A COMPARISON OF VARIOUS ADAPTATION METHODS FOR SPEAKER VERIFICATION WITH LIMITED ENROLLMENT DATA

Man-Wai Mak,[*] *Roger Hsiao*[†] and *Brian Mak*[‡]

[*]Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong
[†]Language Technology Institute, School of Computer Science
Carnegie Mellon University, Pittsburgh, USA
[‡]Dept. of Computer Science,
The Hong Kong University of Science and Technology, Hong Kong

ABSTRACT

One key factor that hinders the widespread deployment of speaker verification technologies is the requirement of long enrollment utterances to guarantee low error rate during verification. To gain user acceptance of speaker verification technologies, adaptation algorithms that can enroll speakers with short utterances are highly essential. To this end, this paper applies kernel eigenspace-based MLLR (KEMLLR) for speaker enrollment and compares its performance against three state-of-the-art model adaptation techniques: maximum a posteriori (MAP), maximum-likelihood linear regression (MLLR), and reference speaker weighting (RSW). The techniques were compared under the NIST2001 SRE framework, with enrollment data vary from 2 to 32 seconds. Experimental results show that KEMLLR is most effective for short enrollment utterances (between 2 to 4 seconds) and that MAP performs better when long utterances (32 seconds) are available.

1. INTRODUCTION

State-of-the-art approaches to text-independent speaker verification use mel-frequency cepstral coefficients (MFCCs) as speaker features and Gaussian mixture models (GMMs) [1] for statistical speaker modeling. To increase the ability to discriminate between client speakers and impostors, a GMM-based background model is typically used to represent the characteristics of impostors. During verification, the ratio of the likelihood that the claimant is a genuine speaker to the likelihood that the claimant is an impostor is compared against a decision threshold for decision making. In case enrollment data for individual speakers are scarce, speaker models can be adapted from the background model using the maximum a posteriori (MAP) technique [2].

So far, most of the speaker verification systems and evaluations use long utterances for enrollment. For example, the evaluation protocols of the NIST SRE use 2-minute enrollment utterances for training and 15 to 45 seconds of speech for verification. No doubt the results of these evaluations are valuable for assessing the practicality of speaker verification systems. However, to maximize user convenience in real applications, short enrollment utterances are highly desirable. This paper aims to compare the performance of different model adaptation techniques (including MAP [2], MLLR [3], and RSW [6]) for speaker enrollment under the practical situations where the amount of enrollment data is very limited. The paper also proposes using a kernel eigenspace-based MLLR (KEMLLR) adaptation approach for creating speaker models. It was found that KEMLLR is significantly better than all other methods investigated when the enrollment utterances contain less than or equal to 4 seconds of speech.

[*]This work was supported by the Research Grant Council of the Hong Kong SAR (Project Nos. CUHK 1/02C and PolyU 5214/04E).

[†]Roger completed this work while he was with the Hong Kong University of Science and Technology before he left for CMU.

[‡]This research is partially supported by the Research Grants Council of the Hong Kong SAR under the grant number CA02/03.EG04.

2. KERNEL EIGENSPACE-BASED MLLR (KEMLLR) ADAPTATION

KEMLLR [9,10] is a kernel version of the eigenspace-based MLLR (EMLLR) adaptation [11]. EMLLR belongs to the group of eigenspace-based adaptation methods, in which an eigenspace is computed from a (hopefully large) set of speaker-dependent (SD) models; any speaker, either training or test speaker, is then represented as a point in the speaker space spanned by the leading eigenvectors of the computed eigenspace. The SD models are represented by vectors, and the eigenspace is found by performing principal component analysis (PCA) on the SD model vectors.

2.1. Eigenvoice and Eigenspace-based MLLR

Though the eigenspace-based adaptation may be considered as a special case of adaptation methods based on speaker clustering, the eigenvoice (EV) adaptation [12] is generally regarded as the first well-known eigenspace-based adaptation method. EV and EMLLR differ in how the training SD models are represented: In EV, each SD model is the result of splicing all Gaussian mean vectors of the speaker's HMM into what is called a *speaker supervector*. In EMLLR, the SD models are created from a single speaker-independent (SI) model using MLLR adaptation. For simplicity, let's assume that only a single global MLLR transform is used. Thus, the gth Gaussian mean vector $\boldsymbol{\mu}_g^{(i)} \in \mathbb{R}^d$ of the ith speaker is given by

$$\boldsymbol{\mu}_g^{(i)} = \mathbf{Y}^{(i)'} \boldsymbol{\xi}_g^{(si)}, \tag{1}$$

where $\mathbf{Y}^{(i)'} \in \mathbb{R}^{d\times(d+1)}$ is the speaker's global MLLR transform, and $\boldsymbol{\xi}_g^{(si)} = [\boldsymbol{\mu}_g^{(si)'}, 1]'$ is the augmented mean vector of the corresponding Gaussian in the SI model. Then each speaker model is represented by his/her vectorized MLLR transform. PCA is performed on the set of vectorized MLLR transforms, and the new speaker's vectorized MLLR transform $vec(\mathbf{Y}^{(emllr)})$ is assumed to be a linear combination of the leading, say, M eigenmatrices $\mathbf{v}_m, m = 1, \ldots, M$ as follows:

$$vec(\mathbf{Y}^{(emllr)}) = \sum_{m=1}^{M} w_m \mathbf{v}_m^{(emllr)} , \qquad (2)$$

where $w_m, m = 1, \ldots, M$, are the eigenmatrix weights. Again, the eigenmatrix weights are usually determined by maximizing the likelihood of the adaptation data.

2.2. Extension of EMLLR to KEMLLR

KEMLLR tries to improve the performance of EMLLR by exploiting possible non-linearity in the speaker transform space. This is achieved by replacing linear PCA by kernel PCA in a way analogous to the use of kernel methods in *kernel eigenvoice* (KEV) adaptation [13].

Readers are referred to [9, 10] for the details of KEMLLR. Instead, we will outline the basic steps below.

Step 1: Conceptually, the speaker MLLR transformation vectors are mapped to a high-dimensional feature space using a mapping function φ. In the actual computation, φ needs not be known. Instead, a kernel function $k(\cdot,\cdot)$ is defined to compute the similarity of the mapped MLLR transformation vectors.

Step 2: Perform kernel PCA to find out the eigenmatrices in the kernel-induced feature space. As in all kernel methods, these eigenmatrices are expressed in terms of the mapped training data.

Step 3: Express the new speaker's transformation vector in the *feature space* in terms of the unknown eigenmatrix weights $w_m, m = 1, \ldots, M$, assuming that the first M leading eigenmatrices are chosen to represent the speaker eigenspace.

Step 4: Express the similarity between the new speaker's transformation vector and any Gaussian mean vector of the SI model in the feature space, again, in terms of w_m.

Step 5: Design a kernel function so that the result of Step 4 may be used to compute the adapted mean vectors of the new speaker, and hence the likelihood of the adaptation speech due to the adapted model.

Step 6: Estimate the eigenmatrix weights by maximizing the likelihood of Step 5 using any gradient-based optimization algorithm.

3. REFERENCE SPEAKER WEIGHTING (RSW)

RSW adaptation [6] is one kind of speaker-clustering-based adaptation methods [7] in which the new speaker's model is assumed to be a linear combination of a small set of reference speaker models. That is, if there are M reference speaker models $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$, then RSW computes the new speaker's model $\mathbf{s}^{(rsw)}$ as

$$\mathbf{s}^{(rsw)} = \sum_{m=1}^{M} w_m \mathbf{x}_m , \qquad (3)$$

where $\sum_{m=1}^{M} w_m = 1.0$. When M is small, fast speaker adaptation is possible due to the small number of parameters—the combination weights $w_m, m = 1, \ldots, M$. In [6], the combination weights are found by maximizing the likelihood of the adaptation data from the new speaker. Moreover, the set of reference speakers for each new speaker is selected from speaker clusters created by a hierarchical speaker clustering algorithm based on the gender and speaking rate of the training speakers.

The implementation of RSW in this paper is different from that in [6] in two aspects:

- The weights are not required to sum to 1.0 so that the new speaker model can be anywhere in the span of the reference speaker models.
- The subset of M training speakers who have the highest likelihood of the adaptation data are taken as the reference speakers; we call them the *maximum likelihood (ML) reference speakers*. The hypothesis is that the new speaker should be closest to those speakers, and, thus, in their span. In [8], we show that RSW using ML reference speakers performs much better than using clustered speakers as in [6].

4. EXPERIMENTAL EVALUATIONS

4.1. Speech Data and Features

The 2001 NIST speaker recognition evaluation set [4], which contains cellular phone speech of 74 male and 100 female target speakers extracted from the SwitchBoard-II Phase IV Corpus, was used in the evaluation. The corpus allows a maximum of 2 minutes of speech for training each target-speaker model (i.e., enrollment), and it provides a total of 850 male and 1,188 female utterances for testing (i.e., verification). There are one target and 10 impostor trials for each verification utterance, which amount to a total of 2,038 target trials and 20,380 impostor attempts for 2,038 verification utterances.

Mel-frequency cepstral coefficients (MFCCs) and their first-order derivatives were computed every 14ms using a Hamming window of 28ms. Cepstral mean subtraction (CMS) was applied to the MFCCs to remove linear channel effects. The MFCCs and delta MFCCs were concatenated to form 24-dimensional feature vectors.

4.2. Enrollment: Model Adaptation

A 1,024-component universal background model (UBM) [2] 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 [2], MLLR transformation (with full transformation matrices) [3], RSW [6], and the proposed KEMLLR. To investigate the performance of these adaptation methods under short-utterance scenarios, enrollment utterances of 2s, 4s, 8s, 16s, and 32s were used. These utterances were created as follows. First, silence segments in the 2-minute enrollment utterances were removed by using an energy- and zero crossing-based speech detector. Then, speech segments of 2 to 32 seconds were randomly extracted from each of the silence-removed speech files.

The parameters for RSW and KEMLLR were set as follows.

- Parameters for RSW adaptation:
  - M = 60 in Eq. 3.

- Parameters for KEMLLR adaptation:
  - initial learning rate = 0.00001.
  - Gaussian kernels are used with $\beta_r = \beta = 0.001$ for $r = 1, \ldots, R$ (see the definition of β_r in [10]). That is, all constituent Gaussian kernels have the same global β value.
  - The eigenmatrix weights $w_m, m = 1, \ldots, M$, were initialized by projecting the following transformation

$$\mathbf{Y}^{(si)} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad (4)$$

  onto each of the M kernel eigenmatrices after it was normalized and φ-mapped to the kernel-induced feature space.

5. RESULTS AND DISCUSSIONS

Figure 1 plots the EERs against the length of enrollment utterances for the four adaptation methods. As expected, MAP adaptation outperforms others when long utterances (e.g. 32 seconds) are available for enrollment.[1] However, as shown in the figure, KEMLLR achieves the lowest EERs under very short-utterance scenarios (e.g., 2 and 4 seconds). The differences between the EERs obtained by KEMLLR and its closest rival for 2-second and 4-second enrollment utterances are statistical significant, as evident from the P-values of McNemar's tests [5] shown in Table 1.

Figure 2 shows the DET performance and minimum DCF (in the legend) of the four adaptation methods for 4-second enrollment utterances. Evidently, KEMLLR outperforms other methods for a wide range of decision thresholds.

Figure 3 explains why KEMLLR outperforms MAP when the amount of adaptation data is limited. The figure shows the projection of the 1024 centers of a speaker model and the background model onto the first two cepstral axes for the case of 4-second enrollment utterances. Evidently, most of the speaker centers in the MAP case (Figure 3(a)) overlap with those of the background centers, suggesting that MAP is not very effective in adapting the centers. This is mainly because in MAP adaptation, only the centers that are sufficiently close to the adaptation data have the chance to be adapted. On the other hand, because of the global adaptation characteristic of KEMLLR, almost all of the speaker centers in Figure 3(b) have been adapted.

6. 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. Short speech segments ranging from 2 to 32 seconds were extracted from the NIST 2001 corpus for creating speaker models. The resulting models were evaluated under the NIST 2001 SRE protocols. It was found that KEMLLR outperforms other adaptation methods when the amount of enrollment data is very limited and that when a large amount of enrollment data is available, MAP is a better candidate for creating speaker models.

[1] Note that our findings with 32s of adaptation data agree well with a previous study using 2 minutes of adaptation data [14].

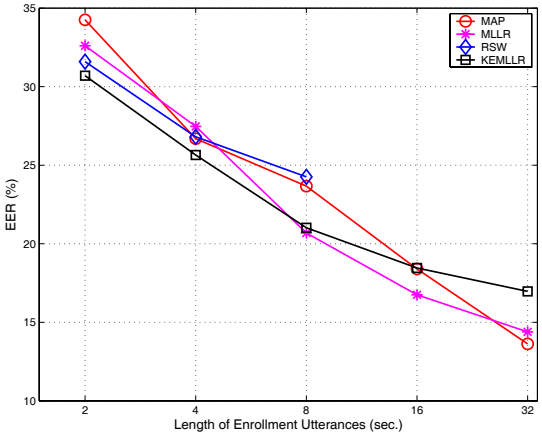

Fig. 1. EER versus length of enrollment utterances. *Note*: For MLLR and KEMLLR, the number of regression classes was set to 1, 2, 4, or 6, and the one that gave the lowest EERs was reported. Because the EER of RSW is the highest among all methods at 8s and its trend also suggests that it will perform poorer than others beyond 8s, no experiments were carried out for RSW beyond 8s.

7. REFERENCES

[1] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communications*, vol. 17, pp. 91–108, 1995.

[2] 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.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[4] *http://www.nist.gov/speech/tests/spk/index.htm*.

[5] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP'89*, 1989, pp. 532–535.

[6] Tim J. Hazen, "A comparison of novel techniques for rapid speaker adaptation," *Speech Communications*, vol. 31, pp. 15–33, May 2000.

[7] T. Kosaka, S. Matsunaga, and S. Sagayama, "Speaker-independent speech recognition based on tree-structured speaker clustering," *Journal of Computer Speech and Language*, vol. 10, pp. 55–74, 1996.

[8] B. Mak, S. Ho, R. Hsiao, and J. T. Kwok, "Embedded kernel eigenvoice speaker adaptation and its implication to reference speaker weighting," *IEEE Transactions on Speech and Audio Processing*, 2005, (in press).

[9] B. Mak and R. Hsiao, "Improving eigenspace-based MLLR adaptation by kernel PCA," in *Proceedings of the ICSLP*, Jeju Island, South Korea, October 14–18 2004, vol. I, pp. 13–16.

[10] R. Hsiao and B. Mak, "Kernel eigenspace-based MLLR adaptation using multiple regression classes," in *Proceedings*

	2-second Enrollment Utterances		
	MLLR	RSW	KEMLLR
MAP	0.002401	0.000000	0.000000
MLLR	–	0.000669	0.000000
RSW	–	–	**0.008735**
	4-second Enrollment Utterances		
	MLLR	RSW	KEMLLR
MAP	0.012517	0.607289	**0.008577**
MLLR	–	0.014289	0.000000
RSW	–	–	0.000020
	8-second Enrollment Utterances		
	MLLR	RSW	KEMLLR
MAP	0.000000	0.214671	0.000000
MLLR	–	0.000000	0.133490
RSW	–	–	0.000000

Table 1. P-values of McNemar's tests [5] for utterance length of 2, 4, and 8 seconds showing the statistical significance between the EERs produced by different adaptation methods. A P-value less than 0.05 means that the EERs produced by the two corresponding adaptation methods are significantly different at a significance level of 5%. For example, when 4-second enrollment utterances were used, the EER of KEMLLR (25.65%) is significantly different from that of its closest rival MAP (26.69%), because $P = 0.008577 < 0.05$.

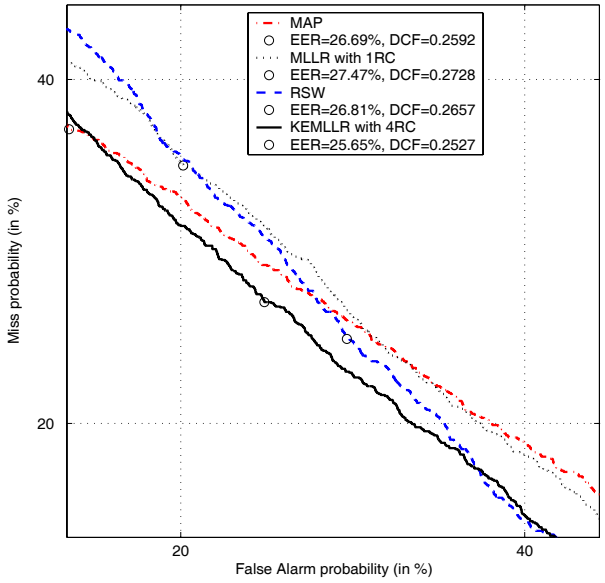

Fig. 2. DET plots based on 4-second enrollment utterances. The numbers of regression classes (RC) that gave the lowest EERs were used for plotting the DET curves of MLLR and KEMLLR.

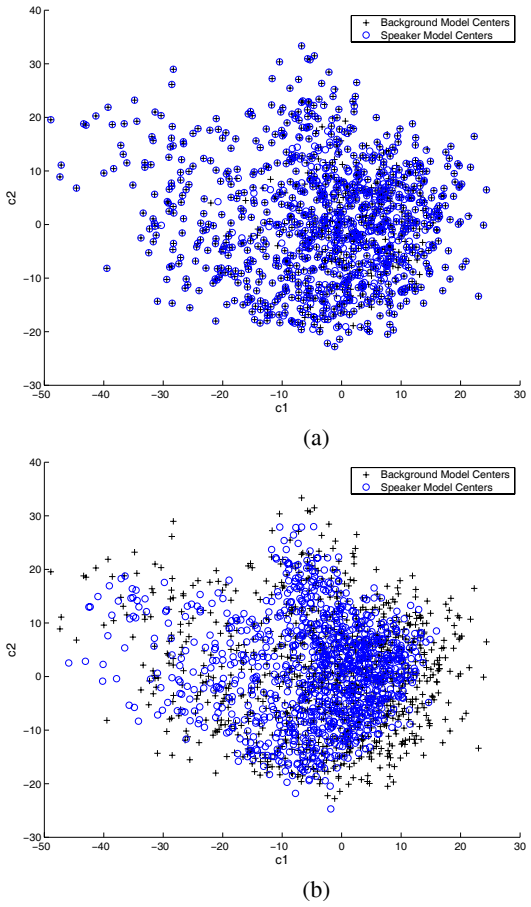

Fig. 3. Cluster plots showing the projection of the speaker model's centers ('○') and background model's centers ('+') onto the first two cepstral axes when the enrollment utterances contain 4 seconds of speech. (a) MAP adaptation and (b) KEMLLR adaptation.

of the IEEE ICASSP, Philadelphia, USA, March 18–23 2005, vol. 1, pp. 985–988.

[11] K. T. Chen, W. W. Liau, H. M. Wang, and L. S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Proceedings of the ICSLP*, 2000, vol. 3, pp. 742–745.

[12] 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.

[13] 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.

[14] J. Mariéthoz and S. Bengio, "A comparative study of adaptation methods for speaker verification," in *Proceedings of the ICSLP*, 2002, pp. 581–584.