Unsupervised Speaker Adaptation using Reference Speaker Weighting

Tsz-Chung Lai and Brian Mak

Department of Computer Science and Engineering
The Hong Kong University of Science & Technology
Clear Water Bay, Hong Kong
{kimo, mak}@cse.ust.hk

Abstract. Recently, we revisited the fast adaptation method called *reference speaker weighting* (RSW), and suggested a few modifications. We then showed that the algorithmically simplest technique actually outperformed conventional adaptation techniques like MAP and MLLR for 5- or 10-second supervised adaptation on the Wall Street Journal 5K task. In this paper, we would like to further investigate the performance of RSW in unsupervised adaptation mode, which is the more natural way of doing adaptation in practice. Moreover, various analyses were carried out on the reference speakers computed by the method.

1 Introduction

In practice, most automatic speech recognition systems come with a speaker-independent (SI) acoustic model that is expected to work sufficiently well with most users in general. However, the recognition performance can be further improved for a particular user if the SI model is fine-tuned to the speaking characteristics of the user through an appropriate speaker adaptation procedure. In particular, fast unsupervised speaker adaptation method that requires only a few seconds of adaptation speech from the users without knowing its content in advance is more desirable, and in some cases (e.g. phone enquiries), is the only feasible adaptation solution. Two similar fast speaker adaptation methods were proposed at about the same time: *reference speaker weighting* (RSW) [1, 2] in 1997 and *eigenvoice* (EV) [3, 4] in 1998. Both methods have their root in *speaker-clustering-based* methods [5]. In both methods, a speaker model is vectorized and a new speaker-adapted (SA) model is required to be a linear combination of a set of reference vectors. In eigenvoice, an orthogonal eigenspace is derived from a set of training speakers by principal component analysis, and the eigenvectors, now called *eigenvoices*, are used as the reference vectors. On the other hand, RSW simply selects a subset of training speakers as the references.

In [6], we revisited RSW with further simplifications. We also suggested to select the reference speakers by their likelihoods on the adaptation speech. Supervised adaptation using 5- and 10-second of speech on the Wall Street Journal (WSJ0) 5K-vocabulary task showed that the algorithmically simplest RSW method actually outperformed conventional adaptation methods like the

Bayesian-based *maximum a posteriori* (MAP) adaptation [7], and the transformation-based *maximum likelihood linear regression* (MLLR) adaptation [8] as well as eigenvoice and eigen-MLLR [9]. Here, we would like to further our investigation on RSW by carrying out unsupervised adaptation which is the more natural way of doing adaptation in practice, as well as performing various analyses on the reference speakers computed by the method.

This paper is organized as follows. We first review the theory of reference speaker weighting (RSW) in the next Section. Unsupervised RSW adaptation was then evaluated on the Wall Street Journal corpus WSJ0 in Section 3. The experiments are followed by various analyses 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 [2] except with a few modifications that we have outlined in [6].

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 [8] 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 $\mathbf{s}$ as depicted in Fig. 1. (Notice that the set of reference speakers, in general, is different for each new speaker.) 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.

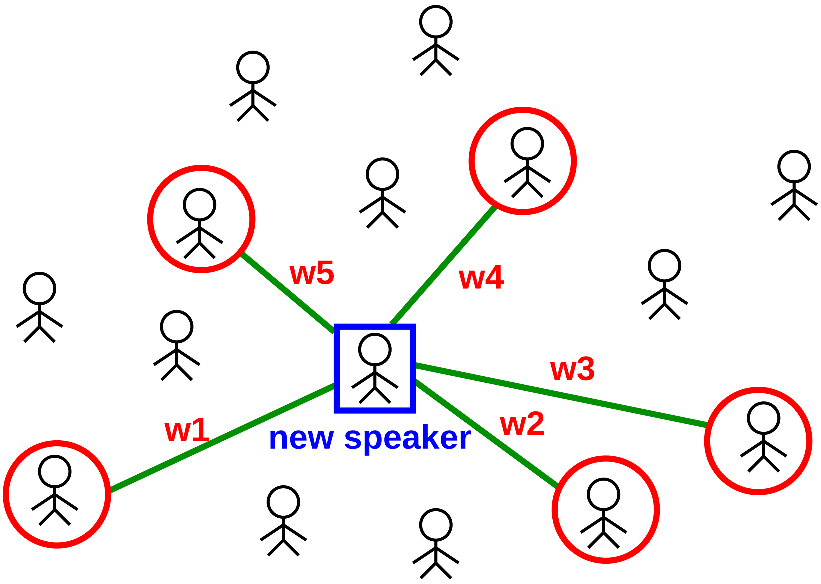

Fig. 1. Concept of reference speaker weighting.

2.1 Maximum-Likelihood Estimation of Weights

Given the adaptation data $\mathbf{O} = \{\mathbf{o}_t, t = 1, \ldots, T\}$, one may estimate $\mathbf{w}$ by maximizing 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:

$$\frac{\partial Q}{\partial \mathbf{w}} = 2\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_t(r)\mathbf{Y}_r'\mathbf{C}_r^{-1}(\mathbf{o}_t - \mathbf{Y}_r\mathbf{w}) = 0$$

$$\Rightarrow \mathbf{w} = \left[\sum_{r=1}^{R}\left(\sum_{t=1}^{T}\gamma_t(r)\right)\mathbf{Y}_r'\mathbf{C}_r^{-1}\mathbf{Y}_r\right]^{-1}\left[\sum_{r=1}^{R}\mathbf{Y}_r'\mathbf{C}_r^{-1}\left(\sum_{t=1}^{T}\gamma_t(r)\mathbf{o}_t\right)\right] . \quad (3)$$

Thus, the weights $\mathbf{w}$ may be obtained by solving a system of M linear equations. The solution requires finding the inverse of an $M \times M$ matrix and has a computational complexity of $O(M^3)$. Notice also that unlike Hazen's formulation in [2], no constraints are imposed on the combination weights.

2.2 Maximum-Likelihood Reference Speakers

In [6], we showed that good RSW adaptation performance could be achieved by selecting those training speakers that gave the highest likelihoods of the adaptation speech from a test speaker as his/her reference speakers. We call these reference speakers the maximum-likelihood (ML) reference speakers. We continue to use ML reference speakers for RSW adaptation evaluation in this paper.

3 Experimental Evaluation

Unsupervised fast speaker adaptation was carried out on the Wall Street Journal WSJ0 [10] 5K-vocabulary task using our modified reference speaker weighting (RSW) method.

Table 1. Duration statistics (in seconds) of the test utterances of each WSJ0 test speaker.

Speaker ID	#Utterances	min	max	mean	std dev
440	40	4.32	12.76	8.23	6.83
441	42	2.82	10.94	6.89	4.14
442	42	2.42	11.85	7.24	4.62
443	40	3.34	14.19	8.01	5.49
444	41	2.36	11.13	7.88	4.49
445	42	2.55	10.70	5.81	4.18
446	40	2.99	12.28	7.14	5.19
447	43	2.06	11.55	7.33	5.78

3.1 WSJ0 Corpus and the Evaluation Procedure

The standard SI-84 training set was used for training the speaker-independent (SI) model and gender-dependent (GD) models. It consists of 83 speakers (41 male speakers and 42 female 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 (5 male and 3 female speakers), each with about 40 utterances. The detailed duration statistics of the test utterances of each speaker is given in Table 1.

During unsupervised adaptation, the content of each test utterance was not assumed to be known in advance. All adaptation methods under investigation were run with 3 EM iterations. During each iteration, the current speaker-adapted (SA) model was used to decode the adaptation utterance and to provide the Gaussian mixture posterior probabilities, then the adaptation method was carried out to get a new SA model. At the first iteration, the SI model was used for decoding instead. The last SA model was used to decode the same utterance again to produce the final recognition output. Results from all speakers and all utterances are pooled together and their average results are reported. Finally, a bigram language model of perplexity 147 was employed in this recognition task.

3.2 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 triphone 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; there are 3,131 tied states in total. The SI model has a word recognition accuracy of 92.60% on the test data[1]. GD models were then created by MAP adaptation from the SI model using gender-specific training data, and they give a word recognition accuracy of 92.92%.

Furthermore, 83 speaker-dependent (SD) models were created by MLLR adaptation using a regression class tree of 32 classes for RSW adaptation methods.

3.3 Effect of the Number of Reference Speakers

We first investigate how many ML reference speakers are sufficient for RSW adaptation. Unsupervised RSW adaptation using only a single utterance at a time was performed with 3 EM iterations. We started with 10 ML reference speakers and then doubled the number until all 83 training speakers were used. The results are plotted in Fig. 2. The figure shows that although using **all** training speakers as reference speakers gives good results, the best adaptation performance actually is obtained with 40 reference speakers, though the difference is small. It also shows that RSW performance saturates fast after about half of training speakers are used as reference speakers.

3.4 Comparative Study

RSW adaptation was compared with the following models and common adaptation methods:

SI: the SI model.
GD: the gender-dependent models.
MAP: the SA model found by MAP adaptation [7].
MLLR: the SA model found by MLLR adaptation [8].

[1] The accuracy of the SI model is better than what we had reported in [6] because better values of grammar factor and insertion penalty are used.

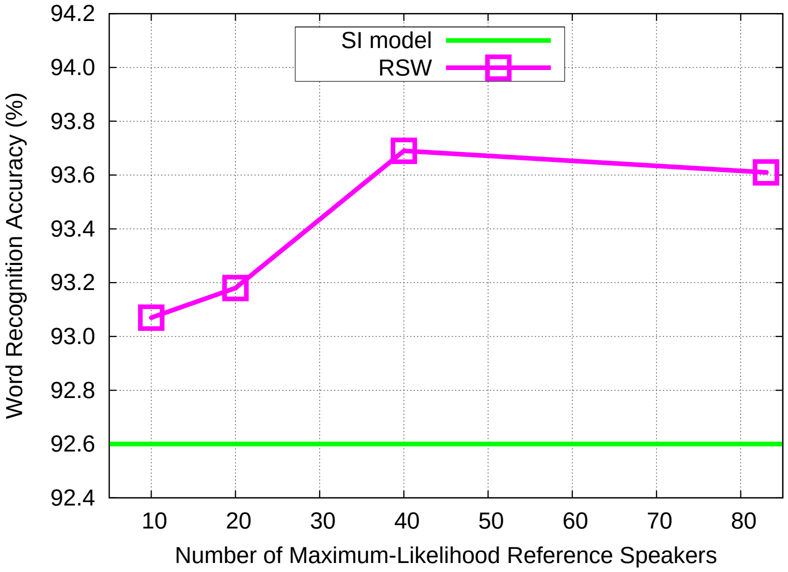

Fig. 2. Effect of the number of ML reference speakers on RSW.

Table 2. Comparing RSW with the SI and GD models, MAP and MLLR adaptation on WSJ0. Results are word accuracies in %. (WERR is word error rate reduction in %, and M is the number of reference speakers.)

Model/Method	Word Accuracy	WERR
SI	92.60	—
GD	92.98	5.14
MAP	92.60	0.0
MLLR (3 blocks)	93.24	8.65
RSW (M=10)	93.07	6.35
RSW (M=20)	93.18	7.84
RSW (M=40)	**93.69**	**14.7**
RSW (M=83)	93.61	13.6

For the evaluation using GD models, the test speaker's gender was assumed known and the GD model of the corresponding gender was applied to his/her utterances; thus, there is no error from gender detection[2]. For each adaptation method, we tried our best effort to get the best performance. MAP and MLLR were performed using HTK. For MAP, scaling factors in the range of 3–15 were attempted, but none of them gave any improvement; MLLR made use of a regres-

[2] As will be explained in Section 4.2, the gender of speaker 442 is actually female.

sion tree of 32 regression classes (though it was found actually in most cases, only a single global transform was employed) and block-diagonal transforms (with 3 blocks) as there were no improvement from using full-MLLR transforms; finally, RSW adaptation using 10, 20, 40, and 83 ML reference speakers was attempted for the comparison. Again, each time, only a single utterance was used for unsupervised adaptation and 3 EM iterations were run. The results are summarized in Table 2.

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

3.5 Saturation Effect of RSW

A more detailed look at the adaptation performance of MLLR and RSW (using 40 ML reference speakers) across the three EM iterations is shown in Fig. 3. It can be seen that MLLR does not improve much after the first iteration and RSW saturates after the second iteration.

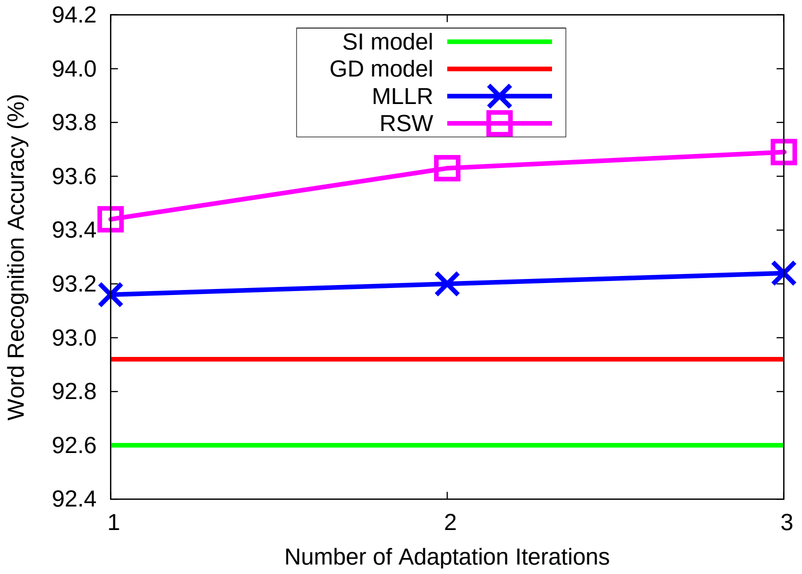

Fig. 3. Saturation effect of MLLR and RSW adaptation on WSJ0.

4 Analysis

In this section, we would like to analyze the maximum-likelihood (ML) reference speakers found for each test speaker from each test utterance.

Table 3. Consistency percentage of ML reference speakers found for each WSJ0 test speaker. (M is the number of reference speakers.)

Speaker ID	M = 10	M = 20	M = 40
440	0.830	0.854	0.908
441	0.790	0.820	0.903
442	0.860	0.863	0.903
443	0.853	0.848	0.978
444	0.868	0.870	0.915
445	0.919	0.920	0.899
446	0.815	0.794	0.887
447	0.865	0.887	0.916
Overall	0.850	0.857	0.913

4.1 Consistency of ML Reference Speakers

Since each test speaker has about 40 test utterances for adaptation, it will be interesting to see how likely that the same ML reference speakers are selected for each test utterance of the same test speaker. To do that, during unsupervised RSW adaptation using M reference speakers, the M ML reference speakers of each test utterance were recorded. Then all the ML reference speakers over all test utterances of the same test speaker are sorted according to their frequencies. Finally the total frequency of the M most frequent reference speakers are found and the percentage of their contribution over all the reference speakers is computed. The percentage is used as a measure of how consistent are the ML reference speakers found by using any utterance of a test speaker. The reference speaker consistency percentages for each test speaker is summarized in Table 3.

From Table 3, we can see that the consistency is quite high. We may conclude that (1) finding reference speakers by maximizing the likelihood of a test speaker's adaptation speech is effective, and (2) one may find the ML reference speakers using *any* utterance of a test speaker.

4.2 Consistency of ML Reference Speakers' Gender

Since gender is generally considered as a major factor affecting one's voicing characteristics, it is interesting to see if one's reference speakers have the same gender as oneself. Here is our analysis procedure: from the RSW unsupervised adaptation using M reference speakers of each of the N test utterances of a test speaker, there are totally MN reference speakers; among those MN reference speakers, count how many of them have the same gender as the test speaker's, and compute their ratio which we call the *gender consistency percentage*. Table 4 lists out the gender consistency percentages of all the 8 test speakers.

Table 4. Consistency pencentage of the gender of ML reference speakers found for each WSJ0 test speaker. (M is the number of reference speakers.)

Speaker ID	Gender	M = 10	M = 20	M = 40
440	male	0.958	0.920	0.773
441	female	0.993	0.956	0.830
442	male	0.260	0.365	0.368
443	male	1.000	1.000	0.899
444	female	0.998	0.968	0.812
445	female	0.902	0.893	0.758
446	male	0.803	0.708	0.613
447	male	0.991	0.986	0.833
Overall	—	0.862	0.849	0.735
442 as female	—	0.923	0.883	0.769

We find that when 10 reference speakers are employed by RSW adaptation, half of the 8 test speakers have a gender consistency percentage close to 100%. The consistency percentage is particular bad for the test speaker labeled as 442. However, after we listened to speaker 442's utterances, we believe that there is an error in the gender label and the speaker is actually a female. The last row of Table 4 is obtained by correcting speaker 442's gender to female. On the other hand, as expected, the gender consistency percentage drops as more reference speakers are employed in RSW adaptation. The high percentages suggest that (1) the common use of gender-dependent models for speech recognition is sensible, and (2) our approach of finding ML reference speakers may be modified to a gender detection method.

5 Conclusions

In this paper, we show that reference speaker weighting is effective for fast speaker adaptation in unsupervised mode as well as in supervised mode (the latter had been investigated in [6]). Its performance is better than MAP and MLLR on WSJ0 when only one utterance is available for unsupervised adaptation. It is also a very simple algorithm. It is also found that it is not necessary to use all training speakers as reference speakers and for this particular task, using half of training speakers actually gives slightly better adaptation results. Analyses on the reference speakers found using ML criterion show that the chosen reference speakers are very consistent across utterances from the same speaker in terms of their identity or gender.

6 Acknowledgments

This research is partially supported by the Research Grants Council of the Hong Kong SAR under the grant number DAG05/06.EG43.

References

1. Tim J. Hazen and James R. Glass, "A comparison of novel techniques for instantaneous speaker adaptation," in *Proceedings of the European Conference on Speech Communication and Technology*, 1997, pp. 2047–2050.
2. Tim J. Hazen, "A comparison of novel techniques for rapid speaker adaptation," *Speech Communications*, vol. 31, pp. 15–33, May 2000.
3. R. Kuhn, P. Nguyen, J.-C. Junqua, et al., "Eigenvoices for speaker adaptation," in *Proceedings of the International Conference on Spoken Language Processing*, 1998, vol. 5, pp. 1771–1774.
4. H. Botterweck, "Very fast adaptation for large vocabulary continuous speech recognition using eigenvoices," in *Proceedings of the International Conference on Spoken Language Processing*, 2000, vol. 4, pp. 354–357.
5. 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.
6. Brian Mak, Tsz-Chung Lai, and Roger Hsiao, "Improving reference speaker weighting adaptation by the use of maximum-likelihood reference speakers," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 14–19 2006.
7. J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.
8. C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Journal of Computer Speech and Language*, vol. 9, pp. 171–185, 1995.
9. 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 International Conference on Spoken Language Processing*, 2000, vol. 3, pp. 742–745.
10. 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.