Speaker Recognition : A Review

Sairam S. Vyawahare
M.E. Student, Govt. Engg. College, Aurangabad-431005

Abstract

Voice communication is the most effective mode of communication used by human beings. The speech processing is an important application of Digital Signal Processing. Speech processing technology consists of analysis/synthesis, recognition, coding etc. And the recognition field may further consist of Speech recognition, Speaker recognition and Language identification. The objective of speaker recognition system is to extract and recognize the information about speaker identity. The speech signal is a slowly time varying signal (so, called quasi-stationary signal), when examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are fairly stationary. However, over long periods of time (0.2s or more) the signal characteristics change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is used to characterize the speech signal. For speaker recognition, features that exhibit high speaker discrimination power, high interspeaker variability, and low intraspeaker variability are desired.

1. Introduction

1.1 Speaker recognition

Speaker recognition is consisting of identification and verification. Speaker verification is the use of a machine to verify a person's claimed identity from his/her voice. This is somewhat different than the speaker identification, which is deciding if a speaker is a specific person or is among a group of persons.

A speaker recognition system mainly consists of two main modules, the speaker specific feature extractor followed by the speaker modeling technique for generalized representation of extracted features [5]. Feature extractor is the first component in speaker recognition system. Feature extraction process transforms the raw speech signal into a compact and effective representation that is more stable and discriminative than the original signal. Extracted features should meet some criteria while dealing with the speech signal [17]. Such as-

- Easy to measure.
- It should show negotiable fluctuations from one speaking environment to another.
- It should be stable over time.
- It should occur frequently and naturally in the speech.

As feature extractor is the first component, it will decide quality of an entire speaker recognition system. In other words, result can be most as accurate as features [18].

Speaker recognition methods can be divided into text-dependent and text-independent methods. In a text-dependent method, the speaker has to say specific word or sentence having the same text for both training and recognition/testing mode. In a text-independent, on the other hand, does not rely on a specific text being spoken. Each method has advantages as well as disadvantages also. And they may require different treatments and techniques for the implementation [4].

Many factors can contribute to verification and identification errors. Some of the human and environmental factors that contribute these errors are as listed below [1]:

- Misspoken or misread prompted phrases.
- Extreme emotional states (e. g., stress, fear, joy, etc.)
- Sickness (e.g., head colds can alter the vocal tract)
- Time varying microphone placement.
- Channel mismatches (e. g., using different microphones for enrollment and verification).

The general approach to speaker verification system consists of five steps: digital speech data acquisition, feature extraction, pattern matching, making an accept/reject decision, and enrollment to generate speaker reference models [1]. A generalized flow diagram of this procedure is shown in Figure 1.

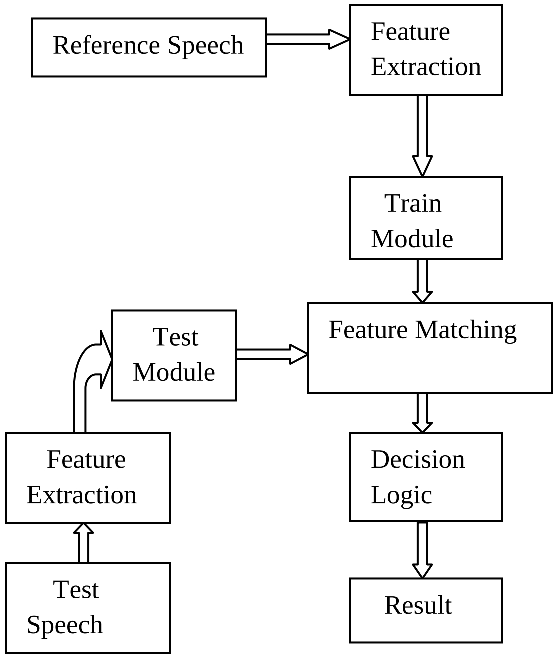

Figure 1. Generalized flow of speaker-verification system.

Feature extraction maps each interval of speech to a multidimensional feature space (A speech interval typically spans 10–30 ms of the speech waveform and is referred to as a frame of speech).

This sequence of feature vectors is then compared to speaker models by pattern matching. This results in a match score for each vector or sequence of vectors. The match score measures the similarity of the computed input feature vectors to models of the claimed speaker or feature vector patterns for the claimed speaker. Last, a decision is made to either accept or reject the claimant according to the match score or sequence of match scores, which is a hypothesis testing problem.

1.2. Necessity

Speech is one of the natural forms of communication. Speech is one of the most important tools for communication between human and his environment. Therefore manufacturing of Automatic System Recognition (ASR) is desire of the time [11]. Recent development has made it possible to use this in the security system. This technique makes it possible to use the speakers voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers.

At a more useful level, speech recognition is increasingly used in automated telephone-based interactive services. For example, it is possible to check the weather forecast, the price of a stock market share, or book a flight using an increasing number of these services. There are advantages to this for both the customer (no waiting for a human operator) and the service supplier (less staff required, can operate 24/7) [16].

1.3. Applications

Some of the applications of speaker verification systems are:

- Telephone-Banking.
- Voice-mail.
- Biometric Login to telephone aided shopping systems.
- Time and Attendance Systems.
- Remote Access to Computers.
- Security control for confidential information.

2. Speaker recognition techniques

Following section gives the review on various techniques for speaker recognition system.

2.1. Speaker recognition using MFCC and VQ approach

Authors [4] have applied Mel Frequency Cepstrum Coefficients (MFCC) technique for speaker identification and Vector Quantization (VQ) technique is used for feature matching. Authors suggest that VQ technique is not only simple but also highly accurate. A traditional triangular shaped filter is used for calculation of MFCC. Here, authors have found out identification rate for different kinds of windows using linear scale and mel scale (The Mel scale is linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz). And experimentally proved that the percentage of an identification rate is maximum for hamming window along with mel scale. Authors also suggest that the linear scale can also have a reasonable identification rate if a comparatively higher number of centroids are used (In other words, the codebook size that constitutes from number of centroids should be increase). However, Mel scale is less vulnerable to the changes of speaker's vocal cord in course of time and so, it is mostly used.

Authors namely Satyanand Singh and Dr. E. G. Rajan [5] have proposed text-independent speaker recognition system using MFCC and VQ techniques. In addition to that an Inverted MFCC is used as one of the performance enhancement parameter for speaker recognition, which contains high frequency region complementary information. Authors have introduced the Gaussian Filter (GF) while calculation of MFCC and Inverted MFCC in place of traditional triangular filters. The main idea behind this is to introduce a higher amount of correlation between sub-band outputs. The authors have proved experimentally that the method gives tremendous improvement and it can detect the correct speaker from much shorter (16% of training length, even 0.5sec of duration) speech samples. It reaches 98.57% identification rates by taking training and testing voice corpus of 2 seconds. Even with a very short test sequence of 0.5 second the proposed method achieved identification rate of 91.42%. And therefore this method is well applicable in real-time systems.

Another authors named Soong, Rosenberg, et.al.[6] have used VQ codebook as an efficient means of characterizing the short-time spectral features of a speaker. A set of such codebooks were then used to recognize the identity of an unknown speaker from his/her unlabelled spoken utterances based on minimum distance (distortion) classification rule. Instead of using word-based VQ codebooks to characterize the phonetic contents of isolated words, authors have proposed a speaker-based VQ codebook approach to speaker recognition application. The VQ codebook was used as an efficient means to characterize a speaker's feature space and was employed as a minimum distance classifier in the proposed speaker recognition system. The results of the 100 talker speaker recognition experiments are good; given a 10 digit long test token and a codebook of 64 vectors a recognition rate of over 98% was achieved. The results obtained by authors are summarizing as follows:

(i) Both larger codebook size and longer test token length (more digits in the test utterance) can be used to improve the recognition performance. Ten different digits, when used as test tokens, outperformed any of the repeated digit test tokens. Moreover, when digit "9" used as a test token outperformed other digits as it has a relatively long duration and it has a strong nasal-vowel co articulation. In short, phonetically rich test tokens give better performance than phonetically poor test tokens.

(ii) It is recommended that all speech frames (both voiced and unvoiced frames) be used during training and recognition/testing. Since, in the training phase, unvoiced frames can’t remove deliberately. (Hence, this nonparametric VQ approach not only eliminates the need to separate voiced frames from the input data, but also improves the speaker recognition performance by using all the speech data.)

(iii) It is recommended that the VQ codebook be updated from time to time to alleviate the performance degradation due to different recording conditions and intra-speaker variations.

2.2. Speaker recognition using MFCC and VQ for Hindi words

Authors namely Nitisha and Ashu Bansal [7] have implemented an automatic speaker recognition system using Mel Frequency Cepstrum Coefficients (MFCC) technique for feature extraction and Vector Quantization (VQ) technique for feature matching. Here, authors have proposed the text-dependent speaker recognition system and instead of English text authors have used Hindi text for a greater degree of recognizing accuracy. About 90% success rate has been achieved during the experiment.

2.3. Speaker recognition using VQ by LBG and KFCG

In this paper, authors [8] have proposed two approaches for text-dependent speaker recognition system based on vector quantization and their performances are compared.

Two methods for codebook generation have been used. In the 1st method, codebooks are generated from the speech samples by using the Linde-Buzo-Gray (LBG) algorithm. While in the 2nd method, the codebooks are generated using the Kekre’s Fast Codebook Generation (KFCG) algorithm.

The results are obtained by varying number of feature vectors (code vectors) with and without overlapping of speech samples. The results show that accuracy decreases as the number of feature vectors are increased with or without overlap for LBG. For KFCG, the results are consistent and also accuracy increases in the number of feature vectors for without overlap approach. Also KFCG is simple and faster as only simple comparisons are required as against Euclidean distance calculation for LBG.

2.4. Robust text-independent speaker identification using GMM

Authors namely Reynolds and Rose [9] have introduced and motivated the use of Gaussian Mixture Models (GMM) for robust text-independent speaker identification. The Gaussian mixture speaker model was specifically evaluated for identification tasks using short duration utterances from unconstrained conversational speech, possibly transmitted over noisy telephone channels. Through experimental evaluation authors examined several aspects of using Gaussian mixture speaker models for text-independent speaker identification. They listed as follows:

- Identification performance of the Gaussian mixture speaker model is insensitive to the method of model initialization.
- Variance limiting is important in training to avoid model singularities.
- There appears to be a minimum model order needed to adequately model speakers and achieve good identification performance.
- The Gaussian mixture speaker model maintains high identification performance with increasing population size (the system attained 98.6% identification accuracy for 5 second clean speech utterances and 80.8% accuracy for 15 second telephone speech utterances for an all-male 49 speaker population).
- Cepstral mean normalization is very effective compensation for telephone spectral variability degradations.
- With nodal variance parameterization, the GMM outperforms the Vector Quantization (VQ), Radial Basis Function (RBF), Tied Gaussian Mixture Model (TGMM) and Gaussian Classifier (GC) speaker modeling techniques on an identical telephone speech task.

These results indicate that Gaussian mixture models provide a robust speaker representation for the difficult task of speaker identification using corrupted, unconstrained speech. The models are computationally inexpensive and easily implemented on a real-time platform.

2.5. Speaker recognition based on short polish sequences

This paper presents results of speaker recognition system carried out using short polish (The Slavic language of Poland) sentences.

According to authors [10], techniques of identification based on the acoustic signals (voice) are less popular and they hold about 3% share in commercial biometrics market. However, speaker identification has a number of advantages and can be used to authorization access during access of multiple services and systems such as voice dialing options, telephone banking, shopping by phone, database access, voicemail, etc.

Authors said that the Vector Quantization (VQ) and Gaussian Mixture Model (GMM) techniques are well suited for text-independent speaker recognition system, while Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) techniques are well suited for text-dependent speaker recognition.

Authors have applied MFCC technique for feature extraction and feature matching is done with VQ and GMM. Finally results of VQ and GMM are compared.

It was observed that identification efficiency for GMM is more than VQ for very short duration of sequences.

2.6. Advanced method for speech recognition

Authors namely Meysam Pour and Farokhi [11] have proposed an advanced method which removes deficiencies of each available technique and is able to classify speech signals with a high accuracy at the minimum time.

The speech recognition process suggested by author contains four main stages-

(i) Acoustic processing- The main task of this unit is filtering of white noise from speech signals and it consists of three parts, Fast Fourier Transform, Mels Scale Bank pass Filtering and Cepstral Analysis.

(ii) Feature extraction using the Discrete Wavelet Transform (DWT) coefficients.

(iii) Classification and recognition using the Multi-Layer Perceptron (MLP) neural network.

(iv) At last, after training of neural network effective features are selected with UTA algorithm.

The UTA algorithm redounded to increase system learning time from 18000 to 6500 epoch and system accuracy average value to 98%. Considerable specification of this system are excellent performance with minimum training samples, fast learning and wide range of recognition and online classification of the receiving signals.

2.7. Speaker recognition based on idiolectal differences between speakers

Idiolect means, the language or speech of one individual at a particular period of life [12]. Word unigrams and bigrams, used in a traditional target/background likelihood ratio framework, are shown to give surprisingly good performance. Performance continues to improve with additional training and/or test data. Bigram performance is also found to be a function of target/model sex and age difference.

According to author George Doddington [12], human listeners can distinguish between speakers who are familiar to them far better than those who are unfamiliar. This increased ability is due no doubt to speaker idiosyncrasies (uniqueness) that are recognized by the listener, either consciously or unconsciously. According to author, these speaker characteristics offer the possibility to significantly improve automatic speaker recognition performance, if only we were able to identify and use them. This study was directed toward the statistics of word sequences as a function of speaker.

The performance of speaker detection based upon bigram statistics is surprisingly good. Surprising from several aspects, not just that speaker detection error rates are low:

- Although performance was observed to continue to improve as the amount of training data was increased, nonetheless good performance was observed for a surprisingly small number of training conversations.
- Performance was maintained while excluding all but a small number of bigrams, on the order of a few thousand. These bigrams are namely those that occur most frequently.

2.8. Channel compensation for SVM speaker recognition

One of the major remaining challenges to improving accuracy in state-of-the-art speaker recognition algorithms is reducing the Impact of channel and handset variations on system performance. To handle this issue Support Vector Machine (SVM) based speaker recognition is the best way [13].

SVMs are two-class hyperplane-based classifiers operating in a (usually) high-dimensional space related nonlinearly to the original (usually lower-dimensional) input space.

SVM based systems have performance close to that of the best GMM based systems. And these systems possess substantial advantages in terms of computational cost, both in training and testing as well. When scores of SVM and GMM based systems are fused, the result is typically much better than either system alone.

3. Conclusion

In this review paper, the various techniques developed for each stage of speaker recognition system have been discussed. It is observed that-
(1) For feature extraction the FFT and DCT techniques are used with linearly spaced filter bank while MFCC technique is used with logarithmically spaced filter bank.
(ii) In case of feature matching, VQ and GMM are suitable for text-independent system while DTW and HMM is best for text-dependent system.

4. References

[1] Joseph P. Campbell, “Speaker recognition: A Tutorial,” *Proc. IEEE*, vol. 85, no.9, pp. 1437-1462, Sept. 1997.

[2] Richard J. Mammone, Xiaoyu Zhang and Ravi P. Ramchandran, “Robust speaker recognition: A feature basedapproach,” *IEEE Signal Processing Mag*., pp. 58-71, Sept. 1996.

[3] Abdul Syafiq B Abdul Sukor, “Speaker identification system using MFCC procedure and noise reduction method,” University Tun Hussein Onn Malaysia, Jan. 2012.

[4] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani & Md.Saifur Rahman, “Speaker identification using mel frequency cepstrum ,” in *Proc. Int. Conf. Electrical and Computer Engineering*, Dhaka, Bangladesh, 2004.

[5] Satyanand Singh and Dr. E.G. Rajan, “Vector quantization approach for speaker recognition using MFCC and Inverted MFCC,”

[6] F. K. Soong, A. E. Rosenberg, L. R. Rabiner and B. H. Juang, “A Vector quantization approach to speaker recognition,” *IEEE Trans*., vol. 85, pp. 387-390, 1985.

[7] Nitisha and Ashu Bansal, “Speaker recognition using MFCC front end analysis and VQ modeling technique for Hindi words using MATLAB,” in *Proc. Int. Journal Computer Applications*, vol. 45, no. 24, May 2012.

[8] Dr. H. B. Kekre and Vaishali Kulkarni, “Performance comparison of speaker recognition using vector quantization by LBG and KFCG,” in *Proc. Int. Journal Computer Applications*, vol. 3, no. 10, July 2010.

[9] Douglas A. Reynolds and Richard C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, Jan 1995.

[10] Tomasz Marciniak, Radoslaw Weychan, Agenieszka Krzykowska, Szymon Drgas and Adam Dabrowski, “Speaker recognition based on short polish sequences,” Poznan, Poland.

[11] Meysam Mohamad Pour and Fardad Farokhi, “An advanced method for speech recognition,” World Academy of Science, Engineering and Technology 49 2009.

[12] George Doddington, “Speaker recognition based on idiolectal differences between speakers,” National Institute of Standards and Technology, USA.

[13] Alex Solomonoff, Carl Quillen, and William M. Campbell, “Channel compensation for SVM speaker recognition,” MIT Lincoln Laboratory Lexington, Massachusetts, USA.

[14] Jamel Price and Ali Eydgahi, “Design of Matlab®-based automatic speaker recognition system,” *Int. Conf. Engineering Education*, session T4J, no.1, July 2006.

[15] IBM (2010) online IBM Research Source:-http://www.research.ibm.com/Viewed 12 Jan 2010.

[16] S. K. Gaikwad, B. W. Gawali, P. Yannawar, “A review on speech recognition technique,” *Int. Journal Computer Applications,* vol. 10, no. 3, Nov. 2010.
[17] http://www.wikipedia.com
[18] Vibha Tiwari, “MFCC and its applications in speaker recognition,”*Int. Journal Emerging Technologies*, Feb.2010.