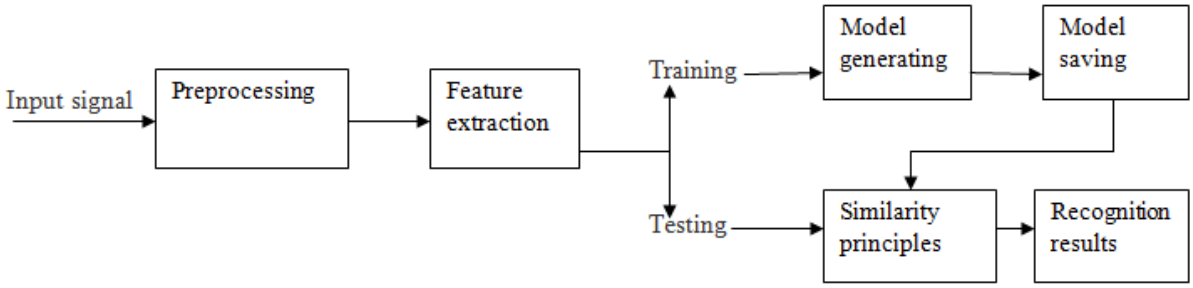

Fig.1 Structure of speaker recognition system [1]

4. SPEECH RECOGNITION TECHNIQUES AND APPLICATIONS

The feature analysis component of a speech recognition system plays an important role in the overall performance of the system. Many feature extraction techniques are available; these include linear predictive coding, Perceptual linear predictive coefficients, Mel frequency cepstral coefficients, Gamma tone frequency cepstral coefficients.

4.1 Linear Predictive Coding (LPC)

It is one of the most powerful speech investigate technique and a useful method for encoding quality speech at a low bit rate. The basic idea behind linear predictive investigate is that a specific speech sample at the present time can be estimated as a linear combination of past speech samples. The principle behind the use of this technique is to minimize the sum of the squared differences between the actual speech signal and the estimate speech signal over a restrictedduration [5].The applications of LPC are detailed below:

1. LPC is uses for a speech analysis and resynthesis. It uses as a form of voice compression by mobile companies, e.g. GSM.
2. LPC also uses for assure wireless, where voice is essentially encrypted and send across narrow voice channel, e.g. US government's Navajo I.
3. LPCsynthesis uses to build voice encoders (vocoders).
4. Predictors of LPC used in Shorten, SILK audio codes and other lossless audio codec.

4.2 Perceptual Linear Predictive Coefficients (PLPC)

The goal of the initial PLP model is to describe the psychophysics of human hearing more precisely in the feature extraction process. PLP and LPC analysis are similar, is based on the short-term spectrum of speech. This technique uses three methods from the psychophysics of hearingto derive an estimate of the auditory idea: (1) The critical-band spectral resolution. (2)The equal-loudness curve.

(3) The intensity-loudness power law.

It analysis is computationally efficient and Low-dimensional representation of speech [6].
Applications-
1. PLPimprove speech recognition rate.
2. PLP used for short speech.

4.3 Mel Frequency Cepstral Coefficients (MFCC)
Mel scale cepstral analysis andperceptual linear predictive analysis of speech are very similar, where the short term spectrum is changed based on psychophysically based spectral transformations. In this concept, however, the spectrum is crooked according to the MEL Scale, whereas in PLP according to the Bark Scale. Cepstral coefficients is the main difference between both techniques. This is done directly by converting of the log power spectrum to the cepstral domain using an inverse Discrete Fourier Transform [5].
Applications-
1. MFCC is used asfeatures in speech recognition system,e.g. automatically accept numbers spoken into telephone.
2. It also used in musical information recovery,e.g.Sort of classification and sound matching measures.

4.4 Gamma toneFrequency Cepstral Coefficients (GFCC)
It informs about rate of change in different spectrum bands, used to determine fundamental frequency of human speech and to analyze radar signal return. Like MFCC, feature vectors in this technique are calculated from the spectra of series of frames. After the logarithm is taken to each filter output and discrete cosine transform applied to filter output.
Application -
1. GFCC algorithm used for extraction of features of speech signals,eg. Emotions such as sad, joyfear etc.
2. GFCC improves validity of speech recognition in noisy conditions.

5. COMPARISON OF SPEECH RECOGNITION TECHNIQUES

This section offers an overview of main characteristics of MFCC and GFCC. The comparative differences of the above techniques of speech recognition system are as given below:

- Pre-emphasis- are LPCC, PLPC, MFCC where GFCC is not.
- Frequency bands- LPCC has 24, PLPC has 23, MFCC has 26 and GFCC has 64.
- Frequency scale- LPCC has no frequency scale, PLPC has Bark, MFCC has Mel and GFCC has ERB.
- Cepstralliftering – LPCC and MFCC has where PLPC and GFCC not
- Non-linear rectification-LPCC and MFCC has logarithmic where PLPC and GFCC has cubic root.

S.No	*Category*	*LPCC*	*PLPC*	*MFCC*	*GFCC*
1	Pre-emphasis	Yes	Yes	Yes	No
2	No. of frequency bands	24	23	26	64
3	Frequency scale	No	Bark	Mel	ERB
4	Cepstrallifte ring	Yes	No	Yes	No
5	Non-linear rectificatio n	Log	Cubic root	Log	Cubic root

6. CONCLUSIONS

We have examined features extraction methods. LPC parameter is average due to its linear computation nature. Because human voice is non-linear in nature, so LPC is not acceptable.PLP and MFCC are most often used features extraction techniques in the field of speech recognition and speaker verification system.PLP and MFCC are derived on the method of logarithmic spaced filter bank with the method of human soundly system has improved response as compare to LPC parameter. GFCC has non-linear rectification mainly accounts for the noise validity differences. In particular, the cubic root rectification provides more validity to the features than log.

The two main differences of GFCC and MFCC are, one is the frequency scale, and other is the non-linear rectification step prior to the DCT. GFCC based on equivalent rectangular bandwidth (ERB) scale, has better resolution at low frequencies than MFCC .MFCC use log whereas GFCC use cubic root. In particular, the cubic root rectification supplies more validity to the character than the log.

REFERENCES

[1] Wu Ju, “Speaker Recognition System Based on Mfcc and Schmm.” Symposium on ICT and Energy Efficiency and Workshop on Information Theory and Security,**2005**, Dublin Ireland, pp. **88 – 92**.

[2] R.Mukherje, I.Tanmoy, and R.Sankar, "Text dependent speaker recognition using shifted MFCC." Southeast on,**2013**, Proceedings ofIEEE,Orlando,FL USA,pp.**1-4**.

[3] D.A.Reynolds, “Speaker Identification and Verification Using Gaussian Mixture Speaker Models,” Speech Communication, Vol.17,**1995**, No. 1-2, pp. **91-108**.

[4] W. Junqin, and Y.Junjun, "An Improved Arithmetic of Mfcc in Speech Recognitions System." Electronics, Communications and Control (ICECC), International Conference on.IEEE,**2011**, ZhejiangChina,pp.**719-722**.

[5] U.Shrawankar, and V.M. Thakare, "Techniques for Feature Extraction In Speech Recognition System: A Comparative Study." International Journal Of Computer Applications In Engineering, Technology And Sciences,Vol. 2, No. 5**,2010**, pp. **412-418**.

[6] H.Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech." Speech Technology Laboratory, Division of Panasonic Technologies, Vol. 87, No. 4, **1990**,pp. **1738-1752**.

[7] N. Wang, and P.C.Ching, "Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features speaker verification," IEEE Transaction on Audio Speech and Language processing, Vol. 19, No. 1,**2011**, pp. **196-205**.

[8] M.I. Faraj, and J.Bigun,"Synergy of lip-motion and acoustic features in biometric speech andspeaker recognition."Computers, IEEE Transactions on computers Vol.56, No.9,**2007**, pp. **1169-1175**.

[9]M.S. Sinith,A.Salim,K. GowriShankar,S. Narayanan, and V. Soman,"A novel method for Text-Independent speaker identification using MFCC and GMM."Audio Language and Image Processing (ICALIP), International Conference on. IEEE,Shanghai,**2010**, pp.**292-296**.

[10] A.Solomon off,. "Channel compensation for SVM speaker recognition." Odyssey.Vol. 4,**2004**, pp.**57-62**.

[11] R.Collobert, andS.Bengio, "SVMTorch: Support vector machines for large-scale regression problems." The Journal of Machine Learning Research ,No.1,**2001**, pp. **143-160**.

[12] D.E.Sturim, and D.A. Reynolds, "Speaker Adaptive Cohort Selection for Tnorm in Text-Independent Speaker Verification."ICASSP ,No.1,USA ,**2005**, pp.**741-744**.

[13] G.S.V.S.Sivaram, Thomas, and H.Hermansky, "Mixture of Auto-Associative Neural Networks for Speaker Verification,"INTERSPEECH, Baltimore, USA,**2011**, pp. **2381-2384**.

[14] S.Gfroerer, "Auditory instrumental forensic speaker recognition" Proceedings of Eurospeech,Geneva, **2003**,pp. **705–708**.

[15] H.R.Bolt,andF.S.Cooper, "Identification of a Speaker by Speech Spectrograms," American Association for the Advancement in Science, Science, Vol. 166, **1969**.pp. **338–344**.

[16] D.Charlet, D.Jouvet, and O.Collin, "An Alternative Normalization Scheme in HMM-based Text-dependent Speaker Verification," Speech Communication, Vol. 31,**2000**, pp. **113-20**.

[17] T.Dutta, "Dynamic Time Warping Based Approach to Text-Dependent Speaker Identification Using Spectrograms," Congress on Image and Signal Processing, Vol. 2,No.8,**2008**, pp. **354-60**.

[18] M.Ben, M.Betser, F.Bimbot, and G.Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs." **2004**,*Proc. ICSLP*, France.

[19] A.A.M. AbushariahT.S,Gunawan, O.O. Khalifa, and M.A.M. Abushariah, "Voice based automatic person identification system using vector quantization". In Computer and Communication Engineering (ICCCE), International Conference, Kuala Lumpur, **2012**,pp. **549–554**.

[20] S.Agrawal, A.Shruti,and A.C.Krishna , "Prosodic Features Based Text Dependent Speaker Recognition Using Machine Learning Algorithms," International Journal of Engineering Science and Technology, Vol. 2,No.10,**2010**, pp. **5150-5157**.

