Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Multiple Background Models for Speaker Verification

Wei-Qiang Zhang, Yuxiang Shan, Jia Liu

Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
wqzhang@tsinghua.edu.cn

Abstract

In Gaussian mixture model - universal background model (GMM-UBM) speaker verification system, UBM training is the first and the most important stage. However, few investigations have been carried out on how to select suitable training data. In this paper, a VTL-based criterion for UBM training data selection is investigated and a multiple background model (MBM) system is proposed. Experimental results on NIST SRE06 evaluation show that the presented method decreases the equal error rate (EER) of about 8% relatively when compared with the baseline.

1. Introduction

The Gaussian mixture model - universal background model (GMM-UBM) system, firstly proposed in [1], is now one of the state-of-the-art text-independent speaker verification systems. It is based on the likelihood ratio test for verification, using GMMs for likelihood computation, using a UBM for alternative speaker modeling, and using maximum a posteriori (MAP) adaptation to derive a speaker model from the UBM. UBM training is the first and the most important stage of the whole system. A high-quality UBM is supposed to represent the speaker-independent feature distribution. To achieve this goal, on the one hand, training data of different types and qualities from thousands of speakers are usually involved to reflect the alternative speech to be encountered during recognition. On the other hand, people often use gender- or channel-dependent UBMs to get better performance on some specific subpopulation of data. Above all, the data selection depends mainly on experience and experiments, which implies the quality of UBM cannot be guaranteed until experimentally tested.

In this paper, we try to resort to people's vocal tract length (VTL) to study this problem. A VTL-based criterion was firstly used to divide the whole training corpus into separate datasets, each of which was used to train a VTL-dependent UBM. Then the performance of each UBM was examined, which to some extent explained why some data were suitable for UBM training but others were not. After that a multiple background model speaker recognition system was proposed by combining the UBMs together. The multiple background model system can be viewed as a natural extension of gender-dependent UBM system. It can benefit from both subpopulation specification and system fusion.

The paper is organized as follows. Section 2 describes vocal tract length and its extraction method. Section 3 describes our experiment setup. In section 4 and section 5, the VTL-based data selection criterion and the multiple background model system are detailed. And the conclusion and future directions are given in section 6.

2. Vocal tract length clue to speaker recognition

The speaker variability extensively lies in many aspects, such as speech rate, speech volume, emotion and so on. But the major difference between the speakers is due to the difference between their average VTL [2]. The average VTL of children and females is shorter than that of males, which leads to formants of children and females move towards higher frequencies and formants of males move towards lower frequencies. In speech recognition fields, the main task is to recognize the content of the speech, so vocal tract length normalization (VTLN) is often used to obtain speaker-independent features. But in the speaker recognition situation, where we try to utilize speaker variabilities, VTL parameters are only extracted instead of normalizing the speech features.

The VTL can be measured by a warping factor α. In speech feature extraction procedure, the frequency axis can be warped by a frequency warping function in the filterbank analysis. The commonly used frequency warping function has several forms. In this paper, we use bilinear warping function, which can be expressed as [3]

$$f^{\alpha} = f + \frac{2(f_u - f_l)}{\pi} \arctan\left(\frac{(1-\alpha)\sin\theta}{1-(1-\alpha)\cos\theta}\right), \quad (1)$$

where

$$\theta = \frac{f - f_l}{f_u - f_l}\pi, \quad (2)$$

and f and f^{α} are the original and warped frequencies, respectively. In VTLN, $\alpha < 1$ corresponds to compressing the spectrum, which means the original speech is female-like; and $\alpha > 1$ corresponds to stretching the spectrum, which implies the original speech is male-like. The warping factor should has continuous values, but in practice, it is often discretized as from 0.88 to 1.12 with step-size 0.02.

The warping factor α can be estimated through maximization of likelihood of warped features $\mathcal{O}^{\alpha}$ against the warping model Λ^{*} [3]

$$\alpha^{*} = \arg\max_{\alpha} p(\mathcal{O}^{\alpha}|\Lambda^{*}). \quad (3)$$

The warping model can be obtained by iterative training the model parameters and estimating the warping factors for the training data [3].

The difference of warping factors reflects the variability between speakers, so it can be used directly for speaker recognition [4]. But in this paper, we took another approach. We employed the warping factor as a criterion to select data for UBM training.

3. Experimental setup

Our work is a mainly a process of exploration, discovery and utilization, each of which was companied with experiments, so we first introduce the experimental setup.

All the experiments were carried out on NIST SRE06 [5] corpora in core test condition (1conv4w-1conv4w) and in cross-channel conditions (1conv4w-1convmic).

The UBM training data were selected from NIST SRE04 1-side (616 utterances) and SRE03, SRE02 corpora (500 utterances).

For the frontend, speech/silence segmentation was performed by a G.723.1 VAD detector [6]. 12 MFCC coefficients plus C0 were computed using 20 ms window and 10 ms shift. Cepstral mean subtraction and feature warping [7] with a 3 s window were applied for channel mismatch compensation. Delta, acceleration and triple-delta coefficients were appended to each feature vector, which resulted in a dimensionality of 52. After that, 25% of low energy frames were discarded using a dynamic threshold. Finally, HLDA was employed to decorrelate features and reduce the dimensionality from 52 to 39 [8].

The performance measure is the same as NIST SRE, using equal error rate (EER) and minimum detection cost function (min DCF).

4. VTL-based data selection

4.1. Dataset partition

We first estimated the warping factors of all the utterances of the UBM training data. The distribution of the warping factors is illustrated in Fig. 1. From Fig. 1, we can observe that the warping factors distribute nonuniformly. The females' and males' means of warping factors are approximately 0.91 and 1.02, respectively.

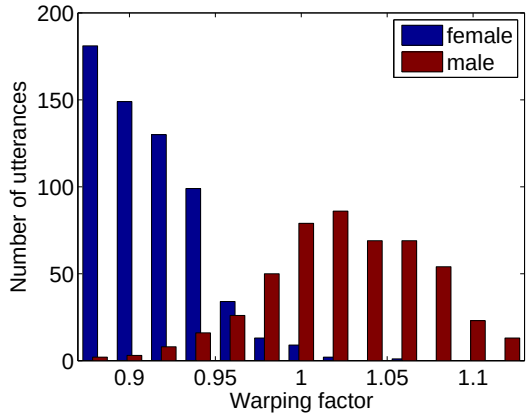


Figure 1: *The VTL distribution of the UBM training data.*

In order to reveal the relation between VTL and the UBM quality, we divided UBM training data into N disjoint datasets according to the warping factors. Considering the data size of each dataset, $N = 8$ was chosen. The detailed partition method is listed in Table 1. 8 UBMs were trained using each of the dataset and were used in next stages.

Table 1: *Dataset partition for UBM training data.*

Dataset	Warp factor	Utterances
1	0.88	183
2	0.90	152
3	0.92	138
4	0.94	115
5	0.96, 0.98	123
6	1.00, 1.02	176
7	1.04, 1.06	139
8	1.08, 1.10, 1.12	90

Table 2: *Performance of baseline gender-independent GMM-UBM system.*

Condition	EER(%)	min DCF×100
female 1conv4w-1conv4w	10.19	4.57
male 1conv4w-1conv4w	9.42	4.23
female 1conv4w-1convmic	11.84	5.69
male 1conv4w-1convmic	9.70	4.73

4.2. Experiments

4.2.1. Baseline performance

A classical GMM-UBM system as described in [1] has been built as baseline for contrastive analysis. A UBM with 1024 mixtures was trained. Speaker models were obtained by MAP adaptation of that UBM, only means were adapted. In the experiments, no channel compensation and no score normalization technologies were used.

The performance of gender-independent GMM-UBM system is listed in Table 2. The EERs for the four test conditions are about 10%, which is relatively high compared with other more powerful GMM-UBM system making use of complicated channel compensation techniques [9].

4.2.2. Performance of gender-dependent UBM

In order to reveal the effect of gender-dependent UBM to speaker recognition, we tested the gender-dependent GMM-UBM system. The results are listed in Table 3. In addition to the matched gender conditions, we also tested the cross gender conditions, i.e., we use female UBM to test male segments and use male UBM to test female segments. We can see that for the matched gender conditions, the performances are better than that of gender-independent GMM-UBM system. But for the cross gender condition, the performances significantly deteriorate, even the relative EERs are more than 100%. This shows that matched UBM training data are very important.

4.2.3. Performance of VTL-dependent UBM

We trained $N = 8$ VTL-dependent UBMs by using each of the dataset listed in Table 1, each of which was used to adapt all the target speaker models and to test all the trials of each conditions in the same way as the baseline. The performance of each UBM is listed in Table 4. We can see that for female condition, the UBM2 obtain the best results and for male condition, the UBM6 obtain the best result. Referring to Fig. 1, we can observe that the warping factors of UBM2 and UBM6 are approximately located in the means of female and male VTL distributions, re-

Table 3: *Performance of gender-dependent GMM-UBM system.*

Condition	Measure	UBM female	UBM male
female 1conv4w-1conv4w	EER(%)	9.69	19.88
	min DCF×100	4.49	7.92
male 1conv4w-1conv4w	EER(%)	20.78	8.38
	min DCF×100	8.20	3.97
female 1conv4w-1convmic	EER(%)	11.65	24.06
	min DCF×100	5.63	10.47
male 1conv4w-1convmic	EER(%)	23.19	10.01
	min DCF×100	8.89	4.42

spectively. This shows that selecting UBM training data with mean VTL for each gender is better than selecting data with other values.

Comparing the UBM2 results for female conditions and the UBM6 results for male conditions with the baseline, we can find that a UBM with far less but well-selected training data can obtain even better performance than the UBM with all the training data.

5. Multiple background models

From the results of the previous section, we can see that selecting data with mean VTL results in better performance. Since we have N UBMs, if we combine the results from all the UBMs together, even better performance may be achieved. Following this idea, we proposed a Gaussian mixture model - multiple background model (GMM-MBM) system for speaker verification. This system consists of N background models, each of which is trained using VTL-dependent data. The speaker enrollment and testing framework are described in the following subsections.

5.1. Speaker enrollment

To enroll a target speaker, all of the N UBMs are adapted using MAP (maximum a posteriori) as shown in Fig. 2. After enrollment, each target speaker is associated with N (speaker GMM, UBM) pairs, each of which is of a specific VTL warping factor.

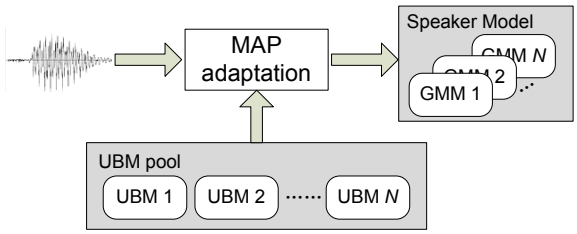


Figure 2: *Speaker enrollment of the GMM-MBM system.*

5.2. Testing framework

During verification, each testing utterance is tested against all the N (speaker GMM, UBM) pairs. Scores are fused to get final result. This procedure is illustrated in Fig. 3.

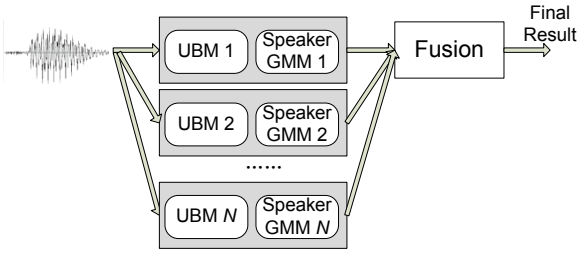


Figure 3: *Testing framework of the GMM-MBM system.*

5.3. Score fusion

For a test utterance, each (speaker GMM, UBM) pair can produce a log-likelihood-ratio score:

$$s_n = \frac{1}{T} \log \frac{p(\mathcal{O}|\mathrm{GMM}_n)}{p(\mathcal{O}|\mathrm{UBM}_n)}, \tag{4}$$

where T is the number of frames of the utterance. The following problem is how can we convert the score vector $\boldsymbol{s} = \begin{bmatrix} s_1 & s_2 & \cdots & s_N \end{bmatrix}^{\mathrm{T}}$ into the final result.

Although more delicate data-driven fusion methods can be used, such as linear fusion, bilinear fusion, GMM fusion etc., we only study the *empirical* fusion method in this paper.

5.3.1. Average method

The simplest fusion method is to average the score over all the (speaker GMM, UBM) pairs, i.e.,

$$s_{\mathrm{avg}} = \frac{1}{N} \sum_{n=1}^{N} s_n. \tag{5}$$

This method, however, without considering differences of each VTL, may not yield a good result.

5.3.2. Maximum likelihood (ML) method

The results in Table 4 remind us that selecting more *matched* UBM leads to better performance. Since more matched UBM may give higher likelihood score, we use maximum likelihood (ML) method to find the matched UBM.

For a test utterance, we first calculate the likelihood using each UBM, then select the UBM with maximum likelihood.

$$n^* = \arg\max_n p(\mathcal{O}|\mathrm{UBM}_n), \tag{6}$$

Table 4: *Performance of each GMM-MBM system.*

Condition	Measure	UBM1	UBM2	UBM3	UBM4	UBM5	UBM6	UBM7	UBM8
female 1conv4w-1conv4w	EER(%)	10.80	**9.81**	10.49	12.12	16.86	20.82	22.37	23.77
	min DCF×100	5.00	**4.37**	5.06	5.53	6.66	7.81	8.41	8.80
male 1conv4w-1conv4w	EER(%)	23.09	20.95	18.96	16.91	11.34	**9.02**	10.06	11.98
	min DCF×100	8.13	7.77	7.42	7.36	5.76	**4.25**	4.81	5.67
female 1conv4w-1convmic	EER(%)	13.01	**11.13**	11.91	13.53	18.65	25.12	26.07	26.16
	min DCF×100	5.77	**5.32**	5.63	6.33	7.70	8.72	8.77	9.05
male 1conv4w-1convmic	EER(%)	25.16	23.67	21.90	20.05	12.94	**9.91**	11.63	13.96
	min DCF×100	8.25	7.91	7.65	7.54	6.45	**4.72**	5.60	6.99

Table 5: *Performance of average fusion method.*

Condition	EER(%)	min DCF×100
female 1conv4w-1conv4w	13.92	5.98
male 1conv4w-1conv4w	12.50	5.48
female 1conv4w-1convmic	15.62	6.33
male 1conv4w-1convmic	14.08	6.37

Table 6: *Performance of ML fusion method.*

Condition	EER(%)	min DCF×100
female 1conv4w-1conv4w	9.77	4.28
male 1conv4w-1conv4w	8.46	3.88
female 1conv4w-1convmic	11.79	5.62
male 1conv4w-1convmic	9.43	4.21

At last, we use the score of corresponding (speaker GMM, UBM) pair to calculate the likelihood rate.

$$s_{\mathrm{ML}} = s_{n^*}. \tag{7}$$

5.3.3. Minimum likelihood ratio (MLR) method

In this method, we want to use the likelihood ratio produced by a (speaker GMM, UBM) pair directly instead of selecting the model via ML method. But intuitively, the speaker GMM likelihood and the UBM likelihood will both increase if a matched test utterance is encountered. How about the likelihood ratio? In order to find the behind law, we calculated the means and standard deviations of likelihood ratios of SRE06 with each (speaker GMM, UBM) pair. The results are plotted in Fig. 4. Compare with Table 4, we can find that the less the likelihood ratio is, the better the performance gets. These results imply that for a matched utterance and model, the speaker GMM and UBM give higher likelihood, but the increment of speaker GMM is less than that of UBM. Other test corpora also give similar results. The underlying reason is just under investigation.

Based on this phenomenon, we can straightforwardly select the minimum likelihood ratio as the last score.

$$s_{\mathrm{MLR}} = \min_n s_n. \tag{8}$$

5.4. Experiments

In this section, we tested the three fusion method. The results of average fusion method are listed in Table 5. We can see that the performance of average method is not as good as, but even worse than that of GMM-UBM baseline system. The reason may be it is unfair to weight each UBM equally.

The results of ML fusion method are listed in Table 6. Compared with Table 4, we can see that it outperforms UBM2 and UBM6.

The results of MLR fusion method are listed in Table 7. This method gives best performance among the three, which shows that MLR method can select more matched models and thus leads better result.

Table 7: *Performance of MLR fusion method.*

Condition	EER(%)	min DCF×100
female 1conv4w-1conv4w	9.40	4.14
male 1conv4w-1conv4w	8.36	3.71
female 1conv4w-1convmic	10.76	5.43
male 1conv4w-1convmic	9.38	4.08

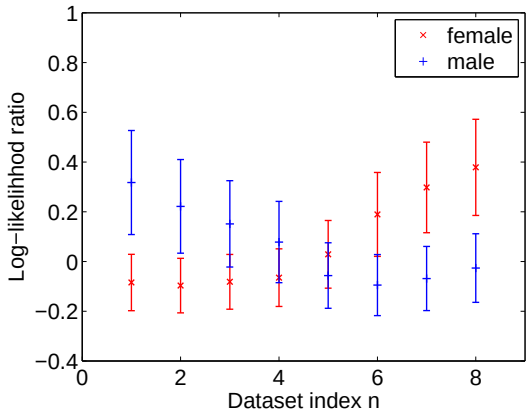
6. Conclusions

In this paper, we first investigated the the VTL-based criterion for UBM training data selection. Experiments showed that the UBM trained with selected mean-VTL data was better than the UBM trained with all the data. Based on this finding, we further proposed a multiple background model system, i.e., using multiple speaker GMM and UBM pairs, for speaker recognition. Through minimum likelihood ratio fusion, the proposed method can improve the performance evidently.

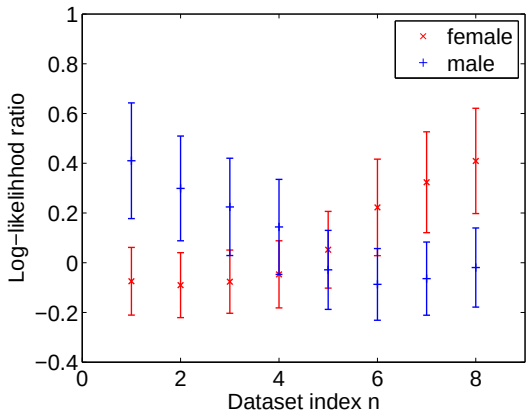
Further works will focus on investigating whether the techniques improve the state-of-the-art systems and developing efficient method to lower the computational cost.

7. Acknowledgements

This work was supported by the National Natural Science Foundation of China and Microsoft Research Asia under Grant No. 60776800, by the National Natural Science Foundation of China under Grant No. 90920302, and in part by the National High Technology Development Program of China (863 Program) under Grant No. 2006AA010101, No. 2007AA04Z223, No. 2008AA02Z414 and No. 2008AA040201.



(a) 1conv4w-1conv4w



(b) 1conv4w-1convmic

Figure 4: *The log-likelihood ratio distribution for each UBM.*

8. References

[1] 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, Jan. 2000.

[2] T. Claes, I. Dologlou, L. ten Bosch *et al.*, "A novel feature transformation for vocal tract length normalization in automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 6, pp. 549 – 557, Nov. 1998.

[3] P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," Carnegie Mellon University, Tech. Rep. CMU-CS-97-148, May 1997.

[4] S. Grashey and C. Geissler, "Using a vocal tract length related parameter for speaker recognition," in *Proc. IEEE Odyssey*, San Juan, Puerto Rico, June 2006.

[5] "2006 NIST speaker recognition evaluation," Available: http://www.itl.nist.gov/iad/mig/tests/spk/2006/index.html, 2006.

[6] ITU-T, "G.723.1 Annex A: Silence compression scheme," Nov. 1996.

[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. IEEE Odyssey*, Crete, Grece, June 2001, pp. 213–218.

[8] N. Brummer, L. Burget, J. H. Cernocky *et al.*, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, Sept. 2007.

[9] H. Li, B. Ma, and K. A. Lee, "NIST SRE 2008 IIR and I4U submissons site presentation," in *2008 NIST Speaker Recognition Evaluation Workshop*, Montreal, June 2008.