Text-independent Speaker Recognition of Mandarin by Big Data Technic of GMM and VQGMM

Su-Tzu HSIEH
Zhaoqing University Economic and Management College,
Zhaoqing City Guangdong province, China

Keywords: Speaker recognition, Vector Quantization Gaussian Mixture Model, VQGMM, Text-independent.

Abstract. Speaker recognition has been long a popular research topic as to identify a specific speaker by his free speech of text-independent. This study uses Mel-frequency Cepstrum technic to simulate and extract speakers features; uses big data technic of Gaussian Mixture Model, and Vector Quantization Gaussian Mixture Model to find out the impact factors that affecting speaker identification hit rate. Research result showed training identification hit rate saturated when Gaussian mixture numbers reached specific level, then start to decreasing. Research result showed VQMM owns triple performance than GMM at training data sets. Research result suggests upcoming scholars do not take GM number as the sore tool to achieve identification hit rat.

Introduction
Speaker recognition is an important field of bio recognitions. Furthermore, speaker recognition has long been a hot topic for the collaboration between human and robot. The dream of robot can recognize me as its boss and only follow its boss’s instruction has been longed for human. This study dedicates on text-independent recognition for speakers to free from specific command words. Hit rate is one of the top issues for text-independent recognition as hit rate could drop dramatically while vocabularies increasing. This study focus on text-independent speaker recognition with high hit rate.

In research field of speaker identification scholars devote themselves to decrease noise [1], to enhance and minimizing noise [2], identifying speakers features [3], categorizing signal to reduce noise [4]. This study combines Douglas A. Reynolds and above scholar’s study to identify Mandarin speaker with Gaussian Mixture Model, GMM and Vector Quantization Gaussian Mixture Model, VQGMM. This study recorded and built database of 200 Mandarin speakers voice then did the experiment. Research result of this study found VQGMM own triple performance of train time saving than mono GMM in Mandarin speaker identification.

Literature Review
Steven F. Boll is one of the first scholar devote on speaker recognition. Steven proposed a speech enhancement of deduction of frequency spectrum to decrease noise [1]. H.L. Van Tree and Y. Ephraim at 1995 enhanced Boll’s method by decomposing a vector space of nosy speech from signal with noise subspace. The decomposing process end with an acceptable minimal noise [2]. Douglas at 1995 proposed Gussian Mixture Model, GMM to identified features of speakers [3]. Mittal and Phamdo at 200 framed speaker signal into two categories of signal dominated frames and noise dominated frames respectively. Then Mittal and Phamdo did the Karhunen-loe’ve transform for signal dominated frame and noise dominated frames separately [4]. Hu and Loizou at 2003 did a pre-whitening process, the scholar use the simultaneous diagonalizations for covariance arrays of signal and noise [5]. Follow above mentioned scholars, this study takes Vector Quantization Gaussian Mixture Model, VQGMM to identify speaker with text-independent from the database of 200 Mandarin speakers.
Research Model

The purpose of this study is to identify speaker of text-independent in a faster way. To achieve the purpose, the research process of this study composed by three steps. First step, Front-end signal process, including end point detecting, detecting the volume of power second, and transforming analog signal to digital signal. Second step is middle process of extracting out speaker’s features by process of Mel-frequency Cepstrum coefficients, Cepstral mean subtraction, and method of principal components. The third and final step of speaker identification of VQGMM.

1. Front-end signal process:
   1. endpoint detection: In order to erase silent noise and find out the start and end point of voice signal, this study detects the short-time energy signal and zero crossing rate.
   2. framing voice signal as a window sector as following definition:

\[
X(n) = S(n)\cdot W(n) \quad 0 \leq n \leq N - 1
\]

where \(S(n)\) denote as a voice signal, \(W(n)\) denote as a function of window, \(X(n)\) is the result of output, while the function of window applying Hamming window and defined as following:

\[
W(n) = 0.54 - 0.46 \cdot \cos \left[ \frac{2n\pi}{N - 1} \right] \quad 0 \leq n \leq N - 1
\]

2. Extracting out speaker’s features

Mel-Frequency Cepstrum Coefficients, MFCC: MFCC is one of most used extracting methods for speaker’s features among others of Cepstrum, Power spectrum Density...etc. Davis and Mermelstein at 1980 proposed MFCC, Mel means human’s perceived Frequency Scale. Davis and Memelstein argued human perceived lesser and lesser while frequency increasing; low frequency can extract out more features than high frequency does. Scholars also defined the Mel-frequency function as following:

\[
mel(f) = 2595 \cdot \log_{10} \left( 1 + \frac{f}{700} \right)
\]

Follow the foot step of previous scholars, this study focus on features of low frequency.

3. Speaker identification Gaussian Mixture Model, GMM, and Vector Quantization Gaussian Mixture Model, VQGMM:

3.1 This study applies GMM as following Gaussian weight average functions:

\[
P(X|\lambda) = \sum_{i=1}^{M} p_i b_i(X)
\]

Where as

\[
b_j(x; \mu_j, \sigma_j) = \frac{1}{2\pi\sigma_j^2} \exp \left\{ -\frac{(x - \mu_j)^2}{2\sigma_j^2} \right\}
\]

While

\[
\lambda = \left( p_1, ..., p_M; \mu_1, ..., \mu_M; \Sigma_1, ..., \Sigma_M \right)
\]

Fulfill condition of

\[
\sum_{i=1}^{M} p_i = 1
\]

Sum up above GMM functions as following research architecture:
3.2 This study did advance Vector Quantization Gaussian Mixture Model, VQGMM.

3.2.1 This study put speaker features into cluster analysis of two steps K-mean.
(1) Random pick up K mean center of
\[ Y = \{y_1, y_2, ..., y_K\} \] (8)

(2) Clustering speaker’s features into K group. Calculating mean vector desistance for new clustered centers.
\[ y_i = \frac{1}{N_i} \sum x_j, x_j \in C_i \] (9)

\( C_i \) is the cluster group \( i \)

(3) Do loop steps of 1-2 until \( D \) is sound.
\[ D = \sum d[x, Q(x)] \] (10)

3.2.2 After Clustering process, this study did GMM for every clustered group. To illustrate the algorithm of how GMM into advanced VQGMM as following step. First below is the function of GMM
\[ O_{GMM} = O(I, D, S, M) \] (11)

Where \( O(\bullet) \) denotes Vector complexity, \( I \) is number of iteration, \( D \) is level of density, \( S \) is vectors of speaker features, \( M \) is the number of mixture,

For VQGMM is defined as following hypotheses.
(1) Compare to GMM, VQGMM did clustering first. Hence speaker features in every clusters should be less than previous without clustering. Hence, \( I \), the number of iteration should be lesser than GMM
(2) \( D \), level of density is unchanged
(3) Every clustering group is fair shared
(4) Total number of mixture is unchanged
Hence, compare to GMM, VQGMM function is
\[ O_{VQGMM} = O_{GMM} \left( \frac{I \bullet D \bullet S}{V} \right) \]
\[ = O_{GMM} \left( \frac{I \bullet D \bullet S \bullet M}{C \bullet V} \right) \] (12)

**Experiment**

This study recorded 200 people at 75 seconds/per person. Separate speech into training and testing database, training database length is 5,10,20,30 seconds, then randomly pick 5,3,1 second for testing
sample. Use Mel-Frequency Cepstrum Coefficients to extract 12 best features, and compare GMM and VQGMM.

**GMM Experiment**

The impact of Gaussian Mixture numbers on speaker recognition rate at different training and test times. Purpose of experiment is to find out good performance of Gaussian Mixture numbers. Setting of GMM experiment 4.1 as below:

| Features | Mel-frequency Cepstrum Coefficients 12 features |
|----------|-----------------------------------------------|
| Gaussian Mixture numbers | 2, 8, 16, 32 |
| Length of training | 30 sec, 20 sec, 10 sec, 5 sec |
| Length of testing | 1 sec, 3 sec, 5 sec |

**VQGMM Experiment**

Compare GMM with VQGMM at same Gaussian Mixture numbers for the impact to testing speed. The performance of speaker recognitions showed no-difference between GMM and VQGMM, however, the training time of VQGMM own triple better than GMM from 110 to 300.

**Discussion**

This study use Mel-Frequency Cepstrum Coefficients to simulate perception of human kind. This study experimenting the impacts of different length of training speech, various Gaussian mixture numbers and different testing length onto identification hit rate.
Research result showed that identification hit raising along with Gaussian mixture numbers increasing while training length fixed. However, training identification hit rate saturated when Gaussian mixture numbers reached specific level; training identification hit rate even got decreasing after Gaussian mixture numbers got higher than the threshold level.

Research result, furthermore found training identification rate easier got saturated at Gaussian mixture small numbers when training data is small. However, testing identification hit rate getting better while increasing testing length of speech.

Research result indicates that speech features could be poor when training data is small. Poor features clustered into too many groups can make identification hit rate getting worse. Hence, research result suggests coming scholars do not take Gaussian mixture numbers as a only way to heightens identification hit rate but length of training.

Reference

[1] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," EEE International Conference on Acoustics, Speech, & Signal Processing, 1979.

[2] Y. H. L. V. Ephraim, "A spectrally-based signal subspace approach for speech enhancement," in International Conference on Acoustics, 1995.

[3] D. A. Reynolds, "Large population speaker identification using clean and telephone speech," IEEE Signal Processing Letters, Vols. 2, issue 3, pp. 46-48, 1995.

[4] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," Speech & Audio Processing IEEE Transactions on, pp. 159-167, 2000.

[5] H. Yi, "A subspace approach for enhancing speech corrupted by colored noise," Speech & Audio Processing IEEE Transactions on, pp. 334-341, 2003.

[6] D. A. Reynolds, "Large population speaker identification using clean and telephone speech," IEEE Signal Processing Letters, pp. 46-48, 1995.