Fast Adaptation for Robust Speech Recognition in Reverberant Environments

L. Couvreur[†], S. Dupont[†], C. Ris[†], J.-M. Boite[†] and C. Couvreur[‡]

[†]Faculté Polytechnique de Mons, Belgium | [‡]Lernout & Hauspie Speech Products, Belgium

{lcouv,dupont,ris,boite}@tcts.fpms.ac.be | christophe.couvreur@lhs.be

Abstract

We present a fast method, i.e. requiring little data, for adapting a hybrid Hidden Markov Model / Multi Layer Perceptron speech recognizer to reverberant environments. Adaptation is performed by a linear transformation of the acoustic feature space. A dimensionality reduction technique similar to the *eigenvoice* approach is also investigated. A pool of adaptation transformations are estimated *a priori* for various reverberant environments. Then, the principal directions of the pool are extracted, the so-called *eigenrooms*. The adaptation transformation for every new reverberant environment is constrained to lay on the subspace spanned by the most significant eigenrooms. Consequently, the adaptation procedure involves estimating only the projection coefficients on the selected eigenrooms, which requires less data than direct estimation of the adaptation transformation. Supervised adaptation experiments for recognition of connected digit sequences (AURORA database) in reverberant environments are carried out. Standard adaptation demonstrates improvements in word error rate higher than 30% for typical reverberation levels. The eigenroom-based adaptation technique implemented so far allows at most 50% reduction of adaptation data for the same improvement.

1. Introduction

In many real applications, automatic speech recognition (ASR) systems have to deal with noise and room reverberation. Since these systems, or more exactly their acoustic models, are commonly trained on clean speech material, i.e. noise-free and echo-free speech, they perform poorly during operation because of the mismatch between the training conditions and the operating conditions. In this work, we are primarily concerned by the mismatch due to room reverberation. Two approaches come out naturally for reducing this mismatch. One can suggest to train the acoustic models on reverberated speech. Such training material can be obtained by convolving clean speech with room impulse responses which are either measured in reverberant enclosures [1] or artificially generated [2]. Alternatively, one can suggest to recover (partially) echo-free acoustic features and keep on using the acoustic models trained on echo-free speech.

We propose here adaptation methods in the framework of connectionist speech recognition [3] to compensate for room reverberation by linear transformation of the acoustic features. The standard adaptation procedure consists in estimating the coefficients of the linear mapping from data recorded in the target operating reverberant environment. Recently, the *eigenvoice* concept has been introduced [4, 5] for reducing the dimensionality problem inherent to such adaptation procedures. The eigenvoice method increases the reliability and the efficiency of the adaptation procedure by limiting the amount of parameters that must be estimated. The method was originally devel-

oped for fast speaker adaptation of recognizers based on Hidden Markov Models / Gaussian Mixture Models (HMM/GMM) [6, 7]. In [8], the method was extended to hybrid Hidden Markov Models / Multi Layer Perceptron (HMM/MLP) recognizers. We generalize here the latter approach to room reverberation adaptation by introducing the *eigenroom* concept.

In the next section, we first review the standard technique for adaptation of a HMM/MLP recognizer by linear transformation of the acoustic features. We then describe how the eigenvoice concept can be generalized for adaptation to room reverberation, and we propose a fast version of the standard adaptation technique. In section 3, results for recognition of connected digit sequences are reported. Conclusions are drawn in section 4.

2. Adaptation Procedure

In this work, we use a Multi Layer Perceptron (MLP) as acoustic model for speech recognition [3]. Actually, a single hidden layer MLP is used. The MLP inputs are acoustic feature vectors computed for successive frames of speech along the utterance to be recognized. The MLP outputs are estimates of *a posteriori* phone probabilities. The resulting lattice of probabilities is then searched for the most likely word sequence given a lexicon of word phonetic transcriptions. Such an acoustic model is commonly trained on a large database consisting of a sequence of acoustic feature vectors and the corresponding sequence of phone labels. The training procedure aims at minimizing the square error between the actual outputs and the expected ones (1 for the output of the desired phone and 0 otherwise). This supervised training of the MLP coefficients can be efficiently implemented via a gradient descent procedure using the popular back-propagation (BP) algorithm [9].

2.1. Standard Adaptation

Unfortunately, the performance of a MLP-based speech recognizer degrades severely when the training acoustic conditions differ from the operating acoustic conditions [2]. In order to recover satisfying performance, the acoustic model has to be adapted. A usual technique for adapting a MLP consists in transforming the input acoustic feature vectors linearly [10]:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \qquad (1)$$

where $\mathbf{x}$, $\mathbf{y}$, and $(\mathbf{A}, \mathbf{b})$ denote the current acoustic feature vector possibly augmented with left and right context acoustic feature vectors, its compensated version, and the adaptation parameters, respectively. The transformed feature vector $\mathbf{y}$ serves then as input to the unchanged existing MLP. Hence, the adaptation procedure consists in estimating the adaptation parameters $(\mathbf{A}, \mathbf{b})$. The linear transformation $(\mathbf{A}, \mathbf{b})$ can be seen as
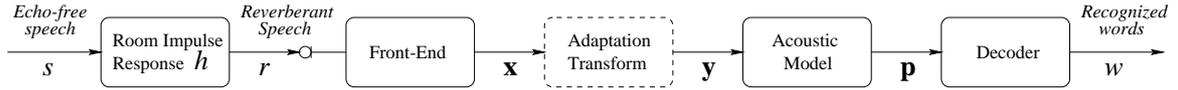
Figure 1: Adaptation scheme for hands-free speech recognition in reverberant environments.

an extra linear input layer appended to the existing MLP. Initializing this layer with the identity matrix ($\mathbf{A} = \mathbf{I}$) and zero biases ($\mathbf{b} = \mathbf{0}$), it can be estimated by resuming the supervised training of the augmented MLP on the available adaptation data, keeping all the other layers frozen [10]. In this work, we apply this procedure for adapting an existing speaker-independent MLP trained on echo-free speech to a reverberant environment (see figure 1).

2.2. Fast Adaptation

As shown in section 3, a significant amount of data is necessary to adapt efficiently the echo-free MLP, and to obtain a room-dependent but still speaker-independent acoustic model. We propose to apply a method similar to the eigenvoice approach [5] in order to reduce the amount of adaptation data. Let define $\mathbf{v}$ as the D-dimensional adaptation vector gathering the D adaptation parameters,

$$\mathbf{v} = [\text{vec}\,\{\mathbf{A}\}\,, \mathbf{b}]\,. \qquad (2)$$

Assume that such a vector may be computed for L reverberant environments with different reverberation levels. That is, a set $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_L\}$ of L adaptation vectors are computed *a priori*. Next, a Principal Component Analysis (PCA) [11] is performed on the set V. The principal directions $U = \{\mathbf{u}_1, \ldots, \mathbf{u}_D\}$ are extracted by eigendecomposition of the covariance matrix $S = \frac{1}{L-1} \sum_{l=1}^{L} (\mathbf{v}_l - \overline{\mathbf{v}})(\mathbf{v}_l - \overline{\mathbf{v}})^T$ with $\overline{\mathbf{v}} = \frac{1}{L} \sum_{l=1}^{L} \mathbf{v}_l$ denoting the mean adaptation vector. The resulting eigenvectors, the so-called *eigenrooms*, are arranged in decreasing order of the associated eigenvalues. Eventually, the estimation of the adaptation transformation for any new reverberant room is constrained to lay on the subspace in $\mathbb{R}^D$ spanned by the eigenrooms corresponding to the largest eigenvalues. That is, any new adaptation transformation $\widetilde{\mathbf{v}}$ is given by

$$\widetilde{\mathbf{v}} = \overline{\mathbf{v}} + \sum_{k=1}^{K} g_k \mathbf{u}_k = \mathbf{W}\mathbf{g} \qquad (3)$$

with $W = [\overline{\mathbf{v}}\ \mathbf{u}_1 \ldots \mathbf{u}_K]$ and $\mathbf{g} = [1\ g_1 \ldots g_K]^T$ where K denotes the number of retained eigenrooms. The adaptation procedure reduces then to estimating only the projection coefficients $\mathbf{g}$ in order to define entirely the adaptation transformation $(\mathbf{A}, \mathbf{b})$. This estimation is likely to require less adaptation material than in the standard adaptation case. The biggest part of the data, i.e. reverberated speech, is needed for building the eigenrooms, which is done *a priori*. Data for *a priori* adaptation can be either recorded in reverberant rooms, or obtained by convolving echo-free speech with impulse responses measured within the reverberant rooms [1]. In a previous work [2], we showed that artificially generated room impulse responses can model efficiently room acoustics with respect to speech recognition. The artificial room impulse responses are computed to match a high-level, perceptually meaningful, acoustic property of the target reverberant room, namely the reverberation time

T_{60}. Hence, no large collection of speech and no measure of impulse responses within reverberant rooms are involved. In this work, the computation of the eigenrooms is based on artificially reverberated material for various T_{60} using the method described in [2]. The number K of retained eigenrooms and the amount of adaptation data should tradeoff between the computational complexity and the inherent modeling error, i.e. the projection error δ, between the eigenroom-based adaptation vector $\widetilde{\mathbf{v}}$ and the standard adaptation vector $\mathbf{v}$ (see figure 2). Using (3), the adaptation equation (1) for the new reverberant environment becomes

$$
\begin{aligned}
\mathbf{y} &= \mathbf{A}(\widetilde{\mathbf{v}})\mathbf{x} + \mathbf{b}(\widetilde{\mathbf{v}}) &\qquad (4)\\
&= \left(\sum_{k=1}^{K} g_k \mathbf{A}_k + \overline{\mathbf{A}}\right)\mathbf{x} + \left(\sum_{k=1}^{K} g_k \mathbf{b}_k + \overline{\mathbf{b}}\right) &\qquad (5)
\end{aligned}
$$

with $\mathbf{A}_k = \mathbf{A}(\mathbf{v}_k)$, $\mathbf{b}_k = \mathbf{b}(\mathbf{v}_k)$, $k = 1, \ldots, K$, and $\overline{\mathbf{A}} = \mathbf{A}(\overline{\mathbf{v}})$, $\overline{\mathbf{b}} = \mathbf{b}(\overline{\mathbf{v}})$. Equivalently, each element y_i of the N-dimensional compensated vector $\mathbf{y}$ is given by, $i = 1, \ldots, N$,

$$y_i = \sum_{k=1}^{K} g_k \left(\sum_{j=1}^{N} a_{ij}^k x_j + b_i^k\right) + \sum_{j=1}^{N} \overline{a}_{ij} x_j + \overline{b}_i \qquad (6)$$

with x_j, a_{ij}^k, b_i^k, $\overline{a}_{ij}$ and $\overline{b}_i$ being the jth element of $\mathbf{x}$, the (i, j)th coefficient of $\mathbf{A}_k$, the ith element of $\mathbf{b}_k$, the (i, j)th coefficient of $\overline{\mathbf{A}}$ and the ith element of $\overline{\mathbf{b}}$, respectively.

As in the standard adaptation scheme, the fast adaptation procedure estimates the eigenroom coefficients g_k, $k = 1, \ldots, K$, by minimizing the output square error E of the resulting acoustic model via a gradient descent procedure. The recursive estimation equation is given by

$$g_k^{(\ell+1)} = g_k^{(\ell)} - \alpha \frac{\partial E}{\partial g_k^{(\ell)}} \qquad (7)$$

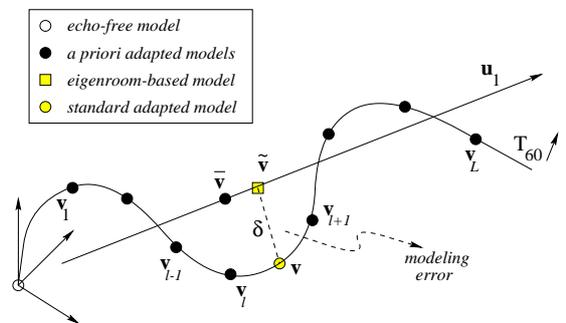where α denotes the learning rate coefficient. The gradient term

Figure 2: Eigenroom-based adaptation: the adaptation transformation is assumed to be modeled by only 3 parameters ($D = 3$) and constrained to lay on the first eigenroom ($K = 1$).

Table 1: Word error rate (WER) as the sum of substitution error rate (SUB), deletion error rate (DEL) and insertion error rate (INS) for the baseline speaker-independent echo-free MLP.

WER [%]	SUB [%]	DEL [%]	INS [%]
1.7	0.7	0.5	0.5

Table 2: Word error rate (WER [%]) for various reverberant environments (T_{60} [ms]) and for various amounts of adaptation data in the case of a full adaptation matrix.

Adapt. data [frame]	Test set				
	T_{60}=200	400	600	800	1000
No	WER=**8.2**	**20.4**	**33.4**	**46.7**	**48.5**
25k	5.5	8.8	14.8	24.3	27.4
50k	4.9	7.4	12.0	20.0	22.1
75k	3.9	6.7	10.0	17.9	19.5
100k	3.4	6.2	9.7	17.1	18.5
125k	3.4	6.0	9.1	15.9	17.1
150k	3.4	5.7	8.8	15.8	17.1
175k	3.5	5.7	8.5	15.3	16.3
200k	3.9	5.4	8.5	15.1	16.1

is computed by applying the chain rule for partial derivative:

$$\frac{\partial E}{\partial g_k^{(\ell)}} = \sum_{i=1}^{N} \frac{\partial E}{\partial y_i^{(\ell)}} \frac{\partial y_i^{(\ell)}}{\partial g_k^{(\ell)}} \qquad (8)$$

where the first term is obtained by the BP algorithm and the second term is easily derived from equation (6), i.e. equal to the output of the kth eigenroom transformation $(\mathbf{A}_k, \mathbf{b}_k)$.

3. Experimental Results

The speech material used in this work comes from the clean part of the AURORA database [12] and consists of English connected digit sequences. The corpus is divided into a training set of 8840 utterances and a test set of 1001 utterances, pronounced by 110 speakers and 104 other speakers, respectively.

First, we train a MLP with a 600-node hidden layer on the echo-free training set. The resulting model is assumed to be speaker-independent. For every speech frame, it estimates the *a posteriori* probabilities for a 33-phoneme set given the acoustic vectors of the current frame augmented with 7-frame left-context and 7-frame right-context acoustic vectors. Each acoustic vector is composed of 12 Mel-warped frequency cepstral coefficients (MFCC) and the Δenergy. The performance of the resulting MLP for recognition of the echo-free test set is given in table 1. Speech decoding is done by Viterbi search, with neither pruning nor grammar constraints.

Then, we try to adapt the echo-free MLP to various reverberant environments. Following the notation of section 2, the compensated vector $\mathbf{y}$ is obtained by linear transformation (see equation (1)) of the N-dimensional vector $\mathbf{x}$ formed by the current acoustic vector with its context, i.e. $N = (7+1+7) \times 13 = 195$. As described in section 2.2, no reverberated data are collected in the reverberant environment to which we want to adapt. Only the reverberation time T_{60} has to be known. Given T_{60}, one can generate adaptation material by convolving the echo-free training set with artificial room impulse responses matching T_{60} [2]. Once the echo-free MLP has been adapted, it is used to recognize a reverberated test set. In this work, the

Table 3: Word error rate (WER [%]) for various reverberant environments (T_{60} [ms]) and for various amounts of adaptation data in the case of a block diagonal adaptation matrix.

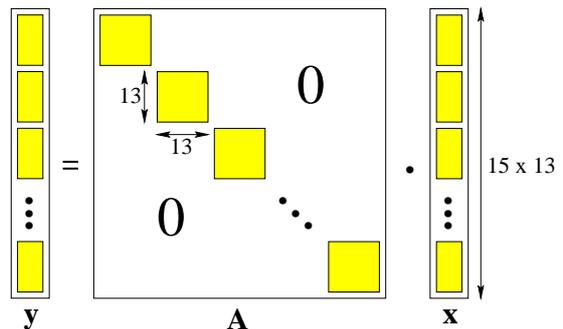| Adapt. data [frame]	Test set				
	T_{60}=200	400	600	800	1000
No	WER=**8.2**	**20.4**	**33.4**	**46.7**	**48.5**
25k	6.1	12.9	22.0	34.6	37.0
50k	5.9	12.2	19.6	30.8	33.9
75k	5.5	11.3	18.1	28.6	32.5
100k	5.5	11.0	17.0	28.4	31.0
125k	5.5	10.4	16.6	27.1	30.2
150k	5.2	10.2	16.7	26.9	29.6
175k	5.2	10.2	16.1	26.7	28.8
200k	5.2	10.3	16.1	26.1	28.5

Figure 3: Block diagonal adaptation matrix.

reverberated test sets are obtained by acoustic room simulation (Image method [13]) which allows us to specify any room configuration and control the reverberation time T_{60}. Table 2 shows the WER for various T_{60} as a function of the number of adaptation frames. The first line corresponds to the performance of the echo-free system, i.e. with no adaptation. As expected, adapting the acoustic model significantly improves the performance of the speech recognizer.

Though the standard adaptation technique provides high WER improvements, especially with large amounts of adaptation data, the number of parameters defining the adaptation transformation is too large ($D = N \times (N + 1) = O(10^4)$) for using it in a fast adaptation framework. Indeed, the computation of the eigenrooms would be highly memory demanding, computationally prohibitive and prone to round-off error. First, we observe that the biases do not help adapting. Besides, we observe that high value coefficients of the adaptation matrix are mostly located along the main diagonal. Hence, we decide to use a block diagonal adaptation matrix instead of a full matrix. That is, the elements which are off the main block diagonal are forced to zero (see figure 3). The number of adaptation parameters is reduced drastically ($D = 15 \times 13 \times 13 = O(10^3)$). For the sake of comparison, table 3 reports the WER for various T_{60} as a function of the number of adaptation frames. As expected, the adaptation procedure with a block diagonal matrix provides less WER improvement than with a full matrix.

Next, we test the eigenroom-based adaptation technique. As a first step, block diagonal adaptation matrices are generated for T_{60} varying from 100ms to 1200ms. For each T_{60}, 200000 frames of adaptation data are obtained by using artificially generated room impulse responses (see section 2.2). Then, the
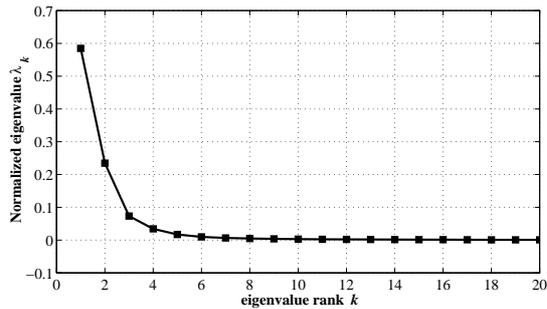
Figure 4: Scree plot for PCA applied to block diagonal matrices for adaptation to room reverberation (limited to first 20 eigenvalues).

adaptation vectors are formed and the principal directions of the resulting vector set are extracted. Figure 4 gives the scree plot of the resulting eigenvalues. It clearly shows that the first few principal directions ($k \leq 10$) account for most of the variability within the adaptation vector set. Finally, we apply the eigenroom-based adaptation approach with various number of eigenrooms. Figure 5 compares WER improvements relatively to the performance of the unadapted MLP: (a) for $T_{60} = 600$ms with the number of adaptation frames varying from 25000 to 200000, and (b) for 50000 adaptation frames with T_{60} varying from 200ms to 1000ms. The eigenroom-based adaptation procedure performs significantly better than the standard adaptation procedure, especially for low amounts of adaptation data.

4. Conclusion and Future Work

We have shown that HMM/MLP recognizers can be efficiently adapted to room reverberation by linear transformation of the input acoustic vectors. Besides, the *eigenroom* concept has been proposed. Similarly to the *eigenvoice* approach for fast speaker adaptation, the eigenroom-based approach requires less data for room adaptation than the standard approach. Unfortunately, the adaptation matrix has to be limited to a block diagonal matrix for computational reasons. Future work will focused on relaxing this constrain. For example, a structured full matrix like a FIR matrix might be used. Furthermore, the promising results obtained for supervised adaptation have to be confirmed in unsupervised adaptation mode.

5. References

[1] D. Giuliani, M. Matassoni, M. Omologo and P. Svaizer, "Training of HMM with Filtered Speech Material for Hands-free Recognition", *Proc. ICASSP'99*, vol. 1, pp. 449–452, Phoenix, USA, Mar. 1999.

[2] L. Couvreur, C. Couvreur and C. Ris, "A Corpus-Based Approach for Robust ASR in Reverberant Environments", *Proc. ICSLP'2000*, vol. 1, pp. 397–400, Beiging, China, Oct. 2000.

[3] H. Bourlard and N. Morgan, "Connectionist Speech Recognition – A Hybrid Approach", Kluwer Academic Publishers, 1994.

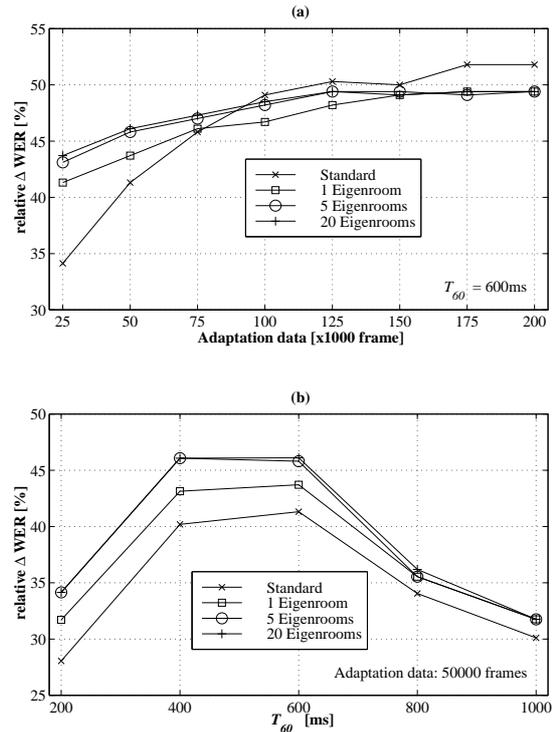[4] P. Nguyen, C. Wellekens and J.-C. Junqua, "Maximum Likelihood Eigenspace and MLLR for Speech Recognition

Figure 5: Relative word error rate improvement (Δ WER [%]) for eigenroom-based adapted MLP's with various numbers K of eigenrooms (a) as a function of the number of adaptation frames and (b) as a function of the reverberation time T_{60}.

in Noisy Environments", *Proc. EUROSPEECH'99*, vol. 6, pp. 2519–2522, Budapest, Hungary, Sep. 1999.

[5] R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space", *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[6] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field and M. Contolini, "Eigenvoices for Speaker Adaptation", *Proc. ICSLP'98*, vol. 5, pp. 1771–1774, Sydney, Australia, Dec. 1998.

[7] P. Nguyen, "Fast Speaker Adaptation", *Technical Report*, Eurécom Institute, Jun. 1998.

[8] S. Dupont and L. Cheboub, "Fast Speaker Adaptation of Artificial Neural Networks for Automatic Speech Recognition", *Proc. ICASSP'2000*, vol. 3, pp. 1795–1798, Istanbul, Turkey, Jun. 2000.

[9] S. Haykin, "Neural Networks: A Comprehensive Foundation", McMillan, 1994.

[10] J. Neto, C. Martins and L. Almeida, "Speaker-Adaptation in a Hybrid HMM-MLP Recognizer", *Proc. ICASSP'96*, vol. 6, pp. 3383–3386, Atlanta, USA, May 1996.

[11] K. Fukunaga, "Introduction to Statistical Pattern Recognition", Academic Press, 1990.

[12] AURORA database - http://www.elda.fr/aurora2.html.

[13] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics", *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.