Sequential Adaptive Learning for Speaker Verification

Jun Wang
CSLT，RIIT，THU
2013-03-01

Outline

1. Introduction
2. Sequential Adaptive Learning
3. Experiments
4. Conclusion
5. Reference

Introduction

- GMM-UBM-based speaker verification heavily relies on a well trained UBM.
- In practice, it is often difficult to collect sufficient channel-matched data to train a fully consistent UBM.
- A multitude of research has been proposed to address channel mismatch or session variation.

Introduction

- Within the GMM-UBM framework,
  - feature transform [3, 4, 5];
  - model compensation [6, 7];
  - score normalization [1, 8];
  - factor analysis [9,10] and it's simple algorithm implementation[11];
  - various feature and model compensation approaches[12];
- Besides GMM-UBM,
  - [13] proposed to reduce channel impact in neural network;

Introduction

- we propose a sequential adaptive learning approach to the channel mismatch problem.
- By this approach, the UBM and speaker models are updated sequentially and gradually, finally converging to the new or dynamic channel with a large amount of enrollments.

Sequential Adaptive Learning

➢Review of MAP estimation

The objective function:

$$\begin{aligned} \mathcal{L}(\mu, \sigma) &= logP(\mu, \sigma | X) \\ &\propto \sum_i log\{\mathcal{N}(x_i; \mu, \sigma)P(\mu, \sigma)\}. \end{aligned}$$

Maximizing this objective leads to the following MAP estimation:

$$\mu = \frac{\sum_i x_i + \frac{\sigma}{\hat{\sigma}}\hat{\mu}}{N + \frac{\sigma}{\hat{\sigma}}} \quad (1)$$

Sequential Adaptive Learning

➢Review of MAP estimation

When extending to GMM:

$$r_i(c) = \frac{\mathcal{N}(x_i; \mu_c, \sigma_c)}{\sum_m \mathcal{N}(x_i; \mu_m, \sigma_m)}. \quad (2)$$

Define the following sufficient statistics:

$$r_c = \sum_i r_i(c) \quad (3)$$

$$z_c = \sum_i r_i(c) x_i, \quad (4)$$

the MAP estimation is given by:

$$\mu_c = \frac{z_c + \frac{\sigma}{\hat{\sigma}}\hat{\mu}}{r_c + \frac{\sigma}{\hat{\sigma}}} \quad (5)$$

Sequential Adaptive Learning

➢Sequential UBM adaptation

Motivation of sequential UBM adaptation:

Use the new enrollment speech data to update the UBM. We start from a 'pool and re-estimation' procedure.

$$\mu_c = \frac{z_c + \hat{z}_c}{r_c + \hat{r}_c} \quad (6)$$

$$= \frac{z_c + \hat{r}_c \hat{\mu}_c}{r_c + \hat{r}_c} \quad (7)$$

$$\hat{r}_c = \frac{\sigma}{\hat{\sigma}}. \quad (8)$$

Sequential Adaptive Learning

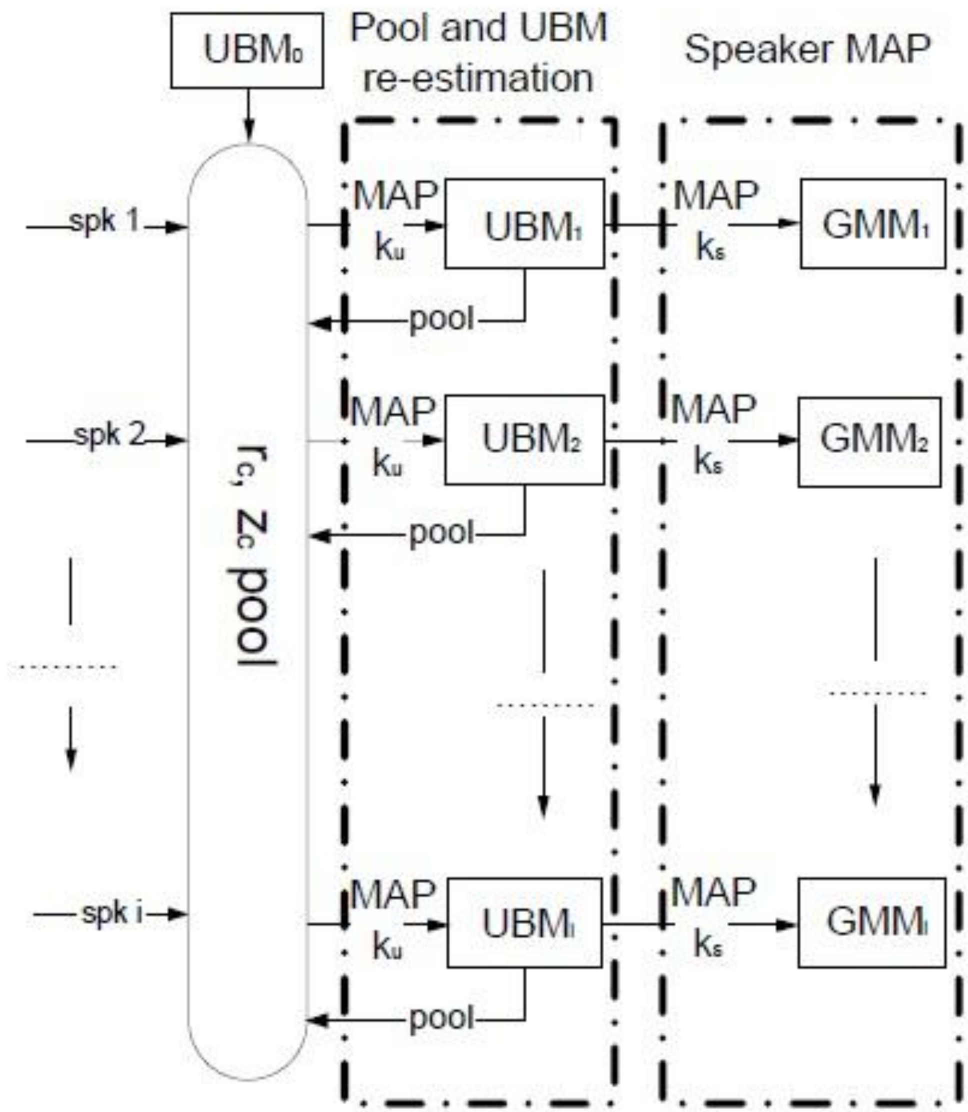

Fig. 1. Sequential UBM MAP adaptation.

Sequential Adaptive Learning

➢ Sequential Speaker Model adaptation

Firstly, we need to save sufficient statistics for each speaker which are defined in equation (3) and (4).

When a new enrollment occurs, sequential UBM adaptation is used to train a new UBM, then we use the new UBM to update each speaker model according to it's sufficient statistics.

$$\mu = \frac{z_c + \frac{\sigma}{\hat{\sigma}}\hat{\mu}_n}{r_c + \frac{\sigma}{\hat{\sigma}}} \qquad (9)$$

Sequential Adaptive Learning

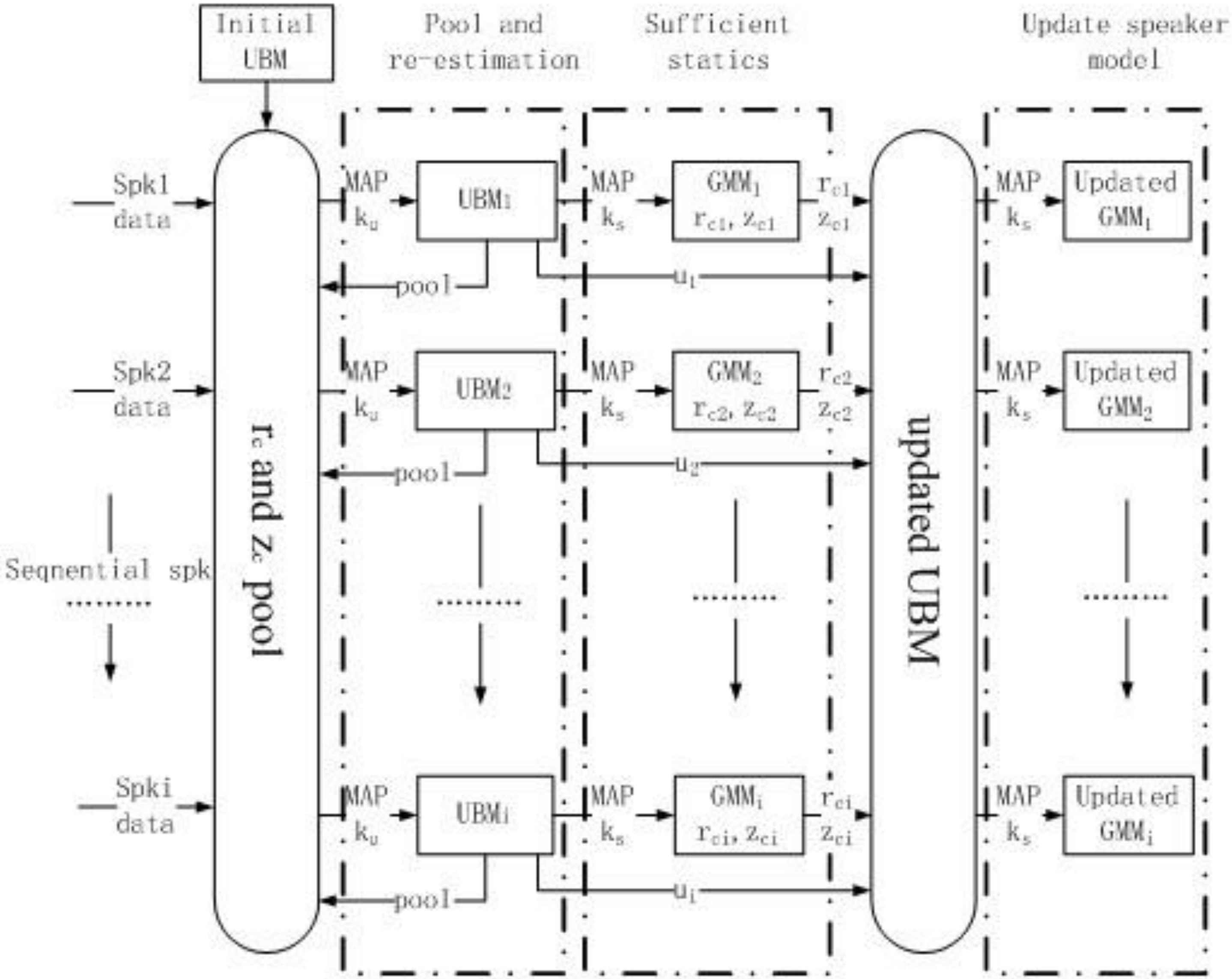

Fig2. Sequential adaptive learning

Experiments

- We conduct the experiments on a time-varying database [14].
- We start the experiments with two initial UBMs.
- The verification performance is evaluated in terms of equal error rates (EER).

Experiments

- Sequential UBM adaptation experiment.

	EER%		
	k_s=0.5	k_s=1	k_s=2
UBM (baseline)	11.75	11.92	11.75
SUBM(k_u =90)	12.69	12.42	12.05
SUBM(k_u=180)	11.67	11.20	11.47
SUBM(k_u=270)	11.19	11.05	11.07
SUBM(k_u=360)	**11.12**	**11.02**	**11.05**

Table 1. Results with UBM_a as the initial.

Experiments

- Sequential UBM adaptation experiment.

	EER%		
	k_s=0.5	k_s=1	k_s=2
UBM (baseline)	10.04	10.22	10.44
SUBM(k_u =90)	9.36	9.20	8.83
SUBM(k_u=180)	8.77	8.88	8.82
SUBM(k_u=270)	**8.72**	8.84	8.79
SUBM(k_u=360)	8.73	**8.82**	**8.79**

Table 2. Results with UBM_b as the initial.

$SUBM(k_u$

Experiments

➢Sequential Adaptive Learning experiment.

	System EER		
	k_s=0.5	k_s=1	k_s=2
UBM(baseline)	11.75%	11.92%	11.75%
$SUBM(k_u$=90)	9.34%	8.90%	8.65%
$SUBM(k_u$=180)	9.37%	9.04%	8.95%
$SUBM(k_u$=270)	9.47%	9.22%	9.08%
$SUBM(k_u$=360)	9.52%	9.35%	9.24%
$SUBM(k_u$=540)	9.74%	9.54%	9.54%

Table 3. Sequential adaptive learning with UBMa

Experiments

- Sequential Adaptive Learning experiment.

	System EER		
	k_s=0.5	k_s=1	k_s=2
UBM(baseline)	10.04%	10.22%	10.44%
$SUBM$ (k_u=90)	6.78%	6.75%	6.57%
$SUBM$ (k_u=180)	**6.78%**	**6.64%**	**6.48%**
$SUBM$ (k_u=270)	6.92%	6.81%	6.67%
$SUBM$ (k_u=360)	7.10%	7.02%	6.93%
$SUBM$ (k_u=540)	7.43%	7.35%	7.29%

Table 4. Sequential adaptive learning with UBMb

Experiments

- Quality of sequential UBM.

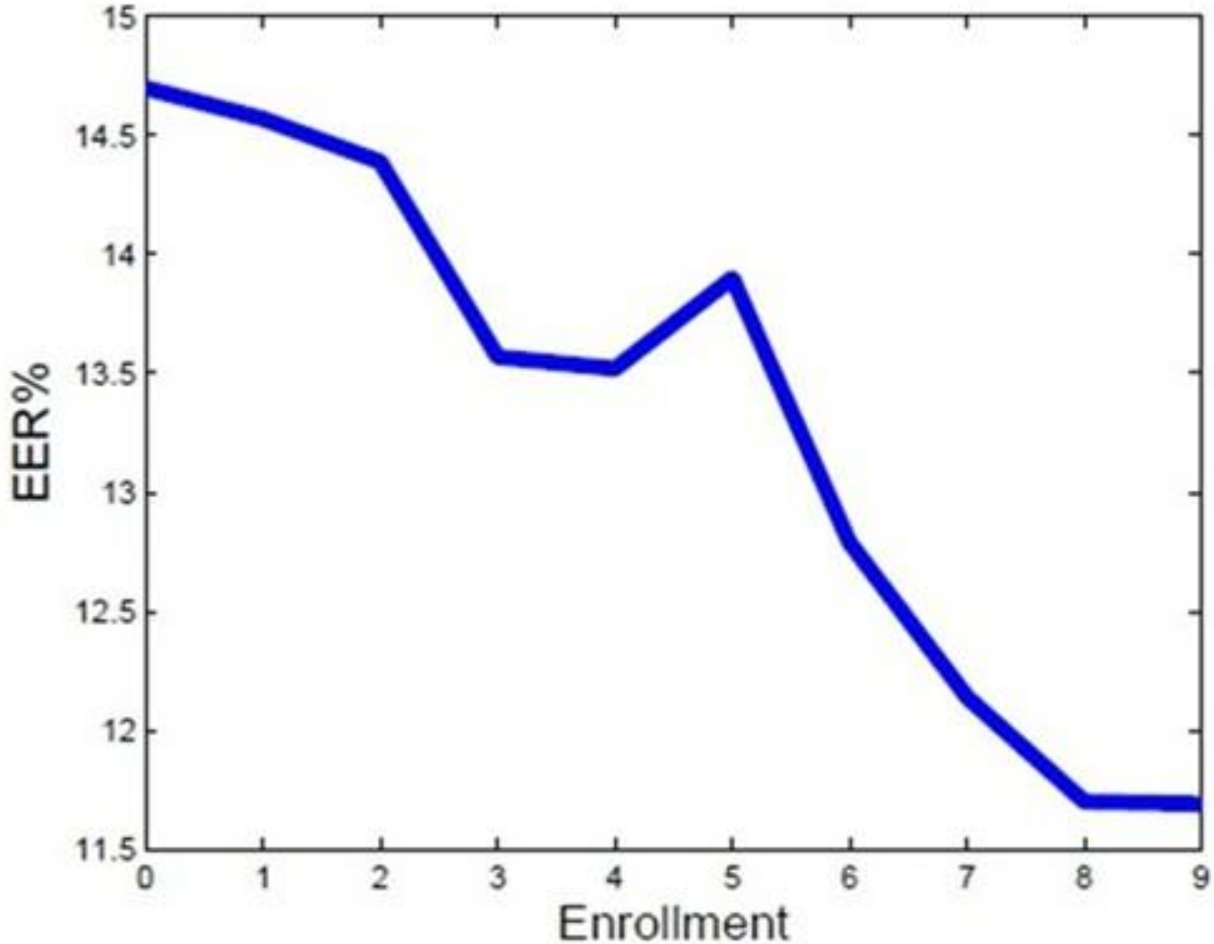

Fig3. Quality of sequentially adapted UBM

Conclusion

- By adapting an initial UBM with a strong prior whenever a new enrollment is available, the UBM learns and converges to the working channel gradually, leading to improved verification performance.
- Use the new UBM to update each speaker model according to it's sufficient statics, leading to improved verification performance.
- In our experiments, this sequential approach provides relative EER reduction of 24.1% and 34.9% for two mismatched UBMs, respectively.

Reference

[1] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, “Speaker verification using adapted gaussian mixture models,” Digital Signal Processing, vol. 10, pp. 19–41, 2000.

[2] J. L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,” IEEE Trans. Speech Audio Process, vol. 2, pp. 291–298, 1994.

[3] D.A. Reynolds, “Channel robust speaker verification via feature mapping,” in ICASSP 2003, 2003, vol. 2, pp. 53–56.

[4] Donglai Zhu, Bin Ma, Haizhou Li, and Qiang Huo, “Handset-dependent background models for robust textindependent speaker recognition,” in ICASSP 2007, 2007, vol. 4, pp. 61–64.

[5] C. Vair, D. Colibro, and P. Laface, “Channel factors compensation in model and feature domain for speaker recognition,” in Odyssey’06, the Speaker Recognition Workshop, 2006.

[6] L. Heck and M. Weintraub, “Handset-dependent background models for robust text-independent speaker recognition,” in ICASSP 1997, 1997, vol. 2, pp. 1071– 1074.

[7] R. Teunen, B. Shahshahani, and L. Heck, “A modelbased transformational approach to robust speaker recognition,” in ICSLP2000, 2000.

Reference

[8] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” Digital Signal Processing, vol. 10, pp. 42–54, 2000.

[9] S.P. Kishore and B. Yegnanarayana, “Speaker verification: minimizing the channel effects using autoassociative neural network models,” in ICASSP2000, 2000, vol. 2, pp. 1101–1104.

[10] A. Solomonoff, C. Quillen, and W. M. Campbell, “Channel compensation for svm speaker recognition,” in Proc Odyssey, Speaker Language Recognition Workshop, 2004, pp. 57–62.

[11] P. Kenny and P. Dumouchel, “Disentangling speaker and channel effects in speaker verification,” in ICASSP2004, 2004, vol. 1, pp. 37–40.

[12] Driss Matrouf, Nicolas Scheffer, Benoit Fauve, and Jean-Franois Bonastre, “A straightforward and efficient implementation of the factor analysis model for speaker verification,” in Interspeech 2007, 2007.

[13] L. Burget, P. Matejka, O. Glembek, and P. Schwarz, “Analysis of feature extraction and channel compensation in gmm speaker recognition system,” IEEE Trans. on Audio, Speech and Language processing, vol. 15, no. 7, pp. 1979–1986, 2007.

[14] Linlin Wang and Thomas Fang Zheng, “Creation of time-varying voiceprint database,” in Oriental- COCOSDA, 2010.

Thanks

Q&A.