GMM-UBM for text-dependent speaker recognition

Conference Paper · July 2012

DOI: 10.1109/ICALIP.2012.6376656

3 authors, including:

Qingyang Hong
Xiamen University
35 PUBLICATIONS **142** CITATIONS

GMM-UBM for Text-Dependent Speaker Recognition

Wanli Chen, Qingyang Hong* and Ximin Li
Cognitive Science Department, Xiamen University, Xiamen, China, 361005
Fujian Key Laboratory of the Brain-like Intelligent Systems (Xiamen University), Xiamen, China, 361005
qyhong@xmu.edu.cn

Abstract

Traditional Text-Dependent Speaker Recognition (TDSR) systems model the user-specific spoken passwords with frame-based features such as mel frequency cepstral coefficient (MFCC) and use Dynamic Time Warping (DTW) or hidden Markov Model (HMM) classifiers to handle the variable length of the feature vector sequence. However, DTW can't deal with cross-channel issue while HMM needs more computational complexity and storage space. In this paper, we introduce text-independent framework GMM-UBM into text-dependent field. It not only solves intersession problem but also a compromise between model accuracy and computational cost. Moreover, a more accurate UBM will get lower EER. A new UBM initialization method, LBG-VQ-EM, will be proposed. Experiments shows that it is better than conventional initialization way like K-means. And we also compare the performance of GMM-UBM and DTW, and two stacked methods of training utterances: frame-based and wave-based. The experimental results showed the performance of GMM-UBM exceeded that of DTW, and that of frame-based outperformed that of wave-based.

1. Introduction

Conventional Text-Dependent Speaker Recognition (TDSR) systems use a unique password for each user and typically obtain the identity from the spoken password by extracting frame-by-frame spectral features like MFCC. Owning to the natural variations of the speaking rate, even if the same speaker says the same password twice, the length of the feature-vector sequence varies from one password to another. To compare such two length-variable password, traditional TDSR systems employ dynamic classification methods, such as DTW [1] and HMM [2]. DTW dynamically compares feature-vector sequence between test data and training data. HMM captures the speaker-specific speech dynamics as adapted HMMs, one for each speaker and calculates log-likelihood between trial feature-vectors and target-speaker HMM. DTW has the advantages of less computational complexity and storage space but is not robust. HMM is a more sophisticated statistical model and needs more training data and corpus to deal with inter-session issue.

To find a compromise between DTW and HMM, the classical GMM-UBM [3] for Text-Independent Speaker Recognition (TISR) is taken into consideration in our paper. Since GMM is a simplified version of HMM, it needs less training and test data. Moreover, the algorithms used in TISR can also be applied, such as NAP [4] and JFA [5], which are the mainstream methods in the recent NIST Speaker Recognition Evaluation.

This paper will be organized as follows: Session 2 has a brief depiction of GMM-UBM framework. In Session 3, an improved UBM initialization method will be proposed to reduce EER effectively. Finally, we carry out some experiments to verify our algorithms. Firstly we compare the UBM initialization method between K-means and LBG-VQ-EM; Secondly we illustrate the results of using three, four or five password utterances to train a target-speaker GMM; Thirdly, two ways of joining training utterances together are compared: frame-based and wave-based; Experiments between GMM-UBM and DTW will also be executed.

2. GMM-UBM for TDSR

GMM-UBM is the predominant approach used in Text-Independent Speaker Recognition systems. Given a segment of speech Y and a target-speaker GMM S, the speaker verification task consists in determining if Y was spoken by S or not. This task is often stated as basic hypothesis test between two hypotheses: Y is from the hypothesized speaker S (H_0), and Y is not

from the hypothesized speaker *S* (H_1). A log-likelihood ratio (LLR) between these two hypotheses is counted and compared to a decision threshold θ . The LLR test is given by:

$$LLR = logP(Y|H_0)\text{-}logP(Y|H_1)$$

Where *Y* is the feature vector sequence. If LLR is not less than decision threholdθ , H_0 is accepted, otherwise H_1 is accepted. As usual H_0 is learned from the corresponding speaker voice, while H_1 is obtained using speech data from a large set of impostor speakers.

The principle of GMM-UBM makes sure that it can be applicable to TDSR. Training data is adapted from UBM via MAP to get a target-speaker GMM representing not only speaker identity but also speech content. Therefore, when training corpus is vocal password, the target GMM will be used as vocal password of speaker. However, a short vocal password may cause singularity in adapting phase. The solution is to collect 3-5 vocal utterances, covering same content and different speech speed and tone, and concatenate them. Two distinct concatenation modes lead to different EERs, which will be explained in detail in session 4.

3. UBM Initialization: LBG-VQ-EM

In TISR, two models are required for the baseline system: one is speaker model and the other is Universal Background Model (UBM) that represents all possible speakers.

In our paper, a better training method for UBM will be proposed. UBM training is divided into two phases: initialization to get a raw model and iteration to obtain an accurate GMM. It is crucial to initialize the UBM to get a good performance. Although the conventional initialization method such as K-means and LBG[6] can converge fastly, they are hard-decision methods which can not initialize the UBM accurately. Therefore, more effective initialization technique is bring forward: LBG-VQ combined with EM [7].

LBG-VQ design algorithm is an iterative algorithm which alternatively solves two optimality criteria: Nearest Neighbor Condition and Centroid Condition. The algorithm requires an initial codebook, obtained by the splitting step. In this method, an initial code vector is set as the average of the entire training feature sequence. This code vector is then split into two. The iterative algorithm is run with two vectors as the initial codebook. After splitting, VQ iteration is carried out to adjust the centroid of each codebook, and then EM is executed for a soft-decision of all the feature vectors. The process is repeated until the desired number of code vectors have been gained.

Empirically, both the number of the VQ iteration and EM iteration are set 3.

The main procedure of initializing a UBM is described as follows:

Algorithm: UBM initialization LBG-VQ-EM

Input: features vectors $X = \{x_1, x_2 \ldots x_T\}, \varepsilon > 0$ to be convergence condition and expected mixture component number M.

Output: an initialized UBM.

Step 1: Let N=1 and $c_1^* = \frac{1}{T}\sum_{m=1}^{T} x_m$, calculate the distortion $D_{ave}^* = \frac{1}{Tk}\sum_{m=1}^{T}\left\|x_m - c_1^*\right\|$ (k is the feature-vector dimension).

Step 2: Split N code vectors to 2N. Splitting formula is shown as $c_i^{(0)} = (1+\varepsilon)c_i^*$, $c_{N+i}^{(0)} = (1-\varepsilon)c_i^*$ (i=1,2…N), set N=2N.

Step 3: VQ iteration. Let $D_{ave}^{(0)} = D_{ave}^*$. Set the iteration index i=0 and MAX_ITERATION_NB to 3.

a) For m=1,2…T, find the minimum value of $\left\|x_m - c_n^{(i)}\right\|$ over all n=1,2…N. Let n^* be the index which achieves the minimum. Set $Q(x_m) = c_{n^*}^{(i)}$

b) For n=1,2,…,N, update the code vector $c_n^{(i+1)} = \frac{\sum_{Q(x_m = c_{n^*}^{(i)})} x_m}{\sum_{Q(x_m = c_{n^*}^{(i)})} 1}$.

c) Set i=i+1.

d) Calculate $D_{ave}^{(i)} = \frac{1}{Tk}\sum_{m=1}^{T}\left\|x_m - Q(x_m)\right\|$

e) If $\left|(D_{ave}^{(i-1)} - D_{ave}^{(i)}) / D_{ave}^{(i-1)}\right| > \varepsilon$, go back to Step a).

f) Set $D_{ave}^* = D_{ave}^{(i)}$. For n=1,2,…N, set $c_n^* = c_n^{(i)}$ as the final code vectors.

g) If i is equal to MAX_ITERATION_NB, then go to Step 4.

Step 4: According to the final code vectors, use EM algorithm to gain a soft-decision classification.

Step 5: If N is equal to M, output the initialized UBM, otherwise go back to Step 2.

As shown above, to acquire a more accurate initialized model, EM iteration is run until it converges

or achieves the expected maximum iteration number. An more representative UBM will be trained with our proposed method.

4. Experiment Setup

All the experimental data are collected with PC microphone by ourselves. There are 147 female utterances and 454 male utterances as UBM background data. In light of imbalance of female-male data, gender-dependent UBMs [3], one for female and one for male, will be trained via LBG and EM algorithm. Due to the amount of development data, The GMM-UBM consists of 128 Gaussian mixture components.

All trials will be same-sex trials. We gather 22 female and 36 male with the same vocal password counting from 0 to 9, each speaker having six utterances. We randomly extract three to five utterances to train target-speaker GMM only the means adapted from UBM via MAP with a relevance factor of r=16, and the remaining will be as positive trials. Statistically, for female there are 66, 44 and 22 positive trials when training utterances number is 3, 4 and 5 respectively, and 2772 impostor trials; for male there are 108, 72 and 36 positive trials when training utterances number is 3, 4 and 5 respectively, and 7560 impostor trials.

Before obtaining the MFCC vectors, Voice Activation Detection (VAD) based on window energy is performed on the raw utterances. A 16-dimensional MFCC vector is extracted from the pre-emphasized speech signal every 10 ms using 20 ms Hamming window. The mel-cepstral vector is computed using a simulated triangular filter-bank on the DFT spectrum. Band-limiting is performed by retraining only the filter-bank outputs from the frequency rang 250 Hz-3800 Hz. Delta-cepstral coefficients are then computed over a 2 frame span and appended to the cepstral vector, producing a 32 dimensional feature vector, followed by Cepstral Mean Subtraction (CMS) [8] and Cepstral Variance Normalization (CVN) [9] to mitigate the channel distortion.

Then two UBM initialization methods are compared: K-means and LBG-VQ combined with EM. The EER result is demonstrated in Figure 1. With the same speech corpus, different EERs illustrate LBG-VQ combined with EM is better than K-means both in principle and in practice. The following experiments adopt LBG-VQ-EM initialization method.

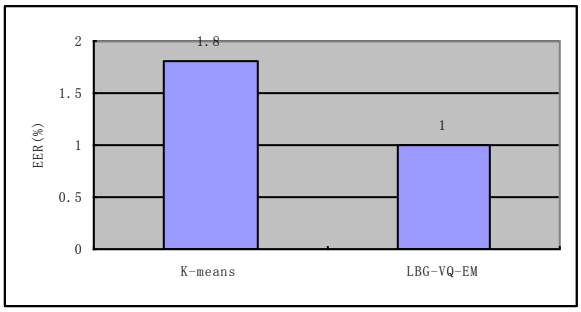

Figure 1. Different UBM initialization methods

Afterwards, we compare using different training utterances number to train target-speaker GMM. Figure 2 demonstrates our expectation that more training utterances can promote recognition performance.

Figure 2. Different number of training utterances

From the figure above, we can learn that when training utterances number increases from 3 to 5, the EER decreases from 2.2% to 1%. In addition, the reason why performance of male outgoes that of female is the background data for training male UBM is more sufficient than female's.

In view of timing sequence of training utterances, two concatenation methods are compared: frame-based and wave-based: The first one is to extract MFCC of each utterance and then join the MFCC together, while the second is to directly concatenate all training utterances followed by extracting the MFCC of the whole utterance. The result is shown in Figure 3 and Figure 4.

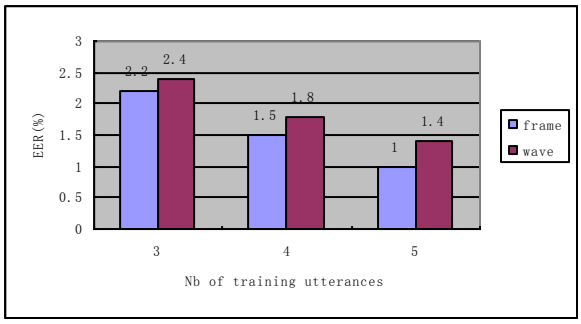

Figure 3. Comparison of Frame-based and wave-based concatenation of male

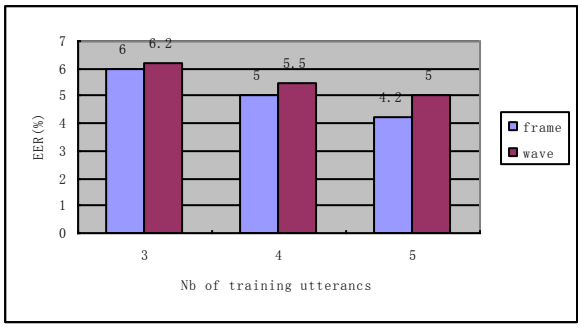

Figure 4. Comparison of Frame-based and wave-based concatenation of female

It is shown that for both female and male trials, the EER of frame-based concatenation method is lower than wave-based one.

Finally, traditional TDSR system based on DTW is compared with GMM-UBM. Based on the principle issued in session 2, GMM-UBM will outperform DTW, which is verified by our experiment below.

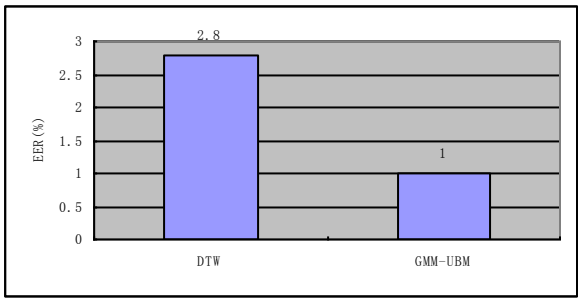

Figure 5. EER of DTW versus GMM-UBM

5. Conclusions

We have demonstrated GMM-UBM can be applied for Text-Dependent Speaker Recognition. And the proposed UBM initialization method LBG-VQ-EM algorithm obtains lower EER than conventional K-means or other clustering algorithms. Different number of training utterances comes to different EER consequences. And five training utterances is the most appropriate in GMM-UBM. At last, comparison between GMM-UBM and DTW shows GMM-UBM is quite better in Text-Dependent Speaker Recognition.

6. References

[1] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoustics, Speech, and Signal Proc., Vol. ASSP-26, pp 43-49 1978.

[2] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition [J]. Proceedings of the IEEE, 1989, 77(2):257-286.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, no. 1-3, pp.19-41,2000.

[4] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in Proc. ICASSP, 2005, pp. 629-632.

[5] D. Matrouf, N. Scheffer, B. Fauve, and J.F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in Interspeech, 2007, pp.1242-1245.

[6] Y. Linde, A. Buzo, R.M. Gray: An algorithm for vector quantization, IEEE Trans. Commun. 28, 94-95(1980).

[7] A. Dempster, N. Larid, D. Rubin: Maximum likelihood from incomplete data via the EM algorithm, J.R.Stat.Soc.39,1-38(1977).

[8] Furui S. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Processing, 1981. 29(2):254-272.

[9] Barras C and Gauvain J L. Feature and score normalization for speaker verification of cellular data, in proc of ICASSP 2003, 2:49-52.