Speech and Speaker Recognition Technology using MFCC and SVM

Anamika Baradiya[#1], Vinay Jain[#2]

[#]*Department of Electronics & Telecommunication Engineering, SSCET Bhilai, Chhattisgarh, India*

Abstract

Speaker recognition is an active field of research with important forensic and security application .The investigation in the field of speaker recognition is in progress almost five decades and also there are several challenges and day to day new opportunities in this field. In observation of the fact that speech is the most natural form of communication for the human being it is also uses to express the sense and identity. A speaker is known through their tone which contained the information of speech signal. Speaker identification is one of the biometric identification technologies and now days it is use in different areas. The principle of Speaker recognition is to recognize the human being through their voice or speech signal. Speaker recognition is categorized into two categories such as speaker identification and speaker verification. The wider range of speaker recognition is in voice dialling, telephone shopping, telephone banking, database access services, voice mail and many others.

Speaker features of the input speech from test subject will be extracted and matched against the speaker model. A probability will evaluate the similarity between the model and the measured observations. The common approach is based on a threshold set for the acoustic likelihood ratio to decide the test speaker is accepted or not. Conventional speaker verification systems use hidden Markov models (HMM) or Gaussian mixture model (GMM) to perform the likelihood ratio test [1-6]. These systems use a generative model for all speaker models. This will result in over-fitting and maybe cannot maximize the discrimination of speaker and impostors.

Keywords- *Speaker verification, Speaker recognition, MFCC, SVM.*

I. INTRODUCTION

The speaker verification is a branch of automatic speaker recognition (ASR) system and can apply to determine whether a person is who he/she claims to be. Therefore, the problem of speaker verification is a true-false (accept-reject) question [1-6]. The speaker verification is widely used in many speech related applications, such as banking by telephone, voice dialing, and biometric security system [1-6]. Meanwhile, depended on the differences of recognition target, the systems of speaker verification fall into two types: text-dependent and text-independent. The former one requires that the speaker should provide keywords or sentences of the same text for both training and recognition, while the latter one does not depend on the specific text being spoken [1-6]. For security consideration, this paper will focus on the problem of the text-dependent speaker verification.

The choice of speaker features is another primary concern in the development of a speaker verification system. The ideal speaker features set should have higher inter-class variance and lower intra-class variability. In addition, the selected speaker features should be independent of each other as in order to minimize redundancy.

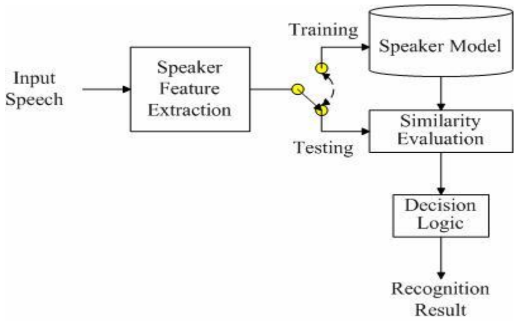

Fig. 1. Block Diagram of Speaker Recognition System

Based on the above discussion, the goal of this paper is to develop a more efficient approach to the text-dependent speaker verification using MFCCs and SVM. Previous researches [4-6] have shown that MFCCs can represent detail characteristics of individual speakers and therefore are mostly usable features for speaker verification. On the other hand, SVM is a two-class classifier based on the principles of structural risk minimization. It is shown that SVM has well generalization ability when compared to hidden Markov model and neural network based classifier [7]. Furthermore, since speaker verification is basically a binary decision, SVM seems to be a promising candidate to perform this task.

II. STUDY OF SPEAKER RECOGNITION

In this paper we try to review a general study about the speaker recognition technology during the last several decades. Even though many technologies have been developed but still several research issues remains which want to be accept as a challenge. Using a machine to recognize any one through speech signal is known as the

Automatic speaker recognition. Speaker/voice recognition is a biometric sensory system (i.e. a classification of proposal on the basis of whether they claim necessity or possibility) which is use peoples voice for recognition process it is differ from speech recognition, in speech recognition words are recognized as they are articulated and it not come in the category of biometric. In other words we can say that speech recognition identifies what you are saying and speaker recognition identifies that who are you. As discussed in many studies [8-10] about the speaker recognition technology that in the early 1960s i.e. after one decade for automatic speech recognition, it was Lawrence Kersta of bell labs who developed the first speaker recognition system which is based on spectrographic voice verification. Since the shape and size of vocal tract is vary from one speaker to another it shows the differences in resonance frequencies. Now days most of the speaker recognition system based on spectral information, these system use spectral information which is extracted from the speech signal segments of size 10-30 ms.

Speaker recognition undergoes development among the speech recognition along with speech synthesis technologies for the reason that the related characteristics in addition to challenges connected with each other. As authors says in [11] that a Swedish professor named Gaunnar Fant make available a model relating the physiological component of acoustic speech production in 1960s. Into 1970s Dr. Josheph Parkell elaborate the font model by using X-rays in addition to included the tongue and jaw.

III. WHY VOICE RECOGNITION

There are many available techniques of biometric which are used for identification to people such as identification through Voice, face, iris, retina, Fingerprint, hand geometry and signature. Biometric technology has many characteristic by which we can be able to distinguish in their applications. In the table provides a comparative study of biometric identification system based on their characteristics (de Luis- Garcia et al., 2003) [12-13].

Biometric Techniques	Comparison Based on Characteristics				
	Accuracy	Ease of Use	Ease of Implementation	Cost	User Acceptance
Voice	M	H	H	L	H
Face	L	L	M	L	H
Iris	M	M	M	H	M
Fingerprint	H	M	H	M	L
Retina	H	L	L	M	L
Hand Geometry	M	H	M	H	M
Signature	M	M	L	M	H

Table I. Comparison of Different Biometric Identification Systems charecteristics

Speaker recognition is the task to recognizing the people by machine using the information obtained by speech signal. Speaker recognition can be classified as speaker identification and speaker verification, speaker identification defined as that it is the procedure to determining the registered speaker from training data or utterance given by speaker while speaker verification is the procedure to accepting or rejecting the identity claim by the speaker. Further speaker recognition method classified in two cases that is text-dependent and text-independent. In case of text-dependent speaker recognition system, speaker gives same utterance or text in both cases that is the time of training and testing while in case of text- independent speaker recognition system utterance differ in training and testing time. As discussed in [14] there are many factors which are responsible for speaker recognition system such as dynamics of articulators, distinctive manner of oral expression, speaking rate, shape of vocal tract, size of vocal tract, rate of vibration of the vocal folds etc. these factors are extracted from the speech signal and the individuality of speaker-specific information due to these factors. Speaker specific information extracted from speech signal and speech signal hold the information regarding identity of speaker as well as language used for communication.

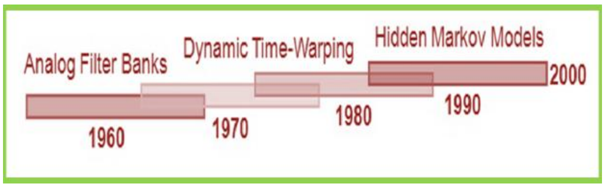

Fig.2. Day to day ASR Technology

IV. USE OF MFCC AND SVM FOR SPEAKER VERIFICATION

A. Support Vector Machine

An SVM is a two-class classifier constructed from sums of a known kernel function K (·, ·) to define a hyperplane.

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x_i}) + b$$

(1)

where $y_i \in \{1, -1\}$ are the target values, $\sum_{i=1}^{N} \alpha_i y_i = 0, and\ \alpha_i > 0$. The vector $\boldsymbol{x_i} \subseteq R^n$ are support vectors and obtained from the training. This hyperplane will separate given points into two predefined classes. Suppose a training set S= {(x_1, y_1), ……,(x_l, y_l)$\}^l_{i=1}$⊆ (*X* xY)land a kernel function $K(x_i, x_j) =< \emptyset(x_i), \emptyset(x_j)$ on *X* x *X* is given, where <·,·> denotes the inner product and ∅ maps the input space *X* to another high dimensional feature space F. With suitably chosen ∅ , the given nonlinearly separable samples S may be linearly separated in F, as shown in Fig. 3. An improved SVM called soft-margin SVM can tolerate minor misclassifications [4] and use in this paper.

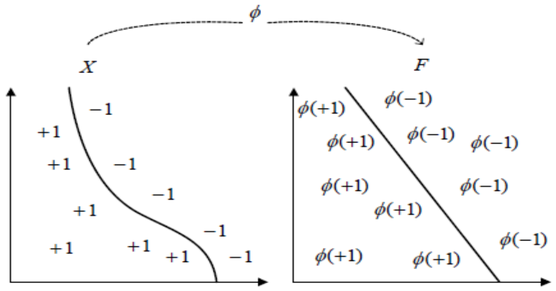

Fig.3. A Features Map Simplifies the Classification Task.

B. Mel Frequency Simplifies the Classification Task

It is shown that MFCC can capture the acoustic characteristics for speech recognition, speaker recognition, and other speech related applications [4-7, 15]. According to psychophysical studies, human perception of the frequency content of sounds follow a subjectively defined nonlinear scale called the "mel" scale [9] defined as,

$$f_{mel} = 1125 \ln(1 + \frac{f}{700}) \qquad (2)$$

where f is the actual frequency in Hz. This leads to the definition of MFCC and its calculation process shown in Fig 4.

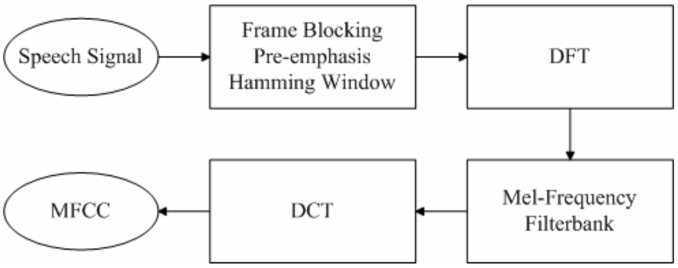

Fig.4. The Block Diagram of MFCC Calculation Process.

V. CONCLUSIONS

The experimental results of the proposed text-dependent speaker verification system are achieved by using 20 male and 20 female speakers selected from the Aurora 2 database [16]. All of the test speech signals are noisy-free and are sampled at 8000 Hz with 16-bit resolution. Each test speech signal consists of 2~8 English digital numbers or English alphabets. Speaker verification performance will be reported using the false acceptance rate (FAR), the false rejection rate (FRR), and the equal error rate (EER). The definitions of FAR and FRR are given as follows:

$$FAR = \frac{\#accepted\ imposter\ claims}{\#imposter\ accesses} X\ 100\% \qquad (3)$$

$$FRR = \frac{\#rejected\ genuine\ claims}{\#genuine\ accesses} X\ 100\% \qquad (4)$$

Once the receiver operating characteristic (ROC) curve of FAR vs. FRR is obtained, one can determine the EER, which FAR and the FRR at this point is the same for both of them.

In this paper, the different settings of MFCC order are studied experimentally for speaker verification. It follows from [8] that the higher-order MFCC does not further reduce the error rate in comparison with the lower-order MFCC. Hence, this paper compared the results obtained on the SVM based speaker verification system with 13 settings of MFCC order, namely p = 2q, q = 1~13. An impostor model was trained on all the MFCCs in the impostor data set while the speaker model was built using the corresponding speaker data set. During speaker verification task, a likelihood ratio was computed between the speaker model and the impostor model. The likelihood ratio was defined as:

LR = log P(x | speaker model) - log P(x | impostor model) (5)

where x is the input test MFCCs vector. Table 2 shows a summary of the experimental results of the proposed text-dependent speaker verification systems. It follows from Table 2 that the better performance could be obtained when MFCC order p = 22. An EER of 0% and average accuracy rate of 94.8% are achieved using the proposed system.

Table II. Comparison of SVM based Text – Dependent Speaker Verification System With Different CC Orders

MFCC order	Average accuracy rate	EER
2	71.8 %	12.4 %
6	85.0 %	2.3 %
10	90.0 %	1.9 %
14	92.5 %	1.1 %
18	93.8 %	0.0 %
22	94.8 %	0.0 %
26	93.0 %	0.4 %

REFERENCES

[1] Peter Day and Asoke K. Nandi, “Robust Text-Independent Speaker Verification Using Genetic Programming,” IEEE Trans. Audio, Speech, and Language Processing, Vol. 15, No. 1, pp. 285-295, 2007.

[2] Minho Jin, Frank K. Soong, and Chang D. Yoo, “A Syllable Lattice Approach to Speaker Verification,” IEEE Trans. Audio, Speech, and Language Processing, Vol. 15, No. 8, pp. 2476-2484, 2007.

[3] A.E. Rosenberg, “Automatic speaker verification: A review,” IEEE Proceedings, Vol. 64, pp. 475-487, 1976.

[4] Guiwen Ou and Dengfeng Ke, “Text-independent speaker verification based on relation of MFCC components,” 2004 International Symposium on Chinese Spoken Language Processing, pp. 57-60, Dec. 2004.

[5] A. Mezghani and D. O'Shaughnessy, “Speaker verification using a new representation based on a combination of MFCC and formants,” 2005 Canadian Conference on Electrical and Computer Engineering, pp. 1461-1464, May 2005.

[6] M.M Homayounpour and I. Rezaian, “Robust Speaker Verification Based on Multi Stage Vector Quantization of MFCC Parameters on Narrow Bandwidth Channels,” ICACT 2008, vol 1, pp.336-340, Feb. 2008

[7] C.C. Lin, S.H. Chen, T. K. Truong, and Yukon Chang, “Audio Classification and Categorization Based on Wavelets and Support Vector Machine,” IEEE Trans. on Speech and Audio Processing, Vol. 13, No. 5, pp. 644-651, Sept. 2005.

[8] Pawlewski, M, and J Jones. "URU Plus – a scalable component-based speaker-verification system for BT’s 21st century network." *BT Technology Journal.* Vol 23 .No 4 (October 2005): 45-53. Print.

[9] Pawlewski, Mark , and James Jones. "Biometric Technology Today." *BT Security Research Centre*. June 2006: 9-11. Print.

[10] Furui, Sadaoki. "50 years of progress in speech and speaker recognition." *Department of Computer Science Tokyo Institute of Technology*. 1-9. Print.

[11] Committee on technology, . "Speaker recognition." national science and technology council. 07 08 2006: 1-9. Print.

[12] Mathur S, Choudhary SK, Vyas JM (2013) Speaker Recognition System and its Forensic Implications. 2: 723 doi: 10.4172/scientificreports.723

[13] Büyük, Osman. "Telephone - based Text - dependent Speaker Verification." Trans. ArrayBoğaziçi University, 2011. 1-134. Print.

[14] Yegnanarayana, B. , S. R. Mahadeva Prasanna, Jinu Mariam Zachariah, and Cheedella S. GuptaPrasanna. "Combining Evidence From Source, Suprasegmental and Spectral Features for a Fixed-Text Speaker Verification System." *IEEE Transactions on Speech and Audio Processing*. Vol. 13.No. 4, (JULY 2005): 575-582. Print.

[15] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993

[16] http://www.elda.org/article52.html. **(**Aurora Database 2.0**)**