Digital Signal Processing: Speaker Recognition Opening Report

Xinyu Zhou, Yuxin Wu, and Tiezheng Li
Tsinghua University

1 Introduction

A **Speaker Recognition** tasks can be classified with respect to different criterion: Text-dependent or Text-independent, Verification (decide whether the person is he claimed to be) or Identification (decide who the person is by its voice).[11]

Speech is a kind of complicated signal produced as a result of several transformations occurring at different levels: semantic, linguistic and acoustic. Differences in these transformations may lead to differences in the acoustic properties of the signals. The recognizability of speaker can be affected not only by the linguistic message but also the age, health, emotional state and effort level of the speaker. Background noise and performance of recording device also interfere the classification process.

Speaker recognition is an important part of Human-Computer Interaction (HCI). As the trend of employing wearable computer reveals, Voice User Interface (VUI) has been a vital part of such computer. As these devices are particularly small, they are more likely to lose and be stolen. In these scenarios, speaker recognition is not only a good HCI, but also a combination of seamless interaction with computer and security guard when the device is lost. The need of personal identity validation will become more acute in the future. Speaker verification may be essential in business telecommunications. Telephone banking and telephone reservation services will develop rapidly when secure means of authentication were available.

Also,the identity of a speaker is quite often at issue in court cases. A crime victim may have heard but not seen the perpetrator, but claim to recognize the perpetrator as someone whose voice was previously familiar; or there may be recordings of a criminal whose identity is unknown. Speaker recognition technique may bring a reliable scientific determination.

Furthermore, these techniques can be used in environment which demands high security. It can be combined with other biological metrics to form a multi-modal authentication system.

In this task, our goal is to build a proof-of-concept text-dependent speaker recognition system with GUI support. Hopefully, we would like to extend its ability to a text-independent speaker recognition system.

2 Approach

The task of speaker recognition can be considered as a task of classification. Thus a number of machine learning techniques can be applied to this task. In general, this task should cover the following three topics:

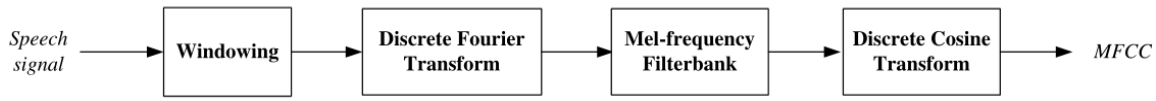
1. Feature Extraction: process the voice signal and extract acoustic features that could describe the acoustic characteristics of speakers, which can be correlated in some extend to model latter used.

2. Acoustic Model: provide the functionality of registration as well as identification or verification.

3. Evaluation: Evaluate our approach using datasets with appropriate metrics.

2.1 Feature Extraction

The task of speaker recognition is highly correlated to the task of speech recognition. In the field of speech recognition, Complex Cepstrum [2] is considered to be a concise description to the original acoustic signal. In particular, Mel-frequency Cepstral Coefficients (MFCC), is a state-of-the-art standard feature widely used in Automatic Speech Recognition (ASR) system. The general procedure for calculating MFCC is as follows:



As for speaker recognition, several different features are suggested and compared by researchers [9]. Cepstrum still proves to be a robust and effective feature in this task. Therefore, a bunch of cepstral features, including MFCC, LPCC (Linear Prediction Cepstrum Coefficients), are commonly used in speaker recognition system.[8]

We plan to implement common cepstral features used by researchers, and use a combination of some of them, according to the further observation and test on how much they contribute to the task.

2.2 Model

According to works in years, Gaussian Mixture Model (GMM) has been a common approach to acoustic modeling.[10] GMM can be used to acquire an acoustic model $P(\mathbf{O}|\mathbf{W})$, which gives the probability of a given observation $\mathbf{O}$ under certain word $\mathbf{W}$. By using GMM, this conditional probability can be well estimated by modeling the distribution as a sum of several normal distribution.

In addition to that, Hidden Markov Model (HMM) is a main-stream approach in ASR system, since it can describe sequential relation of observations.[6] We have noticed a few research-oriented open-source speech recognition tool building on HMM, such as HTK[5]. Such tools might be quite useful in buiding a speaker recognition system.

Recently, some common machine learning model are also applied to the task of speaker recognition. [1] suggested a method of using SVM in speaker recognition. Deep neural networks are also used in speech processing recently.[3]

HMM and GMM will probably be our first attempt, as they are already widely used and proven to be efficient and effective. We might also try to migrate our extracted feature to some other available models.

2.3 Evaluation

2.3.1 Dataset

There are a number of well-known databses for speaker recognition system evaluation, such as KING speaker verification[7]. [4] gives description on several common databases. But most of these databases are not free of charge. We intend to search for an freely available database, otherwise we have to build some simple test cases by our own.

2.3.2 Metrics

As we intend to distinguish different speakers, the primary goal is overall accuracy of the system we built. Furthermore, the performance may differ from speaker to speaker, which may provide with additional feedback to our approach. Thus we shall examine precision, recall and F_1 score for each speaker.

References

[1] William M Campbell et al. "Support vector machines for speaker and language recognition". In: *Computer Speech & Language* 20.2 (2006), pp. 210–229.

[2] *Cepstrum - Wikipedia, the free encyclopedia*. URL: http://en.wikipedia.org/wiki/Cepstrum.

[3] George E Dahl et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 20.1 (2012), pp. 30–42.

[4] John Godfrey, David Graff, and Alvin Martin. "Public databases for speaker recognition and verification". In: *Automatic Speaker Recognition, Identification and Verification*. 1994.

[5] *Hidden Markov Model Toolkit*. URL: http://htk.eng.cam.ac.uk/.

[6] Xuedong Huang, Alejandro Acero, Hsiao-Wuen Hon, et al. *Spoken language processing*. Vol. 15. Prentice Hall PTR New Jersey, 2001.

[7] *KING speaker verification*. URL: http://catalog.ldc.upenn.edu/LDC95S22.

[8] Richard J Mammone, Xiaoyu Zhang, and Ravi P Ramachandran. "Robust speaker recognition: A feature-based approach". In: *Signal Processing Magazine, IEEE* 13.5 (1996), p. 58.

[9] Douglas A Reynolds. "Experimental evaluation of features for robust speaker identification". In: *Speech and Audio Processing, IEEE Transactions on* 2.4 (1994), pp. 639–643.

[10] Douglas A Reynolds and Richard C Rose. "Robust text-independent speaker identification using Gaussian mixture speaker models". In: *Speech and Audio Processing, IEEE Transactions on* 3.1 (1995), pp. 72–83.

[11] *Speaker Recognition - Wikipedia, the free encyclopedia*. URL: http://en.wikipedia.org/wiki/Speaker_recognition.