Tsinghua University
Modern Audio & Speech Technology

Multiple Background Models for Speaker Verification

W.-Q. Zhang, Y. Shan, J. Liu

wqzhang@tsinghua.edu.cn

Contents

Introduction

- The Gaussian mixture model - universal background model (GMM-UBM) is a typical speaker verification system.
- A high-quality UBM is supposed to represent the speaker-independent feature distribution.
- Two methods to guarantee the quality of UBM:
  – training on misc data
  – gender- or channel-dependent UBMs
- Maybe there are other approaches …

Vocal tract length (VTL)

- The speaker variability extensively lies in many aspects, such as speech rate, speech volume, emotion, vocal effect and so on.
- But the major difference between the speakers is due to the difference between their average VTL.
- So, in speech recognition, vocal tract length normalization (VTLN) is often used to obtain speaker-independent features.

VTLN

- A usually used frequency warping function:

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

$$\theta = \frac{f - f_l}{f_u - f_l}\pi$$

  - f original frequency
  - f^{α} warped frequency

Warping factor α

- The warping factor can be estimated by:

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

  - O^{α} : warped features
  - Λ^* : warping model
  - 0.88 to 1.12 with step-size 0.02

Experimental setup – Data

- All the experiments were carried out on NIST SRE06 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).

Experimental setup – Feature

- 12 MFCC + C0
- Cepstral mean subtraction (CMS)
- Feature warping
- Delta, acceleration and triple-delta
- HLDA 52 -> 39

VTL distribution

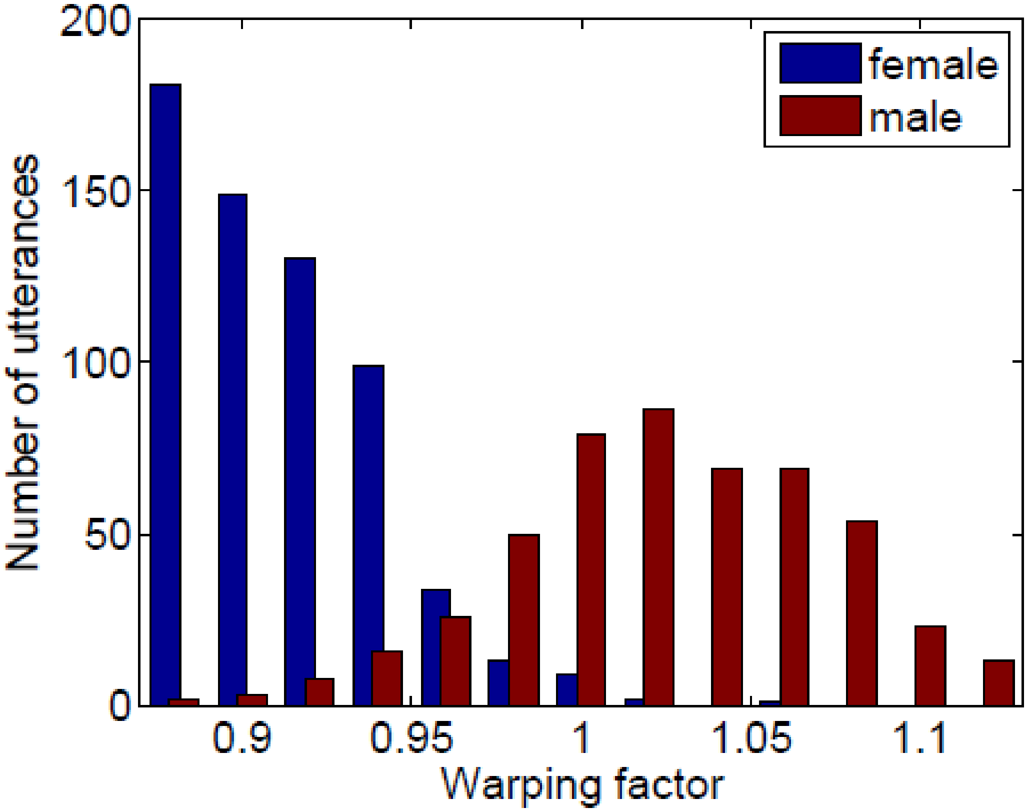


Figure 1: *The VTL distribution of the UBM training data.*

Dataset partition

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

- We divided UBM training data into *N=8* disjoint datasets according to the warping factor.

Baseline performance

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

- The EERs for the four test conditions are about 10%.

Gender-dependent UBM

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

- matched gender condition: slightly improved.
- But the cross gender condition: very bad.

VTL-dependent UBM

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

- For female condition, UBM2 gives the best results.
- For male condition, UBM6 gives the best result.

Comments

- Comparing the UBM2 results for female conditions and the UBM6results 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.

Multiple background models

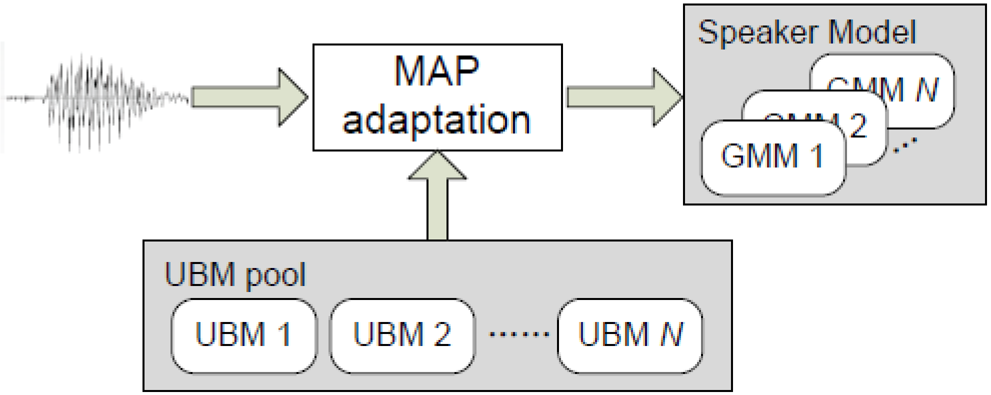


Figure 2: *Speaker enrollment of the GMM-MBM system.*

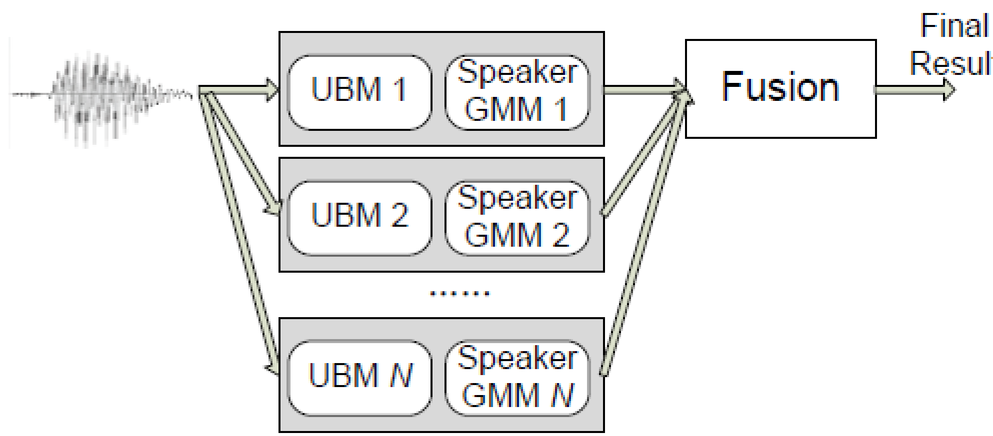


Figure 3: *Testing framework of the GMM-MBM system.*

Score fusion

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

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

- MBM can obtain a score vector:

$$\begin{bmatrix} s_1 & s_2 & \cdots & s_N \end{bmatrix}^{\mathrm{T}}$$

- We can use score fusion method to obtain the final result.

Average method

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

$$s_{\text{avg}} = \frac{1}{N}\sum_{n=1}^{N} s_n$$

- Not good.

Maximum likelihood (ML) method

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

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

$$s_{\mathrm{ML}} = s_{n^*}$$

- Just so so.

Minimum likelihood ratio (MLR) method

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

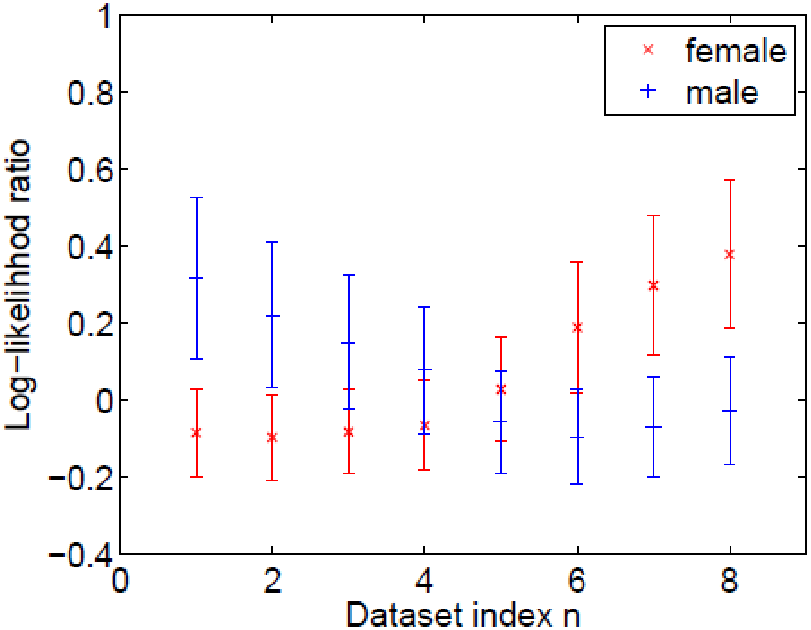
$$s_{\mathrm{MLR}} = \min_n s_n$$

- It gives best result among three methods.

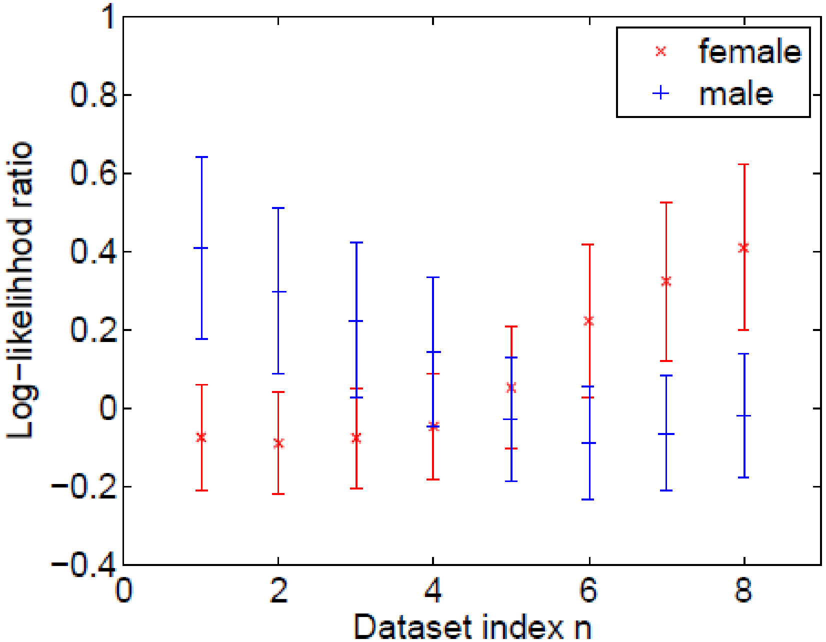
Possible reason

- Why the minimum likelihood ratio method gives best result?
- We haven't known the exact reason.
- Intuitively, the speaker GMM likelihood and the UBM likelihood will both increase if a matched test utterance is encountered.
- We calculated the means and standard deviations of likelihood ratios of SRE06 with each (speaker GMM, UBM) pair.

The llr distribution for each UBM



(a) 1conv4w-1conv4w



(b) 1conv4w-1convmic

- The less the log likelihood ratio (llr) is, the better the performance gets.

Conclusions

- In this paper, we first investigated 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.

Open question

- Why the minimum likelihood ratio method gives best result? Is it just a coincidence?
- Whether the techniques improve the state-of-the-art systems?
- How to lower the computational cost?

Q&A