Text-Independent Speaker Identification Using GMM With Universal Background Model

Jozef Polacký

Audiolab, Dept. of Telecommunications and Multimedia, Faculty of Electrical Engineering, University of Zilina, Univerzitna 8215/1, 010 26 Zilina, Slovakia

jozef.polacky@fel.uniza.sk

Abstract. State-of-the-art of speaker recognition is fully advanced nowadays. There are various well-known technologies used to process voice, including Gaussian mixture models. The paper presents our work on speaker identification from his voice. In our experiment we first extract key features from a speech signal using VOICEBOX [1] toolbox in MATLAB. These features are represented by a matrix of mel frequency cepstral coefficients (MFCC). Then, applying MSR Identity Toolbox, we build an identity for each person enrolled in our system using statistical Gaussian Mixture Model - Universal Background Model (GMM-UBM) and features extracted from speech signals. Universal Background Model improves Gaussian Mixture Model statistical computation for decision logic in speaker verification task. As a corpus, we used TIMIT database for our experiments. Finally, we compared the recognition accuracy for several different scenarios of our experiments.

Keywords

Speaker identification, GMM-UBM, MFCC features, TIMIT, Matlab

1. Introduction

During the past decades, speaker recognition has become a very popular area of research in pattern recognition and machine learning. Speaker recognition includes the identification, verification, classification and with certain extension also the speaker segmentation. It is a general term used for any procedure that involves the knowledge of a person’s identity on the basis of its voice.

Speaker identification is the process by which the identity of the speaker is not known ahead and system has to decide who it is or put the speaker in a particular group. For this identification must exist the reference speaker model.

Verification process is often used to allow or refuse access to certain resources using the spoken word. Compared with the identification of the speaker, in this case the identity of the speaker is known and the role of the system is to verify whether it is really the person for which the speaker is issued.

Speaker classification focuses mainly on gathering information about individuals from their speech varieties. It involves the process of determination whether the speaker is male or female, adult or child, or nationality speaker detection. Similarly, the identity of the speaker may be determined from the speech signal. This task is known as speaker recognition, which is one of the most studied subareas classification of speaker.

Speaker recognition methods can be further divided into two groups according to the types of utterances. Text-independent speaker recognition is a technique that examines the correctness of an identity claim to a text unknown speech utterance. There exist no two persons that sound identical due to shape of their physical parts of voice production organs or speaking manners. Text-dependent recognition requires that the speaker spoke a predetermined text that must be identical in training phase as well as in testing phase. Speaker recognition technique is used in many applications such as telephone banking, information security, forensics and so on [2].

One of the biggest problems in speaker recognition task is the mismatch between training and testing conditions caused by several reasons such as channel distortion, different microphones, transmitting channels or utilized encoder. The additive noise is one of the main causes of the performance degradation in lot of practical applications.

Nowadays, we already know a large number of different solutions to reduce this problem. One of the techniques for noise-robust Automatic Speaker Recognition (ASR) is feature enhancement method which
attempts to normalize the distorted feature or to estimate undistorted feature from the distorted speech and does not require any explicit knowledge about the noise. Cepstral Mean Normalization (CMN) or Cepstral Mean and Variance Normalization (CMVN) can be mentioned as an examples.

In recent years, the Gaussian Mixture Model (GMM) adapted from the Universal Background Model (UBM) [3] by Maximum A Posterior (MAP), has become the most popular method of modeling the acoustical space of speech context. These models represent the core technology of most of the state-of-the-art text-independent speaker recognition systems.

The remaining of the paper is organized as follows. Section 2 briefly describes the GMM-UBM model, features extraction and evaluation data set. Section 3 introduces the experiment carried out in this study. In section 4, the experimental results are presented and discussed. Section 5 concludes the paper and suggests some future studies.

2. GMM-UBM verification system

Gaussian Mixture Models used in combination with MAP adaptation [3] represent the main technology of most of the state-of-the-art text-independent speaker recognition systems. In our system the speaker models are derived from a common GMM root model, the so-called UBM, by means of MAP adaptation. Mean, weight vector and covariance matrix adaptation is performed during model training. A speaker is thus represented by the set of the adapted parameters of all the Gaussians of the UBM.

2.1 Gaussian Mixture Model

Gaussian Mixture Model [4] is stochastic model, which can be considered as a reference method for speaker recognition. The Gaussian mixture probability density function of model $\lambda$ consists of a sum of $K$ weighted component densities, given by the following equation:

$$p(x | \lambda) = \sum_{k=1}^{K} P_k N(x | \mu_k, \Sigma_k)$$

where $K$ is the number of Gaussian components, $P_k$ is the prior probability (mixture weight) of the $k$-th Gaussian component, and

$$N(x | \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right)$$

is the $d$-variate Gaussian density function with mean vector $\mu_k$ and covariance matrix $\Sigma_k$. The prior probabilities $P_k \geq 0$ are constrained as $\sum_{k=1}^{K} P_k = 1$.

For numerical and computational reasons, the covariance matrices of the GMM are usually diagonal.

Training a GMM consists of estimating the parameters $\lambda = \{P_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ from a training sample $X = \{x_1, \ldots, x_T\}$. The basic approach is to maximize likelihood of $X$ with respect to model $\hat{\lambda}$ is defined as:

$$p(X | \hat{\lambda}) = \prod_{t=1}^{T} p(x_t | \hat{\lambda})$$

The goal is to obtain Maximum-likelihood (ML) parameter estimation. The process is an iterative calculation called the Expectation-Maximization (EM) algorithm [5]. Note that K-means [6] can be used as an initialization method for EM algorithm.

In the identification process, a set of test utterances and its model is compared with each model of the training database. From each comparison between test and training model is obtained a likelihood and the model with the highest score corresponds to the unknown speaker.

Let us assume a group of speakers $S_p=1,2,\ldots,S$ represented by GMM’s $\lambda_1, \lambda_2, \ldots, \lambda_S$. Unknown speaker model is identified to each model:

$$\hat{S} = \arg \max_{1 \leq s \leq S} \sum_{t=1}^{T} \log \left( p(x_t | \lambda_s) \right)$$

2.2 Universal background model

Universal Background Model is an improvement in the field of speaker recognition using GMM. It is typically characterized as a single Gaussian Mixture Model trained with a large set of speakers using the EM algorithm.

![Fig. 2. Speaker enrollment.](image-url)

UBM is used for training of the speaker-specific model. The adapted UBM is used as the target speaker model. This process prevents from need for building the speaker model (estimating the parameters) from scratch. There are multiple ways how to adapt the UBM. It is possible to adapt one or more of its parameters as well as all parameters. Adaptation of the parameters is usually done using the MAP. The background model must be built from utterances with common characteristics in the meaning of type and quality of speech. For example, a verification system that uses only telephone channel and female speakers must be trained using only telephone speech spoken by female speakers. For a system where the...
gender composition is an unknown parameter, the model will be trained using both male and female utterances.

The following average log-likelihood formula gives the final score in recognition process [7]:

\[
LLR(X, \lambda_{\text{argmax}}, \lambda_{\text{UBM}}) = \frac{1}{T} \sum_{t=1}^{T} \{ \log(p(x_t | \lambda_{\text{argmax}})) - \log(p(x_t | \lambda_{\text{UBM}})) \}
\]

where \( X = \{x_1, ..., x_T\} \) corresponds to the set of observation or test feature vectors. The higher the score, the more likely the test features belong to the speaker-model with which they are compared.

2.3 Feature Extraction

Feature extraction is the process of extracting the speaker and language related feature vectors from the voice signal which can later be used to represent the speaker and the language that the speaker had spoken. It is also known as speech parameterization. The purpose of feature extraction phase is to extract the speaker-specific information in the form of feature vectors which are more compact and more suitable for statistical modeling and the calculation of score. A good feature vector set should have representation of all components of speaker information.

The most representative vocal tract acoustic features are Mel Frequency Cepstral Coefficients (MFCC). MFCC coefficients are mostly related to the human peripheral auditory system. The main purpose of the MFCC is to simulate the perception of the human ears [8]. According to studies, human hearing is not linear but rather it has the Mel frequency scale which is a linear spacing below 1 kHz and logarithmic scaling above 1 kHz.

At the first step, the pre-emphasized and windowed speech signal frame is converted into spectral domain by the Fast Fourier Transformation (FFT). In the second step, the magnitude spectrum is smoothed by a bank of triangular bandpass filters which emulate the critical bands of processing of the human ear. In the next step, each of the bandpass filters computes a weighted average of that subband, which is then compressed by logarithm. Finally, the log-compressed filter outputs are decorrelated using the Discrete Cosine Transform (DCT) [9].

2.4 Feature Normalization

The objective of the feature normalization technique is to compensate for the effects of environmental mismatch. The components of the fixed feature vector are scaled or warped so as to enable more effective modeling of speaker differences. Here, we used short-time mean and variance normalization (STMVN) in one experimental scenario.

In the STMVN technique, \( m \)-th frame and \( k \)-th feature space \( C(m,k) \) are normalized as:

\[
C_{\text{norm}}(m,k) = \frac{C(m,k) - \mu_s(m,k)}{\sigma_s(m,k)}
\]

where \( m \) and \( k \) represent the frame index and cepstral coefficients index, \( L \) is the sliding window length in frames. \( \mu_s(m,k) \) and \( \sigma_s(m,k) \) are the short-time mean and standard deviation, respectively, defined as [10]:

\[
\mu_s(m,k) = \frac{1}{L} \sum_{j=m-L/2}^{m+L/2} C(j,k)
\]

\[
\sigma_s(m,k) = \left( \frac{1}{L} \sum_{j=m-L/2}^{m+L/2} (C(j,k) - \mu_s(m,k))^2 \right)^{1/2}
\]

2.5 Evaluation data set

The speaker recognition tests described in this paper were evaluated on the TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT – Texas Instruments (TI) and Massachusetts Institute of Technology (MIT)) [11]. TIMIT corpus contains recordings of phonetically-balanced English speech. It was recorded using a Sennheiser close-talking microphone at 16 kHz rate with 16 bit sample resolution. TIMIT contains broadband recordings of 630 speakers of eight major dialects regions of the United States, each reading ten phonetically rich sentences resulting in 6300 sentences. The prompts for the 6300 utterances consist of 2 dialect sentences (SA), 450 phonetically compact sentences (SX) and 1890 phonetically-diverse sentences (SI). TIMIT corpus consists of training and test sets.

The training set contains 4620 utterances, but usually only SI and SX sentences are used. The test set contains 1344 utterances from 168 speakers. This speech corpus was originally designed as standard database for the speech recognition experiments for several decades and nowadays it is still widely exploited corpus for both speech and speaker recognition experiments [12].

2.6 MSR Identity Toolbox

This toolbox consists of a collection of MATLAB tools and routines. These tools includes GMM-UBM and the state-of-the-art i-vector based strategies for speaker recognition tasks. It can be very useful tool for research and development in the area of speaker recognition. This toolbox allows the utilization of ”parfor” loops so that parallel processing can speed up the recognition process [13].

3. Experiments

The experiment consisted of three different scenarios. One of the task was to examine the impact of the speech signal preprocessing and normalization of acoustic features by STMVN on performance of the system for the speaker identification. The second task was to examine the impact of the components number of GMM models of individual
speakers. These models were obtained by adaptation of the relevant GMM-UBM model. For each session, the adaptation of all model parameters was applied. Thus, the vectors of weights, mean vectors and the covariance matrix were adapted. Relevance factor was set to 10. For training model was used EM algorithm. The number of iterations of the K-means algorithm was set to 100.

TIMIT database were used for training the UBMs for the proposed system. 630 speakers in total were used for the background model training. For this training, all sentences from all speakers were exploited. For the testing phase, 190 speakers from the first three dialect regions from training set were used. For the speaker specific model training, 5 out of 10 sentences per speaker were utilized and the remaining 5 sentences were used for testing.

This means that two adapted GMM models were obtained for each speaker. The first was adapted by recordings 1-5 from each speaker. The remaining 6-10 recordings were used to test. Reversely, the second one was adapted using recordings 6-10 per speaker, and recordings 1-5 were used to test. Overall, 10x190 tests for each UBM model (different number of components) were conducted. For each session was evaluated recognition performance by calculating success rate in percentage. These values were statistically averaged and was calculated mean value and variance.

In the first scenario, there were not applied any preprocessing of speech recordings and “raw” recordings were parameterized. For all experiments, 12 MFCC features (excluding log-energy and 0-th coefficient) were used. The analysis frame length was 30 ms with a frame shift of 15 ms with Hamming window.

In the second scenario, the direct component of speech sample was removed, pre-emphasis filter was used and also the normalization of the speech signal was performed. Then silence frames were removed according to the VAD labels. Therein, a basic energy-based VAD was combined with a zero-cross rate based VAD to detect silence frames and MFCC coefficients were calculated.

In the last scenario, feature normalization technique STMVN were additionally applied. Then, a gender-independent UBM models with $2^3-2^{10}$ component GMM models were trained for each scenario. Finally, target speakers models were adapted from this UBM models.

4. Experimental results

Figure 1 shows the dependence between success rate of the recognition process and number of components GMM models. It can be seen that success rate grows with the number of components for scenario 3. For the first two scenarios, success rate grows after 256 component model. Then this value slightly decreases for the other two models constellation. From this perspective, together with the time demands associated with the process of training UBM it can be considered as the most appropriate model the one with 256 components in this case. This means that a system with a large number of components of the GMM model may not have always better performance in comparison with the less components model. The similar effect can be seen at the variance of individual sessions. This implies that the results obtained with the models more components are statistically more balanced and accurate. From the results it can be seen that the features normalization technique STMVN caused a deterioration of performance in comparison with the other two experiments. The best results were achieved with the use raw speech recordings. Table 1 shows the complete results for each session.

<table>
<thead>
<tr>
<th>GMM components</th>
<th>Average success rate [%]</th>
<th>Variance [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>79.84</td>
<td>56.58</td>
</tr>
<tr>
<td>16</td>
<td>87.37</td>
<td>69.21</td>
</tr>
<tr>
<td>32</td>
<td>91.26</td>
<td>76.47</td>
</tr>
<tr>
<td>64</td>
<td>93.95</td>
<td>81.16</td>
</tr>
<tr>
<td>128</td>
<td>95.32</td>
<td>84.63</td>
</tr>
<tr>
<td>256</td>
<td>97.00</td>
<td>85.68</td>
</tr>
<tr>
<td>512</td>
<td>95.95</td>
<td>86.21</td>
</tr>
<tr>
<td>1024</td>
<td>95.47</td>
<td>87.00</td>
</tr>
</tbody>
</table>

Fig. 3. Average success rate for each scenario.

Fig. 4. Variance of success rate.

Tab. 1. Summary of results for individual conditions.
5. Conclusions and future work

In this paper, an analysis of the impact of the number components of GMM (UBM) model on performance of the system for the speaker identification have been executed. We construct a system using the UBM models for different scenarios described above. We plan to do this analysis with i-vector based strategies as our further step in this research.

References


[8] REYNOLDS, D.A. An overview of Automatic Speaker Recognition Technology, MIT Lincoln laboratory, 244 wood St. Lexinton, MA 02140, USA, IEEE, 2002


About Author

Jozef POLACKY was born in Povazska Bystrica, Slovakia in the 1988. In 2013 he finished MSc at University of Zilina, Faculty of Electrical Engineering, Department of Telecommunications and Multimedia. Currently he studies doctor degree. His research is focused on speaker recognition for biometric purposes.