Volume 3, Issue 2, February 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Review on Automatic Speaker Recognition System

Ajit Singh, Taruna Panchal, Mainka Saharan
Department of CSE
Bhagat Phool Singh Mahila Vishwavidayalaya
Khanpur Kalan (Sonipat)

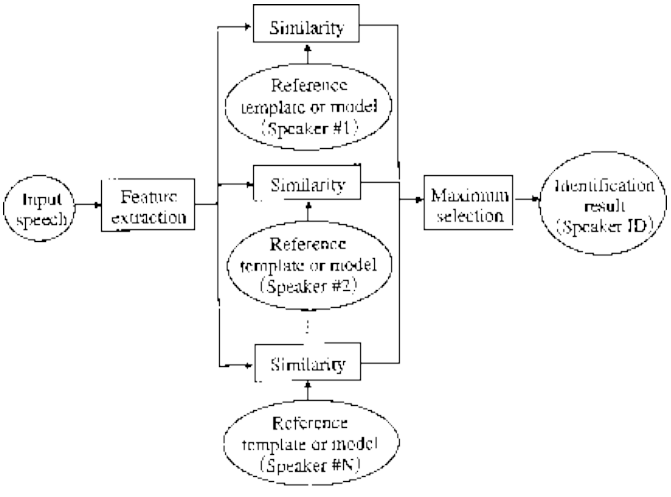
Abstract: Speaker Recognition is a task of identifying a user or verifying a user's claimed identity using the individual information present in their voices. It is successfully implemented in many commercial areas. It provide access control to voice dialing, database access services, information services, voice mail, security control for confidential information areas, remote access to computers. This paper describes the brief overview speaker recognition system , the current technologies used in the area of speaker recognition, development and provide a outline some potential future trends in research .

Keywords— Speech Recognition; Speaker Recognition, Feature Extraction; MFCC; LPC; Hidden Markov Model; Neural Network;

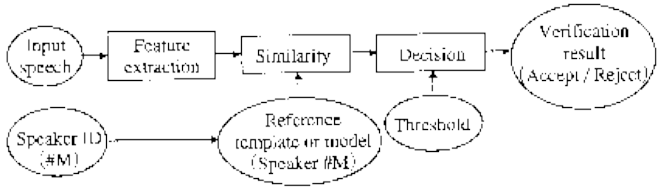
I. Introduction

The speech signal conveys many levels of information to the listener. At the primary level, speech conveys a message via words. But at other levels speech conveys information about the language being spoken and the emotion, gender and, generally, the identity of the speaker. Speaker recognition, or automatic recognition of a speaker, is closely related to speech recognition. The Speaker recognition consist of

(1) ***Speaker identification*** is the task of determining who is talking from a set of known voices or speakers. It is assumed the unknown voice must come from a fixed set of known speakers, thus the task is referred as *closed-set* identification .



(a) Speaker identification



(b) Speaker verification

Fig.1 Basic Structure of Speaker Recognition System

(2) ***Speaker verification*** (also know as speaker authentication or detection) is the task of determining whether a person is who he/she claims to be (a yes/no decision). Since it is generally assumed that imposters (those falsely claiming to be a valid user) are not known to the system, this is referred to as an *open-set* task. The fig. 1 shows the basic structure of Speaker Recognition System. The system extracts speech parameters from the input speech signal to represent vocal characteristics and uses these information to train a speaker specific model. . In training phase we build a reference model for a particular speaker and compared that stored reference model against the input speech. When a test utterance comes with a claim, it tests its parameters under the claimed speaker's model, and calculates a similarity score. If the similarity is above a threshold, the system accepts the speaker, if not, it assumes an illegal access attempt and rejects.

II. LITERATURE

Speaker-recognition is commonly used biomertric in commercial applications. It carried out in industries, national laboratories and universities. The institutions including AT&T and its derivatives (Bolt, Beranek, and Newman) the Dalle Molle Institute for Perceptual Artificial Intelligence (Switzerland); MIT Lincoln Labs; National Tsing Hua University (Taiwan); Nippon Telegraph and Telephone (Japan); Rutgers University and Texas Instruments (TI) Sandia National Laboratories, National Institute of Standards and Technology, the National Security Agency etc. have carried out research activities in speaker recognition domain and conducted evaluations of speaker-recognition systems. Text-independent approaches including Gish‟s segmental Gaussian model, Reynolds‟ Gaussian Mixture Model need to deal with unique problems (e.g. sounds and articulations present in the test material but not in training). To improve the accuracy in identification and verification task databases should be larger and thus enhancing the performance measurement of speaker recognition system. In case of high security applications speaker recognition system can be used with combination of other authenticators eg. smart cards. The existing ASV system performance makes them ideal for commercial applications. Lernout & Hauspie, T-NETIX, Veritel, Voice Control Systems are the existing commercial AVS System . Sprint‟s Voice FONCARD is deployed at very largest scale in commercial applications. Applications of Speech recognition includes control access to services such as voice dialing and voice mail, tele-banking, telephone shopping, database access related services, information services, security control for confidential information areas, forensic applications, and remote access to computers. Automatic Speaker- recognition systems provides great help in reducing crime of fraudulent transactions. The ASV system suffers two kinds of errors : Type:- Type I error (False Acceptance of an invalid user (FA)) and Type II error (False Rejection of a valid user(FR)).)). It uses a pair of subjects: an impostor and a target, to make a false acceptance error. In high security ASV applications these error are main area of concern.

III. CLASSIFICATION OF SPEAKER RECOGNITION

The speaker recognition method is also classified into :

1**. Text- Dependent and**

2**. Text -Independent methods**.

Text-dependent methods: Text dependent methods are usually based on template-matching techniques. In this approach, the input word or sentence is represented by a series of feature vectors or spectral vectors . It is uses Dynamic Time wraping Algorithm for registered speakers /or template matching. The statistical variation in spectral feature vector is efficiently model by Hidden Markov model ,so HMM model is used for better recognition accuracy .

Text-independent methods: Text-Independent methods are based on vector quantization (VQ).In this method speaker–specific features are represented by VQ code or set of training vector and a codebook is obtained for speaker using the training vector

IV. Front end Analysis/ Speech Feature Extraction

The spectral fetures of the speech signal are extracted using the feature extrraction techniques.it is also known as front end analysis. The spectral feature vector of speech signal is efficiently model uses following models , the Hidden markov model (HMM), Dynamic Time wraping (DWT), Neural network and Vector Quantization. these methods are mathematically complex and commonly used in pattern recognition techniques in the area of speech recognition. The spectrum representations of speech signal extremely useful in speaker recognition. Linear Predictive coefficients(LPC), transformation of LPC reflection coefficient, ceptral coefficients, Mel Frequency Ceptral Coefficient (MFCC) are commonly used spectrum representation & extract feature vectors from the input speech signal . The analysis of speech signal is called Front end analysis the speech feature extraction module extracts speech parameters from the input speech signal to represent vocal characteristics, and speech parameters also called as speech vectors. The speech signal undergo various signal pre-processing tasks before being subjected to extracting module. These tasks include:-

- Truncation
- Frame blocking
- Windowing
- Fourier Transform

Truncation : When an audio clip is recorded, the number of samples generated would be around 90000 which are difficult to handle, so we can truncate the signal by selecting a particular threshold value. We can mark the start of the signal where the signal goes above the value while traversing the time axis in positive direction.

Frame blocking : In Frame blocking the continuous speech signal is divided into frames . In this step ,Speech signal is divided into block of I frames of N samples with adjacent frames being separated by N samples with the value N less than that of M(N<M). The first frame consists of the first N samples. The second frame begins from M samples after the first frame, and overlaps it by N - M samples and so on and this process continues for all the speech.

Windowing: In this step at the beginning and end of each frame windowing of frame is done to minimize the discontinuities of speech signal. Hanning or Hamming window function used for weighing of speech samples.

Fourier Transform: Fast Fourier Transform (FFT), which converts each frame of N samples from the time domain into the frequency domain.

1. **Linear coefficients (LPC)**

Linear Predictive coefficient (LPC) is one of the most powerful signal analysis techniques for estimating the parameters of speech signal. In LPC the speech sample represented as linear combination of past speech samples.LPC coefficient are determined through minimizing the sum of squared differences between the actual speech samples and predicted values. LPC provides a accurate estimate of speech feature vectors and useful model of speech. The LPC coefficients are transformed to (LPCC) Perceptual Linear Prediction (PLP) is the another variant of LPC analysis. This technique take advantage of characteristics derived from the psycho-acoustic properties of the human ear and modeled by filter-bank. The steps involved in the LPC feature extraction shown in figure.

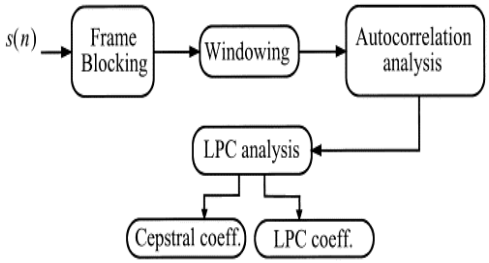


Fig.2 Steps involved in the LPC feature extraction

2. Mel Frequency Cepstral Coefficients (MFCC)

The MFCC have been extensively used in speaker recognition. The steps involved in MFCC feature extraction shown in fig. Firstly speech signal sample is extracted using a window. For the windowing procedure , duration of the window and shift between two consecutive window are two important parameters . Hanning or Hamming window function used for weighing of speech samples. Using fast Fourier transform (FFT) the spectrum magnitude of speech signal is obtain and then processed by band–pass filter bank**.** Triangular filters are used for MFCC computation . **A** logarithmic frequency scale chosen for bank filter known as Mel- Frequency scale. The practical warping is done by using a triangular Mel scale filter bank which handles the warping from Frequency in Hz to frequency in mel scale. Transformation of frequency bins to Mel scale bins by following equation:

$$m_y[b] = a_f w_b[f] \, |[Y][f]|^2,$$

where *wb* [f] is the *b*th Mel-scale filter's weight for the frequency *f* and *Y*[f] is the FFT of the windowed speech signal. Discrete cosine transform (DCT) is used to de-correlate the Mel-scale filter outputs in the last step .The subset of the DCT coefficients are chosen and represent the MFCC features used in the enrollment and the verification phases. By doing the Discrete Cosine Transform the contribution of the pitch is removed.

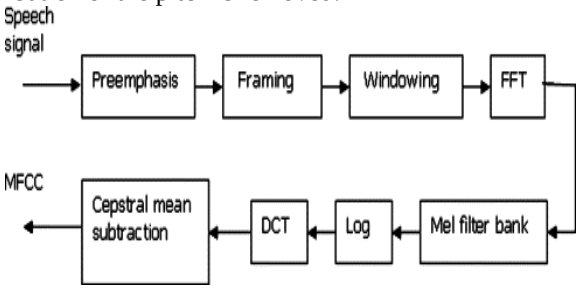


Fig.2 Steps involved in the MFCC feature extraction

V. SPEAKER MODELING APPROACHES

The purpose of speech modeling lies in building a model that can capture variations in a set of features extracted from a given speaker. Depending on Speaker recognition method ,i.e text dependent and text independent method , speaker modeling method can be classified as :

1. Parametric Approaches

2. Non-parametric Approaches

Parametric Approach: Text independent methods use the parametric method of speaker modeling & does not make structural assumptions about speaker model. The modeling techniques such as: 1.) neural networks(NNs) , 2.) support vector machines(SVMs), 3.) Probabilistic models such as the Hidden Markov Model(HMM) and Gaussian Mixture Model(GMM) have been used in speaker recognition tasks. The Probabilistic models such as the Hidden Markov Model(HMM) and Gaussian Mixture Model(GMM) are dominant techniques ,so they can be used for modeling speakers in both the text dependent & text independent speaker recognition.

Non-Parametric Approach: Text- dependent methods use the non- para metric method of speaker modeling. This method make structural assumptions about speaker model .It includes 1.)Template-matching , 2.) Nearest-Neighbour modeling , 3.) Vector Quantization(VQ) .

These various speaker modeling approaches are briefly discussed in this section.

Template-matching : In this approach the extracted feature vectors through the speech feature extraction module are serve as templates for speaker voices and DWT algorithm is used for align and provides measure of similarity between the test speech vectors and templates of feature vector during verification process.Template matching is a text dependent speaker modeling technique.It is very accurate word model. Preparation of templates of feature vector and matching becomes very expensive as the size of words increases beyond few hundred words hence the requirements of processing power and storage makes this method inefficient for speaker identification tasks.

Nearest - Neighbour modeling : In this approach the extracted feature vectors through the speech feature extraction module are store and use as reference feature vectors of input speakers . The match between the test feature vector and reference vector is then cumulated distance of each test vector to its k nearest neighbours .

Hidden Markov Model (HMM): Hidden Markov Model is stochastic model.HMM is simple, computationally feasible and can be trained automatically so they are very popular model in speaker recognition.HMM are simple networks which generate spectral vectors of speech using an number of states for each model and modeling the short term spectra for each associated state ,usually with mixtures of multivariate Gaussian distributions also known as state output distributions. The parameters of the model are the state transition probabilities and the means, variances and mixture weights that characterize the state output distributions. Each word, or each phoneme, will have a different output distribution; a HMM for a sequence of words or phonemes is made by concatenating the individual trained HMM for the separate words and phonemes. Currently HMM based large vocabulary speech recognition systems are trained on large speech data for hours, and during trainning it automatically word and phone boundary information. The advantage of HMM is that it reduce the time and complexity for training the large vocabulary in recognition .

Neural Network modeling approach: Neural Network have also been used for speaker recognition system. They are being used in solving complex recognition tasks.They can handle low quality , noisy data and are speaker independent. The disadvantage of NN approach is that optimal configuration selection is not easy to select.NN based approach used in phoneme recognition.System usinh NN approach based system provides greater accuracy than HMM for limited training data and vocabulary. The NN-HMM hybrid use the neural network part for phoneme recognition and the HMM part for language modeling.

Support Vector Machine: Support vector machine models provides good verification performance and error in performance can be directly estimated from training data.the idea behind SVM approach is the use different types of kernel functions like linear, quadratic, or exponential. which map the feature vectors of speech signal to higher dimensional feature space by using a non-linear transformation.SVM generates a score and scores are integrated over the entire utterance to obtain the final decision score. Recent SVM based approach have applied to super vectors which are single vectors that represent an entire utterance

Vector Quantization (VQ): It is the process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is a cluster and represented by its center called a code word. for each speaker a codebook is generated in VQ method. A codebook is collection of all code words. The codebook then serves as template for the speaker,

and is used when testing a speaker in the system. The advantages of VQ method is it saves a lot of time during the testing phase ,reduces the storage and computation effort in determining the similarity of spectral analysis vectors .

VI. APPLICATIONS

The applications of speaker recognition technology are quite varied and continually growing. There are broad areas where speaker recognition technology has been or is currently used .

Access Control: It used for controlling access to computer networks (add biometric factor to usual password and/or token) or websites (thwart password sharing for access to subscription sites). Also used for automated password reset services.

Transaction Authentication: It is also used for telephone banking, account access control, and can be used for more sensitive transactions which needs higher levels of verification. User verification for remote electronic and mobile purchases (e- and m-commerce).

Law Enforcement: Applications which includes home-parole monitoring (call parolees at random times to verify they are at home) and prison call monitoring (validate inmate prior to outbound call). It is also used in forensic analysis applications for aural/spectral inspections of voice samples .

Speech Data Management: In voice mail browsing or intelligent answering machines, speaker recognition is used to label incoming voice mail with speaker name for browsing and/or action (personal reply). For speech skimming or audio mining applications, annotate recorded meetings or video with speaker labels for quick indexing and filing.

Personalization: It is used in voice-web or device to store and retrieve personal setting/preferences based on user verification for multi-user site or device (car climate and radio settings .)

VII. Conclusion

Speaker recognition can be classified into text dependent and the text independent methods. This paper describes the brief overview speaker recognition system , the current technologies used in the area of speaker recognition, development and provide a outline some potential future trends in research .

References

[1].[Cam97] Joseph P. Campbell, .Speaker recognition: A tutorial., Proc. IEEE, vol. 85, pp. 1437- 1462, September 1997.
[2].An Overview of Automatic Speaker Recognition Technology•1 *Douglas A. Reynolds*
MIT Lincoln Laboratory, Lexington, MA USA
[3] A Review on Speech Recognition Challenges and Approaches *World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 1, 1-7, 2012*
[4][Faq00] ICSI Speech FAQ: 5.1, .What are features? What are their desirable properties?., Answer by: dpwe - 2000-05-26, downloaded from http://www.icsi.berkely.edu/local-cgibin/ man-cgi-bin?ftrs-intro.html on Tue Mar 25 20:52:32, 2003.
[5].[Hor95] Bojan Imperl, Zdravko, and Bogomir Horval, .Use of Harmonic Features in Speaker recognition., Laboratory for Digital Signal Processing, Faculty of Electrical Engineering and Comp. Sci., Smetanova 17, 2000 Maribor, Slovenia.
[6].[Imp94] Bojan Imperl, .Speaker recognition techniques., Laboratory for Digital Signal
Processing, Faculty of Electrical Engineering and Comp. Sci., Smetanova 17, 2000 Maribor, Slovenia.
[6][Rey02] D. A. Reynolds, .An Overview of Automatic Speaker Recognition Technology., Proc. IEEE, pp. 4072-4075, 2002.
[7][Rsc78] L. Rabiner and R. Schafer, *Digital Processing Speech Signals*, Englewood Cliffs, NJ: Prentice Hall, Inc., 1978.
[8][Sha01] D. O.Shaughnessy, *Speech Communications . Human and Machine*, Universities Press (India) Limited, 2001.
[9][Ttp03] Figure downloaded on Mon, Nov 10, 21:34:05, 2003 from http://lumumba.luc.ac.be/jori/thesis/onlinethesis/chapter4.html#fig5.).
[10] Statistical Pattern Recognition Techniques for Speaker Verification Fazel and Shantanu Chakrabartty.

