Text-Independent Speaker Recognition System

Yuanzhong Zheng
Department of Electrical and Computer Engineering, University of Rochester, Rochester, New York 14627
yzheng25@ur.rochester.edu

ABSTRACT

The article introduces a simple, yet complete and representative text-independent speaker recognition system. The system can not only recognize different speaker in the normal condition, but it also can distinguish different speaker in telephone. The system implements Linde–Buzo–Gray algorithm to generate a codebook for training dataset and recognizes different speakers by calculating Euclidean distance.

Key words: speaker recognition system, telephone, MFCC, LBG

1. INTRODUCTION

Everyone has his own unique timbre which known as "voice-print", so people can always distinguish who is on the other end of the phone as soon as they answer the telephone. In computer science, speaker recognition1 refers to identify "who is speaking".

Speaker recognition can be dated back to 1970s. It distinguishes different individuals by acoustic features. Speaker recognition system is difficult to develop due to the highly variant of input speech signals and the principle source of variance is the speaker himself. No two individuals sound are identical because their vocal tract shapes, larynx sizes and other parts of their voice production organs are different. Each speaker has his own characteristic manner of speaking, including particular accent, rhythm, intonation style, vocabulary selection and pronunciation pattern. Moreover, other factors, beyond speaker variability, show a challenge to speaker recognition technology. Examples of these are acoustical noise and variations in recording environments (e.g. speaker uses different telephone handsets). Speaker verification has earned speaker recognition its classification as a "behavioral biometric".

The automatic system, especially in artificial intelligence area, always have two stages. One is training or enrolment stage and another is testing or operational stage. In the training stage, the system should build specific models for each sample in training dataset. In the testing stage, the unknown input source is matched with stored reference models and the system selects a model which has a maximal similarity to the input.

The basic structures of speaker recognition system is shown in Figure 1. It is easy to conclude that feature extraction and feature matching are the key components in this system. Feature extraction is the process that extracts eigenvectors from the audio. Feature matching tries to identify the unknown speaker by comparing features from voice with models trained previous. Section 2 and section 3 will demonstrate them in detail.

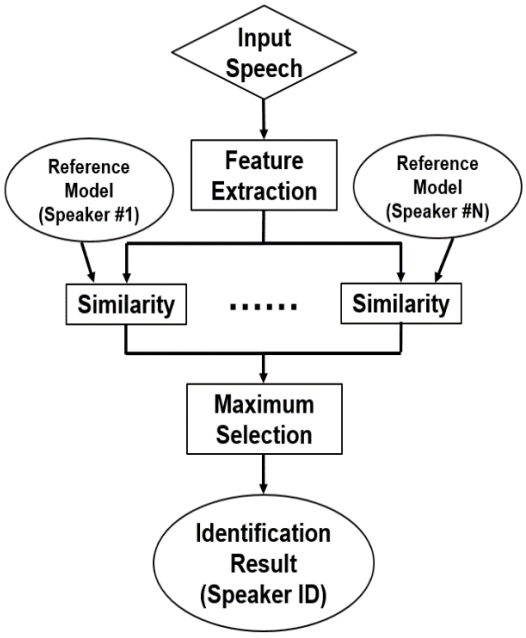

Figure 1. Basic structure speaker recognition system

An important application of speaker recognition technology is forensics. Much of information is exchanged between two parties in telephone conversations, including between criminals, and in recent years there has been increasing interest to integrate automatic speaker recognition to supplement auditory and semi-automatic analysis methods.

Not only forensic analysts but also ordinary persons will benefit from speaker recognition technology. It has been predicted that telephone-based services with integrated

[1] "British English definition of voice recognition". Macmillan Publishers Limited. Retrieved February 21, 2012.

speech recognition, speaker recognition, and language recognition will supplement or even replace human-operated telephone services in the future. An example is automatic password reset over the telephone. The advantages of such automatic services are clear much higher capacity compared to human-operated services with hundreds or thousands of phone calls being processed simultaneously.

Over the last two decades, automatic speaker recognition has made a great progress. Researchers use several features and models to represent voiceprint for specific speaker. For example, CRSS2 recently published an article about using UBM-based LDA for speaker recognition and Sergey Novoselov3 etc. demonstrated the challenge in the NIST i-vector. In addition to exploring new features and new models, people also try to implement the speaker recognition technique into some commercial areas. In fact, the focus of speaker recognition research over the years has been tending towards such telephony-based applications.

This paper not only tries to improve the efficiency of training by using a short fragment whose duration is around 1 second, but also distinguish different speaker in telephone which can be widely used in banking by telephone, telephone shopping.

2. FEATURE EXTRACTION

2.1 Introduction

Digital speech signals use ones and zeros to describe the physical properties of the acoustical waves we hear. The amount of numbers is so huge, 44100 numbers will be used to describe a 1-second audio clips with sampling frequency of 44.1 kHz. So, a set of features are extracted for further analysis from the huge amount of numbers. Selecting features to extract and how to extract them is the most critical decisions in the process of creating an automatic speaker recognition system. Several features exist for parametrically representing the speech signal for audio processing, such as Linear Prediction Coding (LPC), Perceptual Linear Predictive (PLP), Mel-Frequency Cepstrum Coefficient (MFCC), Linear Predictive Cepstrum Coefficient (LPCC), and others. MFCC will be employed in this system.

3.2 Mel-frequency cepstrum coefficients

Mel-frequency cepstrum coefficients (MFCCs) are widely used as features in audio information retrieval. MFCCs are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. Mel-frequency cepstrum (MFC) is made up by MFCCs collectively. In the MFC, the frequency bands are equally spaced on the Mel-frequency scale.

Figure 2 demonstrates procedures of deriving MFCC.

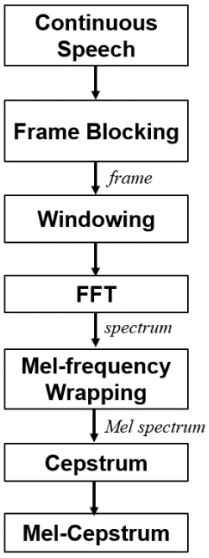

Figure 2. Procedures of deriving MFCCs

2.2.1 Frame Blocking

This process segments the continuous speech signal into frames of N samples, with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N – M samples. This step ends when all the speech is accounted for within one or more frames. In this system, values for N and M are N = 256 and M = 100.

2.2.2 Windowing

It is necessary to window every frames because it can minimize the signal discontinuities at the beginning and end of each frame. If we define the window as $w(n), 0 \le n \le N-1$, where N is the number of samples in each frame, then the result of windowing is the signal:

$$y_i(n) = x_i(n)w(n),\ 0 \le n \le N-1$$

Hamming window is used in this system. The form of Hamming window is:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

[2] Chengzhu Yu, Gang Liu and John H. L. Hansen, "Acoustic Feature Transformation using UBM-based LDA for Speaker Recognition". *Interspeech*, 2014, 1851 - 1854.

[3] Sergey Novoselov, Timur Pekhovsky, Konstantin Simonchik, "STC Speaker Recognition System for the NIST i-Vector Challenge", *The Speaker and Language Recognition Workshop*[J], 16-19 June 2014, 231-240

2.2.3 Fast Fourier Transform

Fast Fourier Transform (FFT) is a fast algorithm to implement the Discrete Fourier Transform (DFT) which converts samples to frequency domain. It is defined on the set of N samples x_n:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N} \quad , \quad k = 0, 1, 2, \ldots, N\text{-}1$$

The resulting sequence $\{X_k\}$ can be explained as follow: positive frequencies $0 \le f \le F_s/2$ responds to values $0 \le n \le N/2 - 1$, while negative frequencies $-F_s/2 \le f \le 0$ responds to values $N/2 + 1 \le n \le N - 1$. F_s represents the sampling frequency of audio.

2.2.4 Mel-frequency Wrapping

Research on psychophysical has proven that human perception of the frequency contents of sounds does not follow a linear scale. Mel-frequency scale which is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000 Hz is defined to describe the human perception.

A filter bank, shown in Figure 3, is created for simulating the mel spectrum. The filter bank are composed by several triangular bandpass filters and the bandwidth is determined by a constant mel-frequency interval. The number of mel spectrum coefficients is 20 in the system.

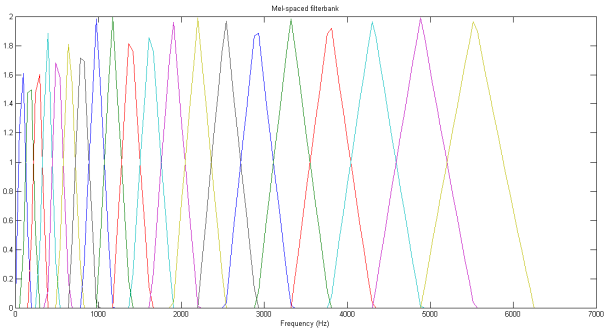

Figure 3. Mel-Spaced Filter Bank

2.2.5 Cepstrum

Log mel Spectrum is transformed back to time domain in this step. For coefficients of mel spectrum are real numbers, Discrete Cosine Transform (DCT) is used for converting. After taking the DCT of the list of mel log powers, the resulting spectrum, MFC, provides MFCCs as its amplitudes.

3. FEATURE MATCHING

3.1 Introduction

The core of the system is pattern recognition. The goal of pattern recognition is to classify patterns into one of a number of categories or classes. In our system, sequences of acoustic vectors which are extracted from speech are patterns and individual speakers are classes. The classification procedure in our case is applied on extracted features, thus it can be also referred to as feature matching.

The feature matching techniques used in speaker recognition include Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). The system will use VQ for its ease of implementation and high accuracy. VQ maps 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 by its center called a codeword. The collection of all codewords is called a codebook. Figure 4 uses two speakers and two dimensional vectors to show a basic structure of VQ process.

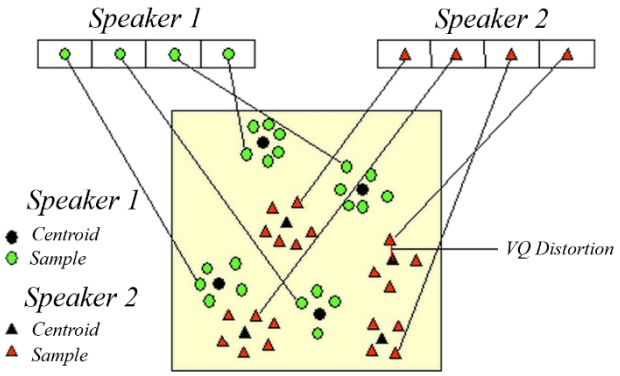

Figure 4. Structure of VQ codebook formation[4]

In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his training acoustic vectors according to the Linde–Buzo–Gray algorithm. In the Figure 4, the result centroids, which are also known as codewords, are black circles and black triangles for speaker 1 and 2. The distance from a vector to the closest codeword of a codebook is called VQ-distortion. Figure 5 shows a codebook construction for vector quantization. The original training set consisting of 5000 vectors is reduced to a set of K = 64 code vectors (centroids).

[4] F.K. Song, A.E. Rosenberg and B.H. Juang, "A vector quantisation approach to speaker recognition", AT&T Technical Journal, Vol. 66-2, pp. 14-26, March 1987.

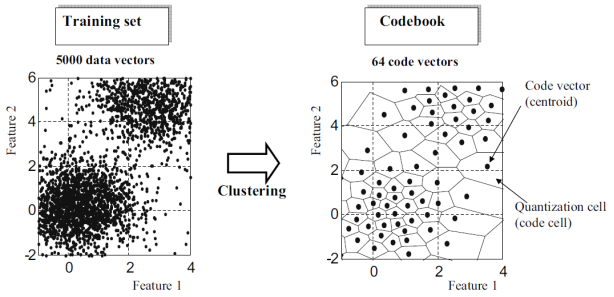

Figure 5. Codebook construction for VQ.

In the recognition phase, an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is calculated. The speaker corresponding to the VQ codebook with smallest total distortion is identified as the speaker of the unknown voice.

3.2 Linde–Buzo–Gray algorithm

After extracting features from training fragments, a specific VQ codebook is built for each speaker using those training features. In 1980, Linde, Buzo, Gray extended k-means algorithm by initializing and achieving better performance in terms of minimizing the total within class distance. The Linde–Buzo–Gray algorithm5, introduced by Yoseph Linde, Andrés Buzo and Robert M. Gray, is a vector quantization algorithm to derive a good codebook.

The stepwise working of Linde–Buzo–Gray algorithm is as follows:

Step 1:

Find the sample mean or what we call the centroid z_1 for the entire data set and it is proven to minimize the total within class distance (total mean square distortion) for a single prototype.

Step 2:

Double the size of the codebook according to splitting each centroid. The splitting rule is:

$$\begin{cases} z_n^+ = z_n(1+\varepsilon) \\ z_n^- = z_n(1-\varepsilon) \end{cases}$$

where n varies from 1 to the current size of the codebook and ε is a constant. In this system, $\varepsilon = 0.01$.

Step 3:

Find the nearest centroid for each training vector and assign the vector to that centroid.

Step 4:

Use the centroid of the training vectors assigned to that cluster to update the centroid of each cluster.

Step 5:

Repeat step 3 and step 4 until the average distance is lower than the threshold.

Step 6:

Repeat step 2, step 3 and step 4 until the size of codebook reaches M which are the number of training speakers.

Figure 5 demonstrates steps of LBG algorithm directly. In Figure 5, "Compute D (distortion)" sums the distances of all training vectors in the nearest-neighbor search. D is the current distance and D' represents the distance in the previous stage.

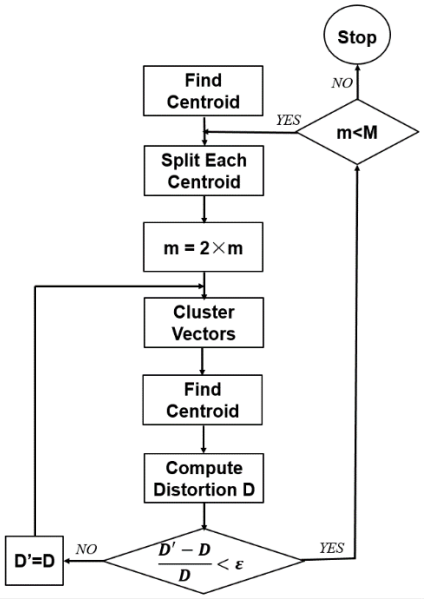

Figure 6. Flow diagram of the LBG algorithm[6]

4. EXPERIMENT AND RESULT

4.1 Dataset

The system employs five datasets: one training dataset and four test dataset. Each dataset contains eight audio clips. Because it is hard to collect a high quality audio when the system works in daily life, all of the audio fragments employ only 8 kHz as their sampling frequency. Details of all of the dataset have been provided in Table 1.

Name of Dataset	Duration	Recording Condition	Random Noise
Training	1s	Studio	NO
OriginTest	4s	Studio	NO

[5] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. on Communication, Vol. COM-28, pp. 84-95, Jan. 1980.dsa

[6] L. R. Rabinerand B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993.

NoiseTest	4s	Synthesis in Matlab	YES
Telephone Test	4s	Synthesis in Matlab	NO
Telephone NoiseTest	4s	Synthesis in Matlab	YES

Table 1. Details of datasets used in the experiment.

To simulate the telephone condition, a bandpass filter is employed for original audio. Intensity of signals out of the frequency range, 100 Hz to 3600 Hz, is decreased. To simulate the noise in daily life, random noise are added in specific dataset.
Another characteristic of datasets is that the texts which speakers say are independent between training and test. So, the system works under a content-independent condition.
Furthermore, to improve the efficiency of system, frames are used less in the training phase. So, the duration of training samples are less than testing samples.

4.2 Results

	Speaker Gender	Original Result	Noise Result	Ground Truth
1	Male	2	2	2
2	Male	3	3	3
3	Male	4	4	4
4	Male	6	6	6
5	Female	6	2	1
6	Female	7	7	7
7	Male	8	8	8
8	Male	3	2	5
Accuracy		75%	75%	

Table 2. Result of the OriginTest and NoiseTest dataset.

	Speaker Gender	Original Result	Noise Result	Ground Truth
1	Male	2	5	2
2	Male	4	5	1
3	Male	4	4	4
4	Male	5	5	5
5	Female	4	4	6
6	Female	4	4	7
7	Male	8	5	8
8	Male	3	3	3
Accuracy		50%	37.5%	

Table 3. Result of the TelephoneTest and TelephoneNoiseTest dataset.

4.3 Discussion

Table 2 shows that the system has a good performance (75%) when recognizing speakers in the recording situation. However, when the system tries to identify specific speakers in telephone, shown in Table 3, the accuracy of recognition decreases (50%). However, after comparing the noisy scene and no-noisy scene, 62.5% for no-noisy scene and 56.25% for noisy scene, it seems that noise has a little effect to the system.

So, the recognition accuracy of automatic speaker recognition system under controlled conditions is high. However, in practical situations many negative factors are encountered including mismatched handsets for training and testing, limited training data, unbalanced text, background noise and non-cooperative users.

Another interesting issue is that the system has a better performance when the speaker is male.

Dataset	Accuracy for Male	Accuracy for Female	Accuracy for All
OriginTest	83.33%	50%	75%
NoiseTest	83.33%	50%	75%
Telephone Test	66.67%	0%	50%
Telephone NoiseTest	50%	0%	37.5%

Table 4. Comparison of accuracy for male and female

The features extracted from audio fragments may be a reason. MFCCs discard most information in the region of high frequency. As we all know, frequency of women's speech is always higher than men's. So, the features used in the system limit the performance of distinguish women's speech. From Table 2 and Table 3, there are eight samples from women, however, the system only recognizes twice successfully. What is more, the system never recognize men's speech as women's. So, a new feature should be explored for recognizing female.

But, we cannot say female is hard to recognize because there are only two female samples in the dataset. Hence, we need to test the system on a larger dataset before making a credible conclusion.

5. CONCLUSION AND FURTHER WORK

5.1 Conclusion

A text-independent speaker identification system for recognizing different speakers in telephone has been presented. The system extracts eigenvectors from training dataset and saves them as special models for specific speaker. Then the system can distinguish different speakers through calculating Euclidean distance. The system can distinguish specific speaker who speaks regardless of what is saying.

5.2 Further work

Firstly, a large dataset for training and test is necessary. Although the results of experiments are acceptable, the limited samples reduces the persuasion of experiments.

Future work will also deal with improving the feature extraction process. Speech signal includes many features of which not all are important for speaker discrimination. While low-level features seem to offer a simple but powerful way of describing the speech, more abstract features are necessary to explain what the organization represents. Several alternatives to estimate the perceived similarity of music have been published recently and a combination might yield superior results. [3, 4] provide more details about features.

Furthermore, an appropriate threshold can be employed for a speaker verification system which shares most modules with this speaker recognition system.

Another interesting subject is cross-language recognition in the future. [5, 6] demonstrate the subject deeply and provide some practical methods.

6. REFERENCES

[1] "British English definition of voice recognition". Macmillan Publishers Limited. Retrieved February 21, 2012.A. Someone, B. Someone, and C. Someone: "The Title of the Journal Paper," *Journal of New Music Research*, Vol. A, No. B, pp. 111–222, 2010.

[2] Chengzhu Yu, Gang Liu and John H. L. Hansen, "Acoustic Feature Transformation using UBM-based LDA for Speaker Recognition". *Interspeech*, 2014, 1851 - 1854.

[3] Sergey Novoselov, Timur Pekhovsky, Konstantin Simonchik, "STC Speaker Recognition System for the NIST i-Vector Challenge", *The Speaker and Language Recognition Workshop*, 16-19 June 2014, 231-240.

[4] F.K. Song, A.E. Rosenberg and B.H. Juang,"A vector quantisation approach to speaker recognition", AT&T Technical Journal, Vol. 66-2, pp. 14-26, March 1987.

[5] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communication*, Vol. COM-28, pp. 84-95, Jan. 1980.dsa

[6] L.R. Rabinerand B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993.

[7] Rose, P. *Forensic Speaker Identification*. Taylor & Francis, London, 2002.

[8] Wolf, J. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustic Society of America*, 51, 6 (Part 2) (1972), 2044–2056.

[9] Campbell, W., Campbell, J., Reynolds, D., Singer, E., and Torres-Carrasquillo, P. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20, 2-3 (April 2006), 210–229.

[10] Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., and Vair, C. Compensation of nuisance factors for speaker and language recognition. *IEEE Trans. Audio, Speech and Language Processing*, 15, 7 (September 2007), 1969–1978.

[11] Adami, A. Modeling prosodic differences for speaker recognition. *Speech Communication*, 49, 4 (April 2007), 277–291.

[12] Adami, A., Mihaescu, R., Reynolds, D., and Godfrey, J. Modeling prosodic dynamics for speaker recognition. *In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003) (Hong Kong, China, April 2003)*, pp. 788–791.

[13] Besacier, L., and Bonastre, J.-F. Subband architecture for automatic speaker recognition. *Signal Processing*, 80 (July 2000), 1245–1259.

[14] Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., and Reynolds, D. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 2004, 4 (2004), 430– 451.