A Comparison of Various Adaptation Methods for Speaker Verification with Limited Enrollment Data

Man-Wai Mak,[1] Roger Hsiao,[2] and Brian Mak[3]

[1]The Hong Kong Polytechnic University
[2]Carnegie Mellon University
[3]The Hong Kong University of Science and Technology

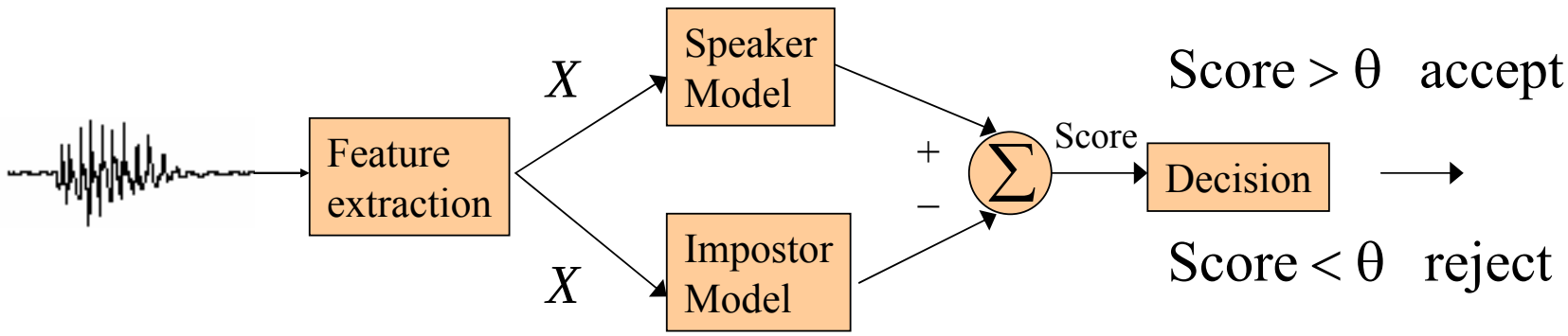
ICASSP-2006

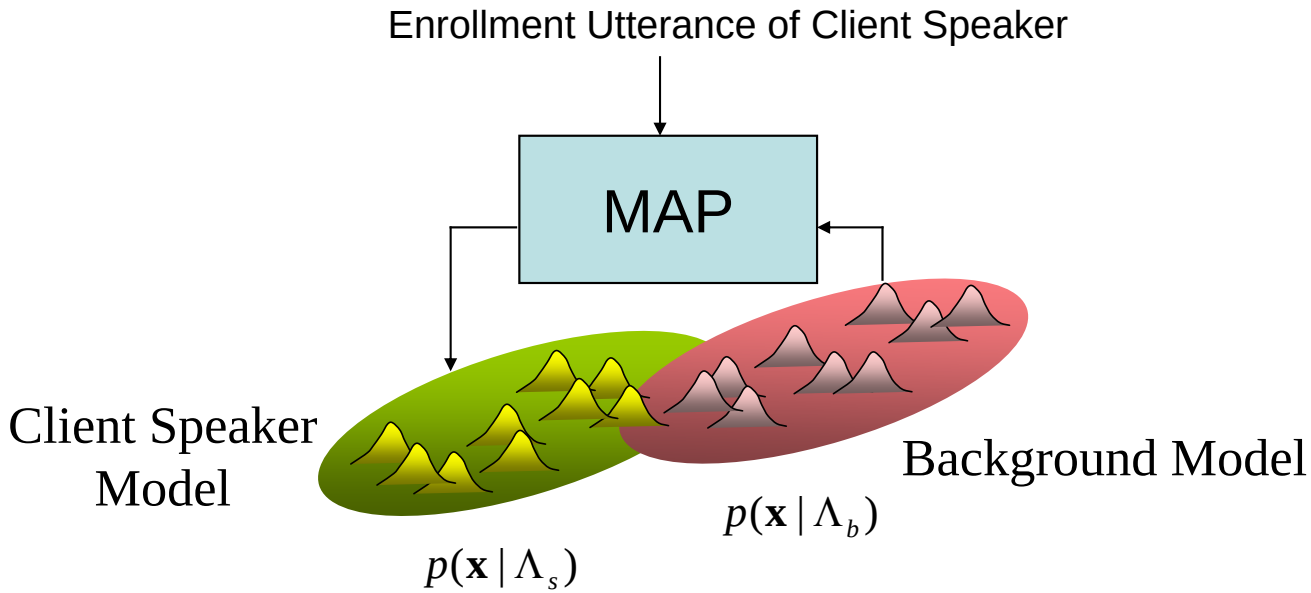
Summary

- To gain user acceptance of speaker verification technologies, adaptation algorithms that can enroll speakers with short utterances are highly essential.
- This paper compares four state-of-the-art model adaptation techniques for speaker model creation:
  - Maximum a Posteriori (MAP)
  - Maximum-Likelihood Linear Regression (MLLR)
  - Reference Speaker Weighting (RSW)
  - Kernel-Eigenspace-Based MLLR (KEMLLR)
- Evaluations based on NIST2001 SRE show that

Adaptation Method	**Best For**
MAP	Long enrollment utterances (32 seconds)
MLLR	Medium-length utterances (8–16 seconds)
KEMLLR	Short enrollment utterances (2–4 seconds)

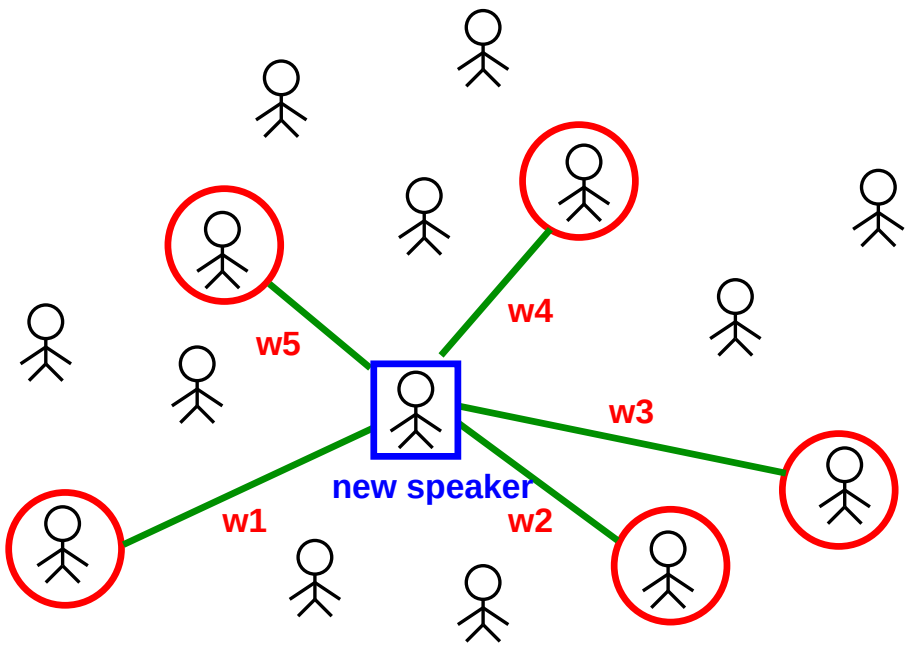
Background





$$\text{Score} = \sum_{\mathbf{x} \in \text{Utterance}} \left[\log p(\mathbf{x} \mid \Lambda_s) - \log p(\mathbf{x} \mid \Lambda_b) \right]$$

Reference Speaker Weighting (RSW)



- In RSW, the new speaker is approximated as:

$$\mathbf{s} \simeq \mathbf{s}^{(rsw)}(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{y}_m^{(rsw)} = \mathbf{Y}^{(rsw)}\mathbf{w}$$

where $\mathbf{w}$ is the combination weight vector and $\mathbf{y}_m$ is the mth reference speaker.

Reference Speaker Weighting (RSW)

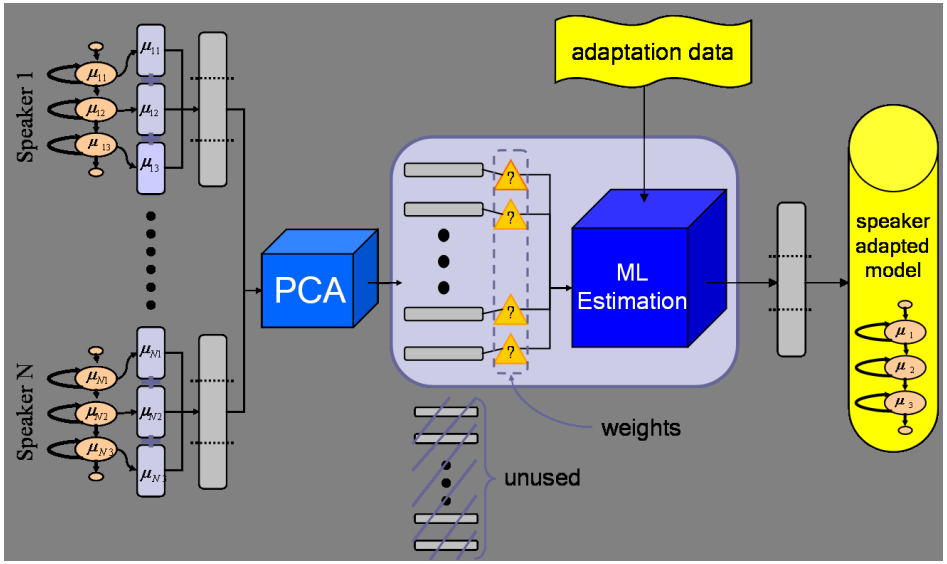
- For the mean vector of the *r*th Gaussian,

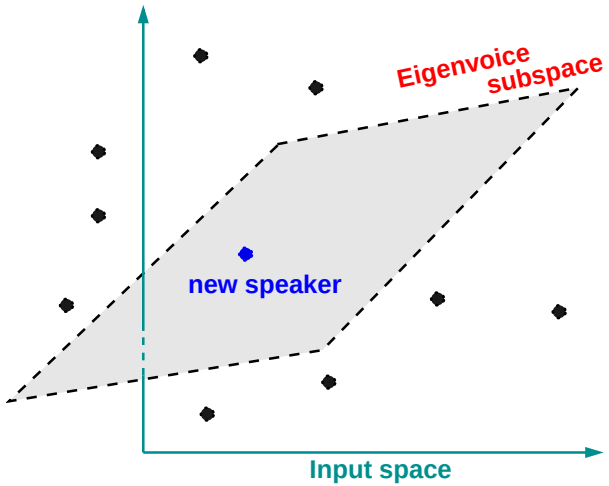
$$\mathbf{s}_r^{(rsw)} = \sum_{m=1}^{M} w_m \mathbf{y}_{mr}^{(rsw)} = \mathbf{Y}_r^{(rsw)} \mathbf{w}$$

- Given adaptation data $\mathbf{O} = \{\mathbf{o}_t, t = 1, ..., T\}$, our aim is to maximize the following Q($\mathbf{w}$) function:

$$Q(\mathbf{w}) = -\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t(r) (\mathbf{o}_t - \mathbf{s}_r^{(rsw)}(\mathbf{w}))' \mathbf{C}_r^{-1} (\mathbf{o}_t - \mathbf{s}_r^{(rsw)}(\mathbf{w}))$$

Eigenvoice Speaker Adaptation (EV)

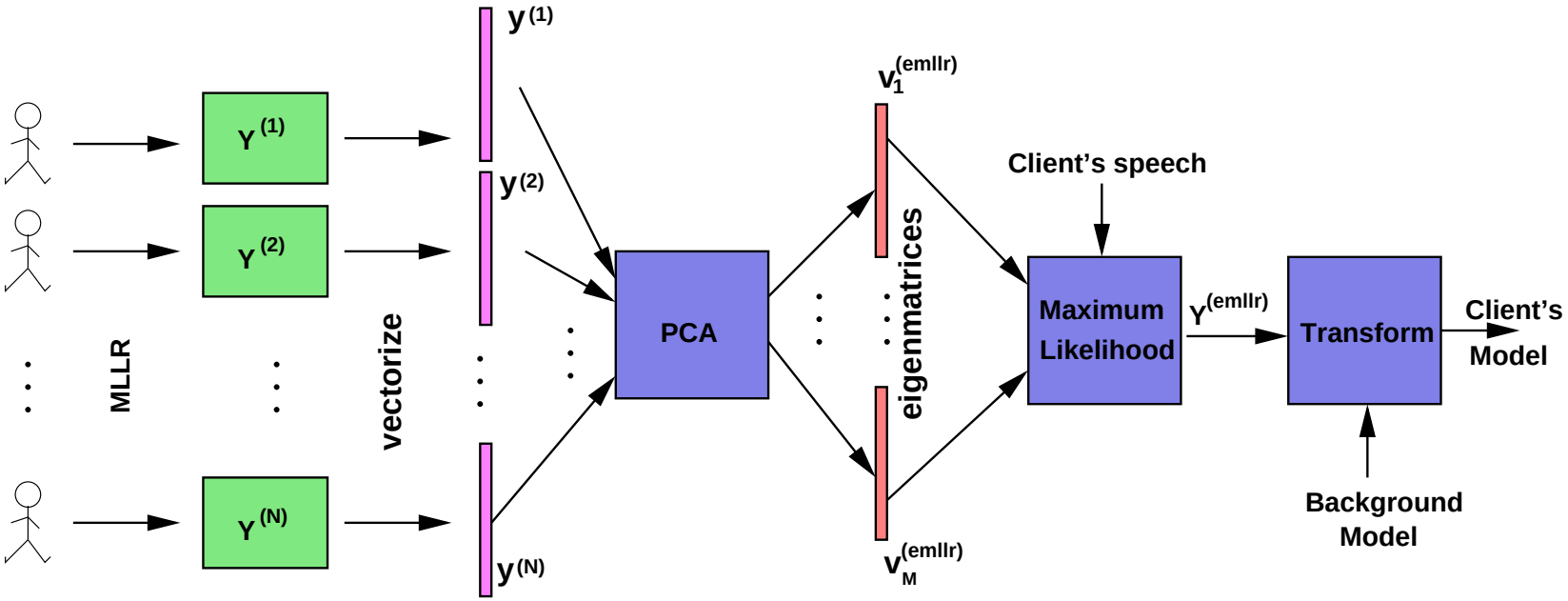




The new speaker model in the input speaker supervector space is:

$$\mathbf{s} \simeq \mathbf{s}^{(ev)} = \bar{\mathbf{x}} + \sum_{m=1}^{M} w_m \mathbf{v}_m^{(ev)}$$

Eigenspace-based MLLR Adaptation



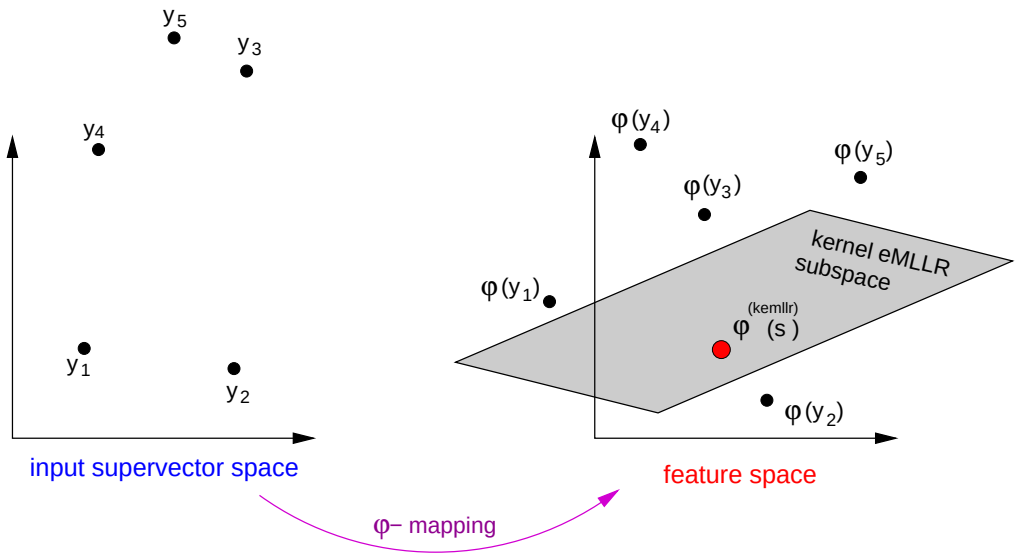
- $\mathbf{Y}^{(i)} \in \mathbb{R}^{(d+1)\times d}$ is the MLLR transformation of the i-th speaker
- The new speaker's vectorized MLLR transformation $vec(\mathbf{Y}^{(emllr)})$ is a weighted sum of the first M eigenmatrices:

$$vec(\mathbf{Y}^{(emllr)}) = \sum_{m=1}^{M} w_m \mathbf{v}_m^{(emllr)}$$

Kernel Eigenspace-based MLLR Speaker Adaptation (KEMLLR)

Step 1: Perform Kernel PCA

- Map data **x** (here, speaker MLLR transformation supervector) in an input space to a high-dimensional kernel-induced feature space.
- Perform PCA in the feature space to determine the eigenmatrices $\mathbf{v}_m^{(kev)}$ using the the kernel matrix **K**: $K_{ij} = k(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}) \equiv \varphi(\mathbf{y}^{(i)})'\varphi(\mathbf{y}^{(j)})$.



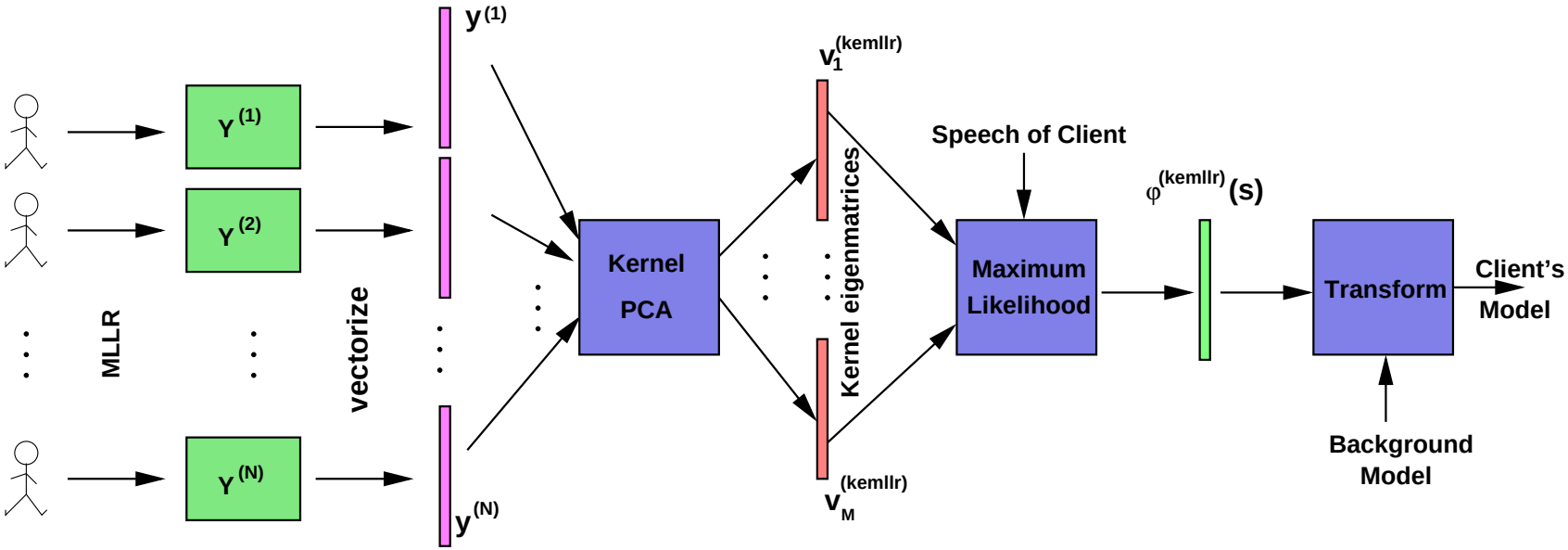
Kernel Eigenspace-based MLLR Speaker Adaptation

Step 2: Determine the eigenmatrix weights w_m

Step 3: Determine the supervector of client speaker **s**

$$\varphi^{(kemllr)}(\mathbf{s}) = \sum_{m=1}^{M} w_m \mathbf{v}_m^{(kemllr)}$$

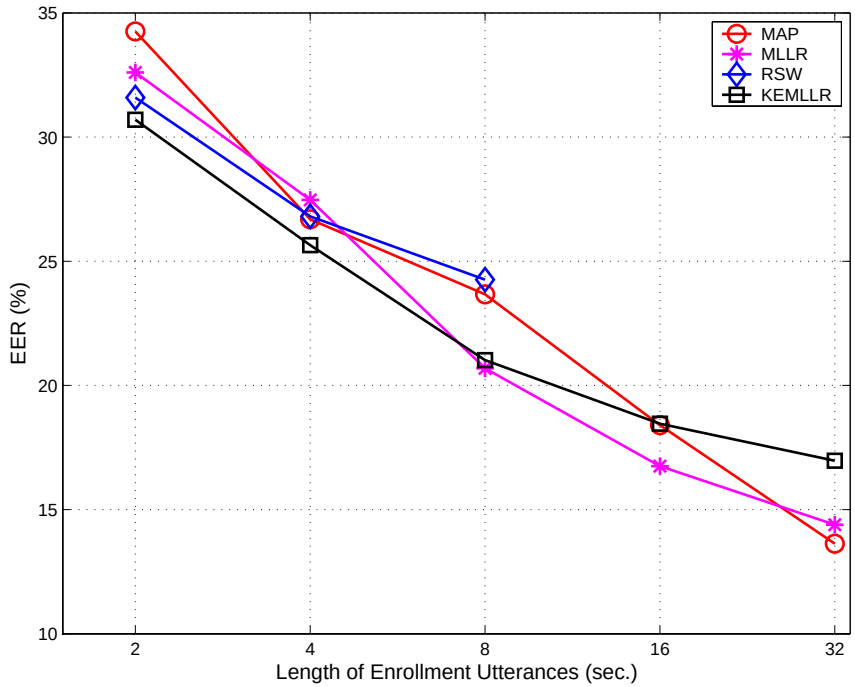
Step 4: Perform MLLR transformation using $\varphi^{(kemllr)}(\mathbf{s})$



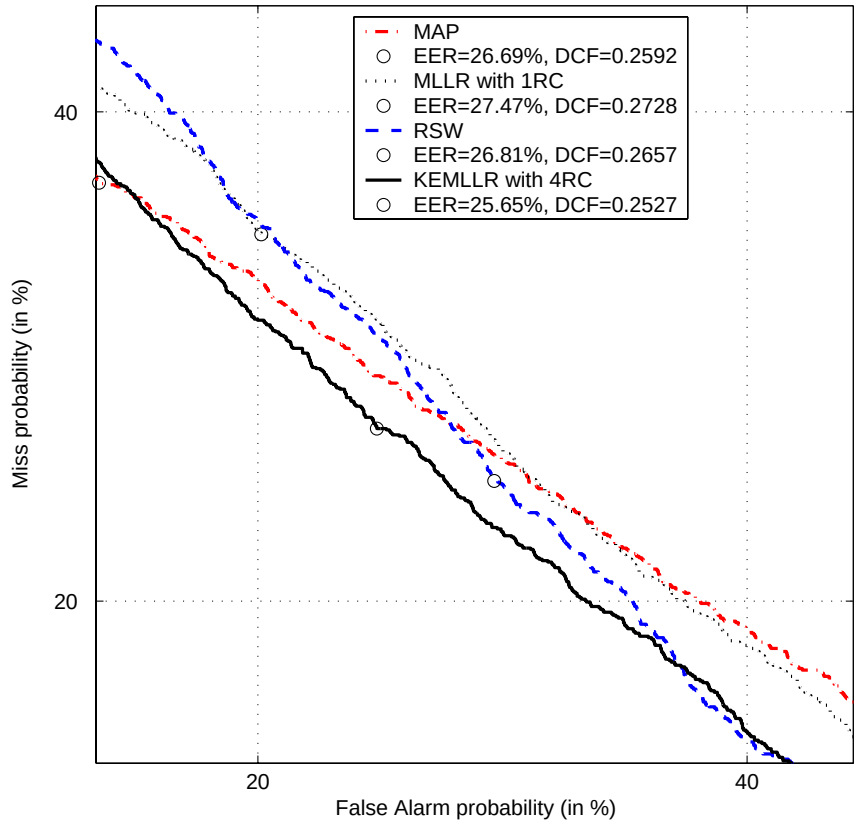
Evaluations on 2001 NIST SRE

- Cellular phone speech of 74 male 100 female target speakers
- 2,038 target trials and 20,380 impostor attempts
- MFCC + ΔMFCC as features
- 1,024-component universal background model (UBM) was trained using the training utterances of all 60 speakers in the development set of NIST01.
- For each target speaker, four 1,024-component speaker-dependent GMMs were created by adapting the UBM using MAP adaptation, MLLR transformation, RSW, and KEMLLR.
- Enrollment utterances of 2s, 4s, 8s, 16s, and 32s were used.

Results: EER and DET



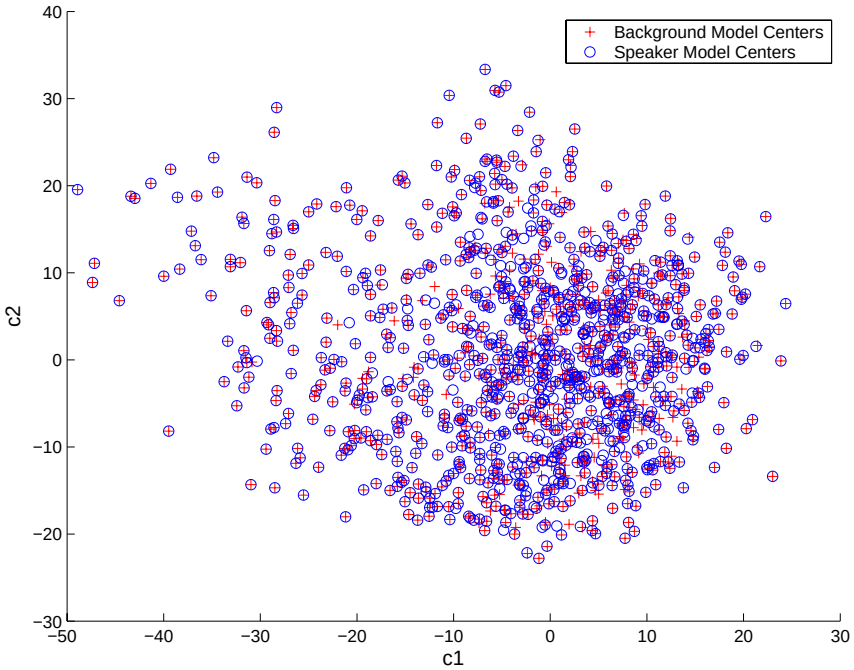
EER versus enrollment utterance length



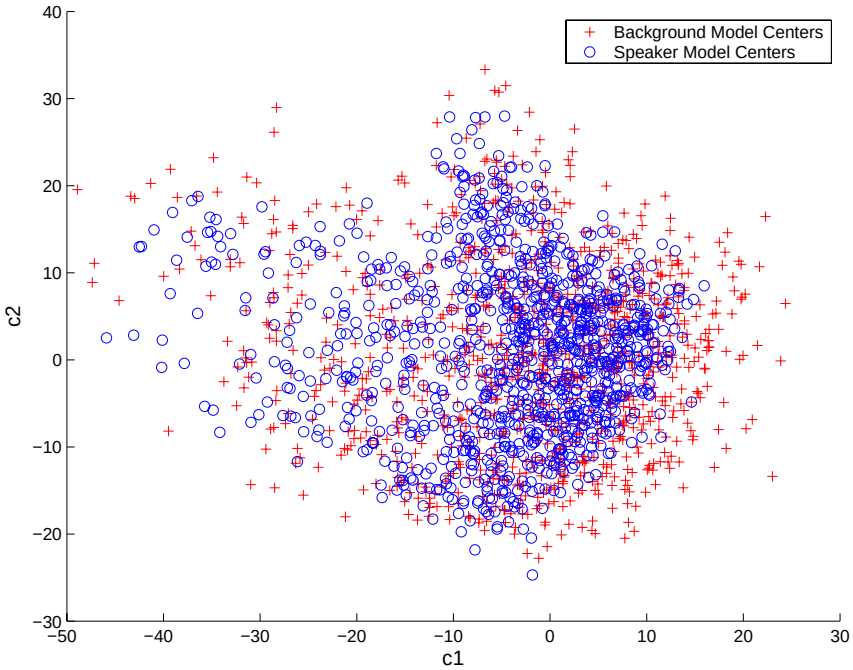
DET based on 4-second enrollment utterances

Results: Positions of Adapted Centers

4-second enrollment utterances



MAP Adaptation



KEMLLR Adaptation

Conclusions

This paper has compared the performance (in terms of EERs, DET, and minimum DCF) of MAP, MLLR, RSW, and KEMLLR for speaker enrollment under short-utterance scenarios. It was found that (1) KEMLLR outperforms other adaptation methods when the amount of enrollment data is very limited, (2) MLLR performs better for medium-length utterances, and (3) when a large amount of enrollment data is available, MAP is a better candidate for creating speaker models.

References

- D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
- Tim J. Hazen, "A comparison of novel techniques for rapid speaker adaptation," *Speech Communications*, vol. 31, pp. 15–33, May 2000.
- R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 695–707, Nov 2000.
- B. Mak, J. T. Kwok, and S. Ho, "Kernel eigenvoice speaker adaptation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 984–992, September 2005.
- B. Mak, R. Hsiao, S. Ho, and J. T. Kwok, "Embedded Kernel Eigenvoice Speaker Adaptation and Its Implication to Reference Speaker Weighting *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.