I-Vector Speaker Verification Based on MAP

Huizong Feng, Yunfang Wang [a]

School of Chongqing University of Posts and Telecommunications, Chongqing 400069, China

[a]416240202@qq.com

Abstract

This paper presents a text-independent speaker identification approach based on i-vector and maximum a posteriori (MAP) algorithm. The preprocessed speech signals are divided into some chunks, then enhanced by MAP. After denoising, we train and test i-vector model. The numerical experiments are carried out to verify the theoretical results and clearly show that our identification method has an acceptable accuracy.

Keywords

Speaker Identification, I-Vector, Maximum A Posteriori.

1. Introduction

Speaker verification is the process to accept or reject an identity claim by comparing two speech samples: one that is used as reference of the identity and the other that is collected during the test from the person who makes the claim [1]. In the domain of speaker verification, many models have been proposed over time, such as Gaussian mixture models (GMM), GMM with universal background model (GMM-UBM), joint factor analysis, etc.

Among these models, the concept of i-vectors [2] has become quite popular in recent years. This approach tries to improve the JFA by combining the inter and intra domain variability and modeling it in same low dimensional total variability space. The recently developed paradigm of i-vector extraction provides an elegant way to obtain a low dimensional fixed-length representation of a speech utterance that preserves the speaker-specific information. A factor analysis (FA) model is used to learn a low-dimensional subspace from a large collection of data. A speech utterance is then projected into this subspace and its coordinates vector is denoted as i-vector. The low dimensional nature of this representation is very appealing and has opened the door for new ways to explore one of the key problems in speaker verification, that is, how to decompose a speech signal into a speaker-specific component and an undesired variability component, which is often referred to as the channel component [3].

In the speech processing stage, single channel speech enhancement technique is used for enhancement of the speech degraded by additive background noises [4]. Different types of algorithms are proposed for speech enhancement such as spectral subtraction, wiener filtering, statistical based methods, noise estimation algorithms [5]. As one of the methods proposed for enhance the noisy speech signal, maximum a posteriori (MAP) has been widely used in automatic speaker recognition since it obtains point estimation of those volumes hard to observe according to the experience.

In this paper, we aim to investigate a text-independent speaker verification approach based on i-vector and MAP algorithm. The input speech signals are pre-emphasized, windows, to complete preprocessing procedure. These signals are enhanced by MAP, extracted Mel-frequency cepstral coefficients (MFCC) [6] as features, then trained and tested in i-vector model to complete speaker verification.

2. Proposed Speaker Verification Method

2.1 Maximum a Posteriori (MPA) Estimators

This section introduces the proposed method used in text-independent speaker verification. The process of our method is shown in Fig. 1.

As shown in Fig. 1, the collected utterances are enhanced by MAP. MFCC features are used to fit a Gaussian mixture model (GMM) to observations, and compute the sufficient statistics for observations given the UBM. Finally computes verification scores for i-vector trials using the PLDA model. The main techniques involved in this process are given specifically as below.

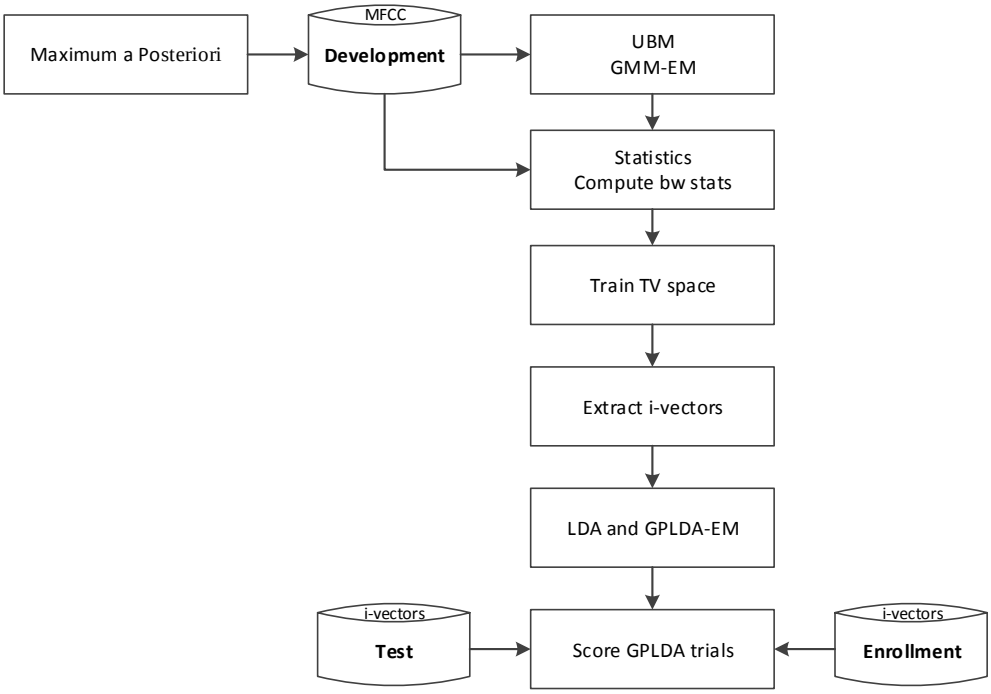

Fig.1 Process of the proposed speaker identification technique

Assume that the speech with noise signal as:

$$y(n) = x(n) + d(n) \tag{1}$$

where $x(n)$ and $d(n)$ denotes pure speech signal and noise signal, respectively.

A proportional cost function results in an optimal estimator which is median of posterior probability density function (PDF). For a "Hit-or-Miss" cost function the mode or maximum location of posterior PDF is the optimal estimator. The latter is termed the maximum a posterior (MAP) estimator. In the MAP estimation approach we choose $\hat{\theta}$ to maximize posterior PDF or

$$\hat{\theta} = \arg\max_{\theta} p(\theta / x) \tag{2}$$

This is shown to minimize the Bayes risk for a "Hit-or-Miss" cost function. In finding maximum of $p(\theta / x)$ we can observe that

$$p(\theta / x) = \frac{p(x / \theta)\, p(\theta)}{p(x)} \tag{3}$$

So an equivalent maximization is of $p(x / \theta)\, p(\theta)$. This is reminiscent of Maximum Likelihood Estimator (MLE) except for the presence of prior PDF. Hence the MAP estimator is given by

$$\hat{\theta} = \arg\max_{\theta} p(x / \theta)\, p(\theta) \tag{4}$$

or

$$\hat{\theta} = \arg\max_{\theta}[\ln(p(\theta / x)) + \ln(p(\theta))] \quad (5)$$

2.2 I-Vector Approach.

Dehak [2] observed that the channel dependent supervector also models the speaker features. To take this finding into account, a new approach was proposed where there was no distinction between channel and speaker variabilities.

A new low dimensional total variability space T was introduced to account for both the variabilities, where M is given by:

$$M = m + Tw \quad (6)$$

where m is the UBM supervector (speaker and channel independent supervector), w is a normally distributed random vector in this space and T is a low ranked rectangular matrix. The factors of w are also called as total factors. The new vectors are known as identity vectors or i-vectors.

In this approach, it is assumed that M is distributed normally with T as its covariance matrix and m as its mean vector. Here total variability matrix T is assumed that the utterances of the same speaker are produced by different speakers.

3. Numerical Experiments

In this section, some numerical experiments are carried out to evaluate the performance of the proposed speaker verification approach. We use TIMIT corpus for training and testing i-vector model. There are 630 speakers in TIMIT, where 192 females and 438 males. We choose 530 speakers to train background model, and the remaining 100 speakers (30 females and 70 males) for testing. Each speaker has 10 speeches. All these speech signals are framed by a series of Hamming windows. The frame size is 30 ms and the overlapped length is 15 ms. The sampling rate is set as 16 kHz.

We choose white noise, speech babble, pink noise and volvo in NOISEX-92 as noise signal. Fig. 2 shows the DET curves of i-vector speaker verification with and without MAP, which the red one is without MAP procedure and the blue one is preprocessed by MAP.

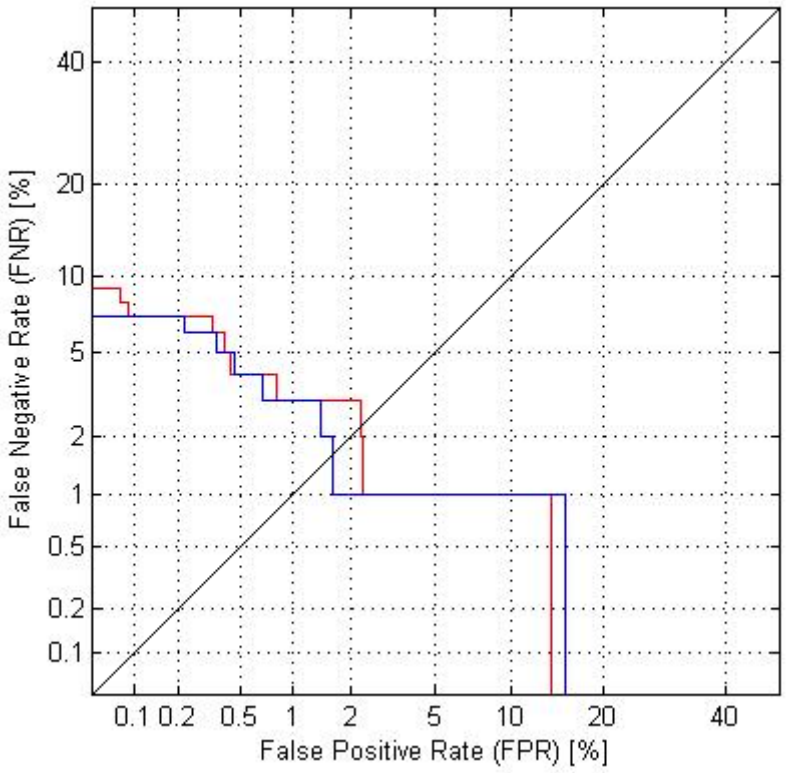

Fig. 2 DET curves of i-vector speaker verification with and without MAP

4. Conclusion

This paper has focused on the scenario of a systematic approach to perform speaker verification based on MAP and i-vector. Through this elastic and robust sequential data matching, the results of the numerical experiments indicate that the proposed method has an acceptable recognition rate with high accuracy.

References

[1] Larcher, Anthony, et al, Text-dependent speaker verification: Classifiers, databases and RSR2015, Speech Communication 60 (2014) 56-77.
[2] Dehak, Najim, et al, Front-end factor analysis for speaker verification, Audio, Speech, and Language Processing, IEEE Transactions on 19.4 (2011) 788-798.
[3] Garcia-Romero, Daniel, and Carol Y. Espy-Wilson, Analysis of i-vector Length Normalization in Speaker Recognition Systems, Interspeech. 2011.
[4] Sivaprasad, N., & Kumar, T. K., A Survey on Statistical Based Single Channel Speech Enhancement Techniques, International Journal of Intelligent Systems and Applications (IJISA) 6.12 (2014) 69.
[5] Loizou, Philipos C, Speech enhancement: theory and practice, CRC press, 2013.
[6] Srinivasan, A, Speaker identification and Verification using Vector quantization and Mel frequency Cepstral Coefficients, Research Journal of Applied Sciences, Engineering and Technology 4.1 (2012) 33-40.