IMPROVING REFERENCE SPEAKER WEIGHTING ADAPTATION BY THE USE OF MAXIMUM-LIKELIHOOD REFERENCE SPEAKERS

Brian Mak[*] and *Tsz-Chung Lai*

Department of Computer Science
Hong Kong University of Science & Technology
Clear Water Bay, Hong Kong

Roger Hsiao[†]

Language Technology Institute
School of Computer Science
Carnegie Mellon University, Pittsburgh, USA

ABSTRACT

We would like to revisit a simple fast adaptation technique called *reference speaker weighting* (RSW). RSW is similar to eigenvoice (EV) adaptation, and simply requires the model of a new speaker to lie on the span of a set of reference speaker vectors. In the original RSW, the reference speakers are computed through a hierarchical speaker clustering (HSC) algorithm using information such as the gender and speaking rate. We show in this paper that RSW adaptation may be improved if those training speakers that have the highest likelihoods of the adaptation data are selected as the reference speakers; we call them the *maximum-likelihood (ML) reference speakers*. When RSW adaptation was evaluated on WSJ0 using 5s of adaptation speech, the word error rate reduction can be boosted from 2.54% to 9.15% by using 10 ML reference speakers instead of reference speakers determined from HSC. Moreover, when compared with EV, MAP, MLLR, and eKEV on fast adaptation, we are surprised that the algorithmically simplest RSW technique actually gives the best performance.

1. INTRODUCTION

Model-based adaptation methods like the *speaker-clustering-based* methods [1], the Bayesian-based *maximum a posteriori* (MAP) adaptation [2], and the transformation-based *maximum likelihood linear regression* (MLLR) adaptation [3] have been popular for many years. Nevertheless, when the amount of available adaptation speech is really small — for example, only a few seconds — other techniques are required to further reduce the number of adaptation parameters.

Two similar fast-speaker adaptation methods were proposed at around the same time: *reference speaker weighting* (RSW) [4, 5] in 1997 and *eigenvoice* (EV) [6, 7] in 1998. Both methods require the model of a new speaker to lie on the span of some reference vectors; they differ only in the ways the reference vectors are computed. Eigenvoice employs principal component analysis to find a set of orthogonal basis vectors for the purpose, and these eigenvectors are commonly known as eigenvoices. On the other hand, RSW, in its simplest form, simply selects a subset of training speakers and uses their models as the references. It has been shown that when there are only a few seconds of adaptation data, both adaptation approaches may improve the recognition performance of the speaker-independent model significantly by using only a small set of reference vectors (say, fewer than 10). However, it seems to us that while EV has drawn a lot of attention and spawns a myriad of eigenspace-based adaptation methods such as eigen-MLLR[8], eigenspace mapping [9], and kernel eigenvoice [10, 11], etc., the simpler RSW adaptation technique has not been as well known as it should be.

In this paper, we would like to revisit the *reference speaker weighting* technique, and suggest the use of *maximum-likelihood (ML) reference speakers* to further improve its performance. The use of ML reference speakers is motivated by our previous work on the embedded kernel eigenvoice (eKEV) adaptation [12, 11]. In eKEV adaptation, a speaker-adapted model is first formulated in a high-dimensional kernel-induced feature space, and is then mapped back to an approximate *pre-image* in the input speaker space. The pre-imaging process is guided by the principle of multi-dimensional scaling with the use of distance constraints between the new speaker and his "closest" neighbors. In [11], we showed that the use of a few ML neighbors in eKEV resulted in good adaptation performance.

This paper is organized as follows. We first review the theory of reference speaker weighting (RSW) in the next Section, and discuss two different ways of defining the reference speakers in Section 3. RSW was evaluated on the Wall Street Journal corpus WSJ0 in Section 4. Finally, in Section 5, we present some concluding remarks.

2. REFERENCE SPEAKER WEIGHTING (RSW)

In this section, we will review the theory of reference speaker weighting in its simplest form. It is basically the same as that in [5] except with some minor modifications that will be

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

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

pointed out later.

Let's consider a speech corpus consisting of N training speakers with diverse speaking or voicing characteristics. A speaker-independent (SI) model is first estimated from the whole corpus. The SI model is a hidden Markov model (HMM), and its state probability density functions are modeled by mixtures of Gaussians. Let's further assume that there are a total of R Gaussians in the SI HMM. Then, a speaker-dependent (SD) model is created for each of the N training speakers by MLLR transformation [3] of the SI model, so that all SD models have the same topology. To perform RSW adaptation, each SD model is represented by what is called a *speaker supervector* that is composed by splicing all its R Gaussian mean vectors together.

In RSW adaptation, a subset of M reference speakers $\Omega(\mathbf{s})$ is chosen among the N training speaker with $M \leq N$ for the adaptation of a new speaker $\mathbf{s}$. (Notice that the set of reference speakers, in general, is different for different new speakers.) Let $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M\}$ be the set of reference speaker supervectors. Then the RSW estimate of the new speaker's supervector is

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

and for the mean vector of the rth Gaussian,

$$\mathbf{s}_r^{(rsw)} = \sum_{m=1}^{M} w_m \mathbf{y}_{mr} = \mathbf{Y}_r \mathbf{w} \,. \tag{2}$$

where $\mathbf{w} = [w_1, w_2, \ldots, w_M]'$ is the combination weight vector.

In the ML estimation of $\mathbf{w}$, given the adaptation data $\mathbf{O} = \{\mathbf{o}_t, t = 1, \ldots, T\}$, one maximizes the following $Q(\mathbf{w})$ function:

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

where $\gamma_t(r)$ is the posterior probability of observing $\mathbf{o}_t$ in the rth Gaussian, and $\mathbf{C}_r$ is the covariance matrix of the rth Gaussian. The optimal weight vector may be found by simple calculus as follows:

$$\begin{aligned}
&\frac{\partial Q}{\partial \mathbf{w}} = 2\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_t(r)\mathbf{Y}_r'\mathbf{C}_r^{-1}(\mathbf{o}_t - \mathbf{Y}_r\mathbf{w}) = 0 \\
\Rightarrow \quad &\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_t(r)\mathbf{Y}_r'\mathbf{C}_r^{-1}\mathbf{o}_t = \sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_t(r)\mathbf{Y}_r'\mathbf{C}_r^{-1}\mathbf{Y}_r\mathbf{w} \\
\Rightarrow \quad &\mathbf{w} = \left[\sum_{r=1}^{R}\left(\sum_{t=1}^{T}\gamma_t(r)\right)\mathbf{Y}_r'\mathbf{C}_r^{-1}\mathbf{Y}_r\right]^{-1} \\
&\qquad \left[\sum_{r=1}^{R}\mathbf{Y}_r'\mathbf{C}_r^{-1}\left(\sum_{t=1}^{T}\gamma_t(r)\mathbf{o}_t\right)\right] \,.
\end{aligned} \tag{3}$$

Thus, the weights $\mathbf{w}$ may be obtained by solving a system of M linear equations.

Our description is simpler than [5] in that

- the speaker model is simply represented by a speaker supervector as commonly used in eigenvoice adaptation. In [5], it is represented by some centroid of the Gaussian components of each HMM state.
- [5] also requires $\sum_{m=1}^{M} w_m = 1$. We remove this constraint and allow the new speaker to be anywhere in the span of the M reference speaker supervectors.

Notice that the reference speaker vectors in the description above may be generalized to any set of reference vectors.

3. REFERENCE SPEAKERS SELECTION

In this paper, we will investigate two ways of selecting the reference speakers for a new speaker.

1. Hierarchical Speaker Clustering (HSC)
   The training speakers are hierarchically clustered offline onto a tree structure [1] using criteria such as speaking rate, gender, voice characteristics, etc. During the RSW adaptation of a new speaker, his adaptation data are first classified into one of the leaf clusters of the HSC tree, and the training speakers belonging to that leaf cluster are his reference speakers.

2. Maximum-Likelihood (ML) Reference Speakers
   At the beginning of RSW adaptation of a new speaker, the likelihood of his adaptation speech with respect to each training speaker model is computed and sorted in descending order. The top M training speakers who have the highest likelihood of the adaptation data are taken as the reference speakers of the new speaker. The hypothesis is that the new speaker should be closest to those speakers, and, thus, in their span.

4. EXPERIMENTAL EVALUATION

The fast speaker adaptation performance of reference speaker weighting (RSW) was tested on the Wall Street Journal speech corpus WSJ0 [13] using 5s and 10s of adaptation data in the supervised mode.

4.1. WSJ0 Corpus

The standard SI-84 training set was used for training the speaker-independent (SI) model. It consists of 83 speakers and 7138 utterances for a total of about 14 hours of training speech. The standard nov'92 5K non-verbalized test set was used for evaluation. It consists of 8 speakers, each with about 40 utterances. During evaluation, for each of the 8 testing speakers, 1–3 utterances of his speech were randomly selected so that

the amount of adaptation speech is about 5s or 10s (or, 4s and 8s respectively if one excludes the silence portions), and his adapted model was tested on his remaining speech in the test set. Notice that all test data are not endpointed before recognition. This was repeated three times and the three adaptation results were averaged before they were reported. Finally, a bigram language model of perplexity 147 was employed in this recognition task.

4.2. Acoustic Modeling

The traditional 39-dimensional MFCC vectors were extracted at every 10ms over a window of 25ms. The speaker-independent (SI) model consists of 15,449 cross-word triphones based on 39 base phonemes. Each of them was modeled as a continuous density HMM (CDHMM) which is strictly left-to-right and has three states with a Gaussian mixture density of 16 components per state. The SI model has a word recognition accuracy of 92.13% on the test data.

The SD models were created by MLLR adaptation using a regression class tree of 32 classes.

Table 1. RSW performance on WSJ0 using different types of reference speakers. Results are word accuracies in %. (Figures in parentheses are the WER reductions in %.)

Reference Speakers	#Speakers	5s	10s
HSC	˜14	92.33 (2.54)	92.41 (3.43)
ML	10	92.85 (9.15)	92.78 (8.26)

4.3. Effect of Different Reference Speaker Selections

RSW was tested with two different definitions of reference speakers:

- All 83 training speakers were clustered by hierarchical speaker clustering (HSC) as in [5]. Thus, the speakers were first clustered according to their gender and then their speaking rate. Three speaking rates were defined: slow, medium, and fast. As a result, we got a clustering tree with six leaves — speaker clusters — and each cluster consists of roughly 14 training speakers.

- M maximum-likelihood (ML) reference speakers as described in Section 3. M is set to 10.

The results are shown in Table 1. It can be seen that the definition of reference speakers is essential to the performance of RSW adaptation. For example, with only 5s of adaptation speech, the clustered speaker groups based on gender and speaking rate give only a small improvement of 2.54% reduction in the word error rate (WER); on the other hand, the use of 10 ML reference speakers boosts the WER reduction to 9.15%.

Hereafter, ML reference speakers are used in all RSW adaptation experiments.

4.4. Effect of the Number of ML Reference Speakers

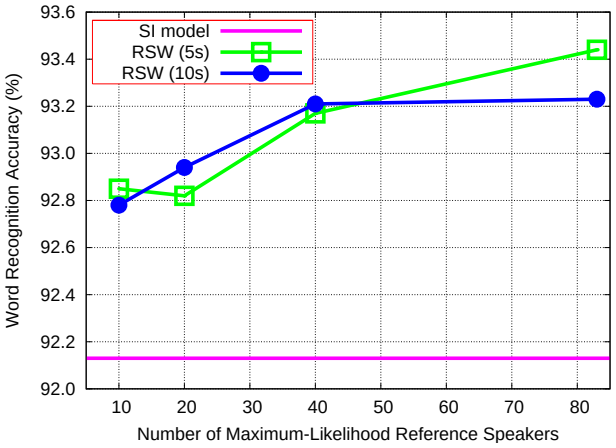

Fig. 1. Effect of the number of ML reference speakers on RSW.

The idea of using HSC or ML reference speakers is to make use of the most important local information to reduce the number of estimation parameters. In this experiment, we would like to investigate the effect of additional reference speakers by doubling the number of reference speakers until all 83 training speakers were used. The results are plotted in Fig. 1. The figure shows that

- The performance of RSW with 5s adaptation and 10s adaptation is very similar, indicating that the method saturates very fast between 5s and 10s.

- The 10s adaptation performance seems to be more steady as it improves monotonically with additional reference speakers until it saturates with 40 ML reference speakers.

- However, in both cases, it shows that, in this task, using **all** training speakers as the reference speakers gives the best adaptation performance.

4.5. Comparison with Other Adaptation Methods

Finally, RSW adaptation was compared with the SI model and the following common adaptation methods:

EV: the speaker-adapted (SA) model found by EV adaptation [14].

MAP: the SA model found by MAP adaptation [2].

MLLR: the SA model found by MLLR adaptation [3].

eKEV: the SA model found by the embedded kernel eigenvoice adaptation method [11].

Table 2. Comparing RSW with other common adaptation methods on WSJ0. Results are word accuracies in %. (Figures in parentheses are the WER reductions in %.)

Model/Method	5s	10s
SI	92.13	92.13
EV	92.46 (4.19)	92.51 (4.83)
MAP	92.48 (4.45)	92.47 (4.32)
MLLR	92.32 (2.41)	92.98 (10.8)
eKEV	92.86 (9.28)	92.92 (10.0)
RSW	93.44 (16.6)	93.23 (14.0)

For each adaptation method, we tried our best effort to get the best performance. MAP and MLLR were performed using HTK, and we implemented the other adaptation methods. For MAP, the best results with a scaling factor in the range of 3–12 were reported; MLLR made use of a regression tree of 32 regression classes, and the better results of using diagonal- or full-MLLR transforms were reported; the basic EV was implemented and 10 eigenvoices were found to give good results; for eKEV adaptation, 10 ML reference speakers gave the best results; finally, the best RSW results using all 83 training speakers as reference speakers were used for the comparison. The results are summarized in Table 2.

From Table 2, we are surprised that the algorithmically simplest RSW technique actually gives the best fast adaptation performance.

5. CONCLUSIONS

In this paper, we revisit the use of reference speaker weighting for fast speaker adaptation on a large-vocabulary task WSJ0. We also simply select a subset of training speakers as the reference speakers instead of more complicated speaker-clustered models. We find that the maximum-likelihood of the adaptation data can be a good measure to select the reference speakers among the training speakers. Moreover, we found that for WSJ0 with only 83 training speakers, using all training speakers will give the best RSW performance for both 5s and 10s adaptation. However, we believe that for a larger speech corpus with many more training speakers, one may still want to use a subset of M maximum-likelihood reference speakers for RSW adaptation in order to reduce the computation in the estimation of the combination weights which has a complexity of $O(M^3)$. This is further supported by our finding on WSJ0 that for 10s adaptation, RSW performance using 40 ML reference speakers is as good as that using 83 ML reference speakers. Further experiments with more speech corpora of various sizes of training speakers are needed.

6. REFERENCES

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

[2] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on SAP*, vol. 2, no. 2, pp. 291–298, April 1994.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Journal of CSL*, vol. 9, pp. 171–185, 1995.

[4] Tim J. Hazen and James R. Glass, "A comparison of novel techniques for instantaneous speaker adaptation," in *Proc. of Eurospeech*, 1997, pp. 2047–2050.

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

[6] R. Kuhn, P. Nguyen, J.-C. Junqua, et al., "Eigenvoices for speaker adaptation," in *Proc. of ICSLP*, 1998, vol. 5, pp. 1771–1774.

[7] H. Botterweck, "Very fast adaptation for large vocabulary continuous speech recognition using eigenvoices," in *Proc. of ICSLP*, 2000, vol. 4, pp. 354–357.

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

[9] B. Zhou and J. Hansen, "Rapid discriminative acoustic model based on eigenspace mapping for fast speaker adaptation," *IEEE Trans. on SAP*, vol. 13, no. 4, pp. 554–564, July 2005.

[10] B. Mak, J. T. Kwok, and S. Ho, "Kernel eigenvoice speaker adaptation," *IEEE Trans. on SAP*, vol. 13, no. 5, pp. 984–992, September 2005.

[11] B. Mak and S. Ho, "Various reference speakers determination methods for embedded kernel eigenvoice speaker adaptation," in *Proc. of ICASSP*, 2005, vol. 1, pp. 981–984.

[12] B. Mak, S. Ho, and J. T. Kwok, "Speedup of kernel eigenvoice speaker adaptation by embedded kernel PCA," in *Proc. of ICSLP*, 2004, vol. IV, pp. 2913–2916.

[13] D. B. Paul and J. M. Baker, "The design of the Wall Street Journal-based CSR corpus," in *Proceedings of the DARPA Speech and Natural Language Workshop*, Feb. 1992.

[14] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on SAP*, vol. 8, no. 4, pp. 695–707, Nov 2000.