FACTORED ADAPTATION OF SPEAKER AND ENVIRONMENT USING ORTHOGONAL SUBSPACE TRANSFORMS

Hyunson Seo[*] *, Hong-Goo Kang*

Yonsei University
DSP Lab, Department of E.E.
Seoul, South Korea

Michael L. Seltzer

Microsoft Research
Redmond, WA 98052 USA

ABSTRACT

This paper presents a subspace-based *acoustic factorization* framework to transform-based adaptation in speech recognition. In the proposed method, adaptation transforms are projected onto factor-dependent low-rank subspaces in a way that decouples the combined extrinsic factors affecting the speech signals. Usually, mismatch between the observed speech and the acoustic models is caused by multiple acoustic factors simultaneously, such as the speaker and environment. Data-driven adaptation methods, such as constrained MLLR, compensate for all sources of mismatch jointly. In many scenarios, however, it is highly desirable to separate the sources of mismatch in order to adapt to speaker and environment variability independently. This adds flexibility to the model adaptation framework. For example, a speaker transform obtained in one environment can be reused when the same speaker is in different environments. Or, an environment transform obtained during training, independently of speaker identities, can be applied to a speaker in deployment. One way to achieve this factorization is to construct each set of transforms such that they are orthogonal to each other, so that any change in one acoustic factor keeps other factors intact. The proposed subspace approach provides a straightforward factor analysis framework while allows us to explicitly formulate the independence among the estimated factor transforms. A series of experiments performed on the Aurora 4 corpus validates our approach.

Index Terms— acoustic factorization, speaker and environment adaptation, orthogonal subspace projection, subspace transforms

1. INTRODUCTION

The performance of automatic speech recognition (ASR) systems degrades as a result of acoustic mismatch caused by extrinsic variabilities, such as speaker characteristics and environment differences. Assuming that a canonical model represents the intrinsic phonetic variability of speech signals, model-based adaptation approaches in conventional recognizers introduce a set of transforms to compensate the possible extrinsic variabilities. Because speech signals are typically affected by multiple acoustic factors simultaneously, the combined effect of these factors is modeled by a single transform. In many scenarios, however, the ability to adapt the recognizer to each source of variability independently is highly desirable. For example, consider a speaker in a range of different environments. If it is possible, recycling the speaker transform obtained in one noise condition even if the speaker's environment later changes increases the system's efficiency and reduces the required adaptation data. Alternatively, an environment transform can be applied across different users in the same environment.

The underlying concept of *acoustic factorization* which separates such combined acoustic factors using a set of transforms was first proposed by Gales in [1] and has been further developed using different methods [2–8]. Common to all of these methods is the notion that the set of transforms are constructed such that each transform is related to only one acoustic factor. In [3], two constrained maximum likelihood linear regression (CMLLR) transforms were cascaded to represent the speaker and environment variability respectively. They were jointly estimated in an iterative expectation-maximization (EM) framework by alternating the target transform to be optimized. In [5], each set of transforms were developed in different domains so that the environmental effects were compensated in the model domain using MLLR transform while the speaker attributes were adapted by CMLLR transform in the feature domain. Combining entirely different adaptation strategies for speaker and noise compensation was also proposed, in the hope that each set of transforms models the specific factor independently to some extent. In [9], Wang and Gales used vector Taylor series (VTS) for environment compensation and MLLR transforms for speaker adaptation. More recently, the independence between factorized transforms was enforced using an explicit mathematical constraint [8].

In this work, we present a subspace-based acoustic factorization approach which provides a straightforward factor analysis framework which allows us to explicitly formulate the independence among the estimated subspace transforms. It is assumed that each source of variability is contained within a corresponding low-rank subspace so that the latent factor measured in the specified subspace can quantify the decoupled speaker or environment characteristics. This approach is motivated by joint factor analysis [10], where a speaker- and channel-dependent supervector is decomposed into a sum of two supervectors: a speaker supervector and a channel supervector. We efficiently integrate this concept onto eigenspace-based MLLR (EMLLR) framework [11] so that the transforms estimated in a complex acoustic environment are separated into different sets of factor-dependent transforms. In contrast to JFA, the key idea of the proposed approach is that the speaker and environment subspaces are constructed to be orthogonal so that the factored transforms lying on different subspaces are forced to be independent.

In Section 3.1, we review the concept of acoustic factorization. In Section 3, the proposed subspace-based acoustic factorization framework is presented with the subspace modeling approach based on orthogonal subspace projection. Then, experiments and results are presented and discussed in Section 4 with conclusions in Section 5.

[*] Part of this work was performed while the first author worked as an intern at Microsoft Research, Redmond.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. ACOUSTIC FACTORIZATION

The most common form of acoustic factorization framework assumes that there are two acoustic factors, speaker characteristics s and environment differences e, that simultaneously affect the speech signal. In a conventional model adaptation framework, the canonical model is adapted by the transform $W^{(se)}$ which models the combined effect of s and e such as [7]

$$\Lambda^{(se)} = F\left(\Lambda_c, W^{(se)}\right), \tag{1}$$

where $\Lambda^{(se)}$ is the adapted acoustic model for condition (s, e) and F is a mapping function. Normally, the transform $W^{(se)}$ is estimated using maximum likelihood (ML) criterion

$$W^{(se)} = \underset{W}{\text{argmax}} \left\{p\left(O^{(se)}|\Lambda_c, W\right)\right\}, \tag{2}$$

where $O^{(se)}$ is a sequence of feature vectors observed in the acoustic condition (s, e).

To effectively deal with the complex acoustic environments, the acoustic factorization framework proposes to factorize a single transform $W^{(se)}$ into two components, each associated with one distinct acoustic factor, i.e.,

$$W^{(se)} = W^{(s)} \times W^{(e)}, \tag{3}$$

where $W^{(s)}$ and $W^{(e)}$ are the transforms associated with the factors s and e, respectively. This attribute offers additional flexibility for the models to be used in complex environments. For example, consider a speaker s in a range of n acoustic conditions, $(s, e_1), ..., (s, e_n)$. In a conventional model adaptation framework, it is necessary to estimate a set of transforms $W^{(se_1)}, ..., W^{(se_n)}$ using the data $O^{(se_1)}, ..., O^{(se_n)}$ from each of these conditions. With the factorization, however, only a single speaker transform $W^{(s)}$ and a set of environment transforms $W^{(e_1)}, ..., W^{(e_n)}$ are required, and the speaker transform can be reused. The environment transforms can be potentially be estimated during a training phase, or from different users' utterances recorded in the same environment. Therefore, theoretically, to achieve acoustic factorization, it is crucial to keep factor transforms independent of each other. Most previous work relies on using different forms of factored transforms and/or EM-based optimization schemes to enforce this independence. In recent work by *Wang*, an explicit constraint for the independence between sets of transforms is formulated [8].

3. ACOUSTIC FACTORIZATION IN SUBSPACE

In this work, we propose a subspace-based approach for acoustic factorization which provides straightforward factor analysis framework and allows an explicit formulation for the independence of each factor. It is assumed that each source of variability is contained within a corresponding low-rank subspace and the latent factor measured in the specified subspace can represent the decoupled speaker or environment characteristics. Theoretically, it is motivated from joint factor analysis [10] model of speaker and channel variability in speech recognition system. Though an equivalent technique may naturally be created using the eigenvoice framework [12] – both algorithms represent a speaker by a supervector that is composed by mean vectors of the GMMs/HMMs, we efficiently extend the concept to eigenspace-based MLLR (EMLLR) adaptation framework [11]. EMLLR approaches have been successfully applied to large vocabulary continuous speech recognition [11, 13].

3.1. Subspace Representation of Linear Transforms

Suppose we have a speech corpus consisting of various speakers in many recording environments, and a set of transforms associated with each speaker/environment condition that are estimated from a canonical model using a linear transform such as MLLR or constrained MLLR [14]. The typical size of such matrices is $d \times (d+1)$, where d represents the speech feature dimension. We assume that each speaker/environment condition is indirectly represented by the transformation matrix. To simplify the notation, the columns of this transformation matrix are stacked into a single vector $\mathbf{w}$ with the dimension of $D = d(d+1)$. The training corpus, then, is represented by a matrix $\mathbf{E}$ whose columns are transform supervectors $\{\mathbf{w}_i\}$ that are estimated from the complete set of training data.

To begin with, let us assume that there is a single acoustic factor that affects the speech signal, i.e., either speaker characteristics or environment differences. In this case, the problem reduces to finding a single low-rank subspace which can represent the corresponding latent factor such as

$$\mathbf{w} \approx \bar{\mathbf{w}} + \mathbf{U}\mathbf{x}, \tag{4}$$

where $\mathbf{U}$ is a $D \times r$ ($r \ll D$) matrix whose columns represent the principle directions of variability in the data which can be estimated by applying principle component analysis (PCA). The r-dimensional vector $\mathbf{x}$ represents the factor-dependent parameter in the estimated subspace, and $\bar{\mathbf{w}}$ the offset mean supervector of the entire training population. If we consider the latent factor to be speaker variability, then this becomes equivalent to the EMLLR approach in [11], where the transform $\mathbf{w}$ in (4) indirectly represents a speaker and, any other extrinsic variabilities, such as environment variability, are not considered.

In order to incorproate the impact of both environmental variability and speaker attributes, we assume that there are two distinct sets of subspaces, each related to only one latent factor as follows:

$$\begin{aligned}\mathbf{w}^{(se)} &\approx \bar{\mathbf{w}} + [\mathbf{U}\,\mathbf{V}]\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \quad s.t.\ \ \mathbf{U}\perp\mathbf{V} \\ &= \bar{\mathbf{w}} + \sum_{i=1}^{r_s}\mathbf{u}_i x_i + \sum_{j=1}^{r_e}\mathbf{v}_j y_j \\ &= \bar{\mathbf{w}} + \mathbf{w}^{(s)} + \mathbf{w}^{(e)},\end{aligned} \tag{5}$$

where $\mathbf{U}$ ($D \times r_s$) represents a low-rank matrix related to speaker variability and $\mathbf{V}$ ($D \times r_e$) relates to environment variability. The subspace specific weight vectors, $\mathbf{x}$ and $\mathbf{y}$, quantify the amount of impact from speaker and environment, respectively. Thus, $\mathbf{w}^{(s)}$ can be regarded as the speaker factor transform and $\mathbf{w}^{(e)}$ be the environment factor transform. Notice that compared to (3), the proposed subspace-based factorization regards the transform $W^{(se)}$ as the *sum* of speaker- and environment-dependent transforms. This form of decomposition is similar to that of the joint factor analysis [10] in the speaker recognition field. In both frameworks, it is assumed that each factor plays a different and independent role, i.e., the speaker factors are constant for all speakers while the environment factors can vary depending on the recording environment. Notice that, however, while JFA seeks to separate the speaker and the session variability via subspace analysis, it has no structural way to guarantee their two subspaces orthogonal. Rather, they rely on the data balance to factor out the speaker variability.

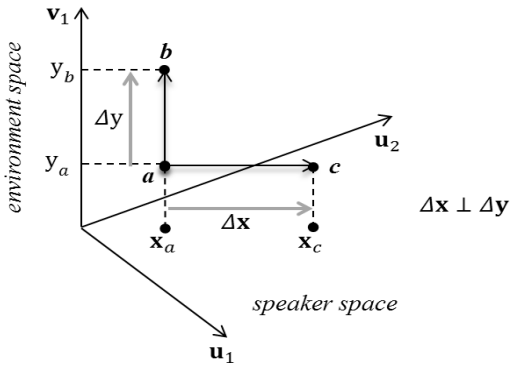

Fig. 1. Factorized adaptation in subspaces. Speaker subspace and environment subspace is orthogonal: $\{\mathbf{u}_1, \mathbf{u}_2\} \perp \{\mathbf{v}_1\}$

3.2. Factored Adaptation Using Subspace Transforms

We utilize the proposed orthogonal subspace model in (5) for separately adapting the speech recognizer to a specific speaker or/and environment independently. The basic idea is illustrated in Fig.1, where the speaker subspace is realized in $\mathbb{R}^2$ and environment subspace in $\mathbb{R}^1$ for visualization. In the figure, the impact of speaker and environment attributes modeled in a transform is represented by the vector $(\mathbf{x}, \mathbf{y})$. For example, the transform for acoustic condition ***a*** is represented by the point $(\mathbf{x}_a, \mathbf{y}_a)$. Suppose that we want to adapt the model to another target condition ***b***. In this framework, the only requirement to reflect the environment transition is an update of $\mathbf{y}_a$, which is $\Delta\mathbf{y}$, while the speaker factor $\mathbf{x}_a$ is fixed. This is possible due to the orthogonality between the speaker and environment subspaces, i.e., $\mathbf{U} \perp \mathbf{V}$. Ideally, any changes of one factor in the specified subspace are assumed to have no impact on the other factor. Similarly, to adapt the model to another operating point ***c***, the speaker factor $\mathbf{x}_c$ needs to be updated while the environment factor $\mathbf{y}_a$ is fixed. From the speaker adaptation point of view, it enables a more robust speaker transform to be obtained in terms of environment immunity. By projecting the transform onto the speaker subspace, we can remove the noise factors specified for a particular recording environment from the transform. It allows fast speaker adaptation in rapidly changing acoustic environments by reusing the speaker information obtained in a particular environment even if the speaker's environment later changes.

3.3. Subspace Modeling via Orthogonal Subspace Projection

The first step in deriving the speaker subspace is to eliminate the impact of the interfering environmental variability, which is represented by the columns of $\mathbf{V}$. The approach is to form an operator that projects each data point in training data set, $\mathbf{E}$, onto a subspace that is orthogonal to the columns of $\mathbf{V}$. The vector resulting from such an operation is ideally more resistant to environment factors. In the least squares sense, the optimal interference rejection operator is given by the $D \times D$ matrix $\mathbf{P}$ as follows:

$$\mathbf{P} = \mathbf{I} - \mathbf{V}\mathbf{V}^{\dagger}\,, \tag{6}$$

where $\mathbf{V}$ is a low-rank matrix whose orthonormal columns define the dimensions to be removed from the space $\mathbf{E}$. In this work, those dimensions are related to environment variability. $\mathbf{V}^{\dagger} = (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T$ is the pseudo-inverse of $\mathbf{V}$, where $\mathbf{V}^{\dagger} = \mathbf{V}\mathbf{V}^T$ if ${\mathbf{v}_i}^T\mathbf{v}_j = 0$, $\mathbf{v}_i \in \mathbf{V}$, $i, j = 1, .., r_e$. This operator has the same structure as the orthogonal complement projector from the theory of least squares [15].

Operating the orthogonal projector $\mathbf{P}$ on (5), we have

$$\begin{aligned}\mathbf{P}\mathbf{w} &= \mathbf{P}\bar{\mathbf{w}} + \mathbf{P}\mathbf{U}\mathbf{x} + \mathbf{P}\mathbf{V}\mathbf{y} \\ &= \mathbf{P}\bar{\mathbf{w}} + \mathbf{U}\mathbf{x},\end{aligned} \tag{7}$$

since $\mathbf{U} \perp \mathbf{V}$, $\mathbf{V}^{\mathbf{T}}\mathbf{V} = \mathbf{I}$. It is clear that, for the purposes of acoustic factorization, this approach optimally rejects the interfering environment attributes from $\mathbf{w}^{(se)}$, in that $\mathbf{P}$ reduces the contribution of $\mathbf{V}\mathbf{y}$ to zero while keeping the desired components of $\mathbf{U}\mathbf{x}$ intact. Thus, in the subspace that is orthogonal to the columns of $\mathbf{V}$, the speaker subspace $\mathbf{U}$ can be readily derived using PCA with the correlation matrix of $\{\mathbf{P}\mathbf{w}_i | \mathbf{w}_i \in \mathbf{E}\}$. It derives a set eigenvectors, $\mathbf{u}_n \in \mathbb{R}^D$, part of which comprises column vectors of $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_{r_s}]$ to span the speaker subspace.

The second step is then to find the matrix $\mathbf{V}$. Our approach to obtain the orthogonal projection matrix $\mathbf{P}$ and the corresponding low-rank matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{r_e}]$ is borrowed from the idea of Nuisance Attribute Projection (NAP) [16]. Though it was originally designed to develop a modified kernel matrix for a support vector machine (SVM), without loss of generality, we apply the basic concept to the proposed subspace modeling framework such as

$$\mathbf{v}^* = \underset{\mathbf{v}, \|\mathbf{v}\|_2 = 1}{\arg\min} \sum_{i,j} \mathbf{A}\, \|\mathbf{P}(\mathbf{w}_i - \mathbf{w}_j)\|_2^2\,, \tag{8}$$

where $\mathbf{w}_i$ and $\mathbf{w}_j$ represent any pair of transforms in a background dataset $\mathbf{E}$. $\mathbf{A}$ is a symmetric matrix consisting of weight parameters a_{ij} to efficiently minimize the average distance of transform pairs $(\mathbf{w}_i, \mathbf{w}_j)$ in the projection space, where a_{ij} can be selected in several different ways [17]. In this paper, we set $a_{ij} = 1$ for the transforms whose speaker identities are the same, and $a_{ij} = 0$ otherwise. It intends to minimize the average distance of cross-environment samples. Specifically, by removing the subspace defined by $\mathbf{v}^*$, any transform pairs of (s, e_1) and (s, e_2), $e_1 \neq e_2$, are pulled together in the projection space. Thus, the vector set $\{\mathbf{v}^*\}$ can be naturally regarded as column vectors spanning the environment subspace. A somewhat lengthy calculation shows that the $\mathbf{v}^*$ is obtained from the generalized eigenvalue problem $\mathbf{K}\mathbf{Z}\mathbf{K}\mathbf{v} = \lambda\mathbf{K}\mathbf{v}$, where $\mathbf{K} = (\mathbf{P}\mathbf{E})^T(\mathbf{P}\mathbf{E})$ and $\mathbf{Z} = diag(\mathbf{V}\mathbf{1}) - \mathbf{V}$. The detailed derivation can be found in [18].

4. EXPERIMENTS AND ANALYSIS

The proposed subspace-based acoustic factorization framework was evaluated on the Aurora 4 corpus [19], which consists of the Wall Street Journal (WSJ0) 5k-word corpus degraded by six types of noise - car, babble, restaurant, street, airport and train. SNRs in the training set range from 10 to 20 dB and from 5 to 15 dB for test set. There are 7,138 utterances in the training set, produced by 83 speakers. The evaluation set includes 7 subsets grouped by noise type, regardless of the SNR, each consisting of 330 utterances produced from eight speakers. Standard 39-dimensional MFCC features consisting of 13 static, first and second order dynamic features including C0 were used with cepstral mean normalization. Cross-word triphone models with 6916 distinct tied-states and 16 components per state were used for acoustic modeling and the standard bi-gram language model provided for the Aurora 4 evaluation were used in decoding. To estimate the low-rank matrices, $\mathbf{U}$ and $\mathbf{V}$, for subspace construction, the same training set was used. As an initial investigation, they were trained using both speaker and environment labels. Each of training

Table 1. Word accuracy (%) of three adaptation schemes: *None, CMLLR, batch*. Speaker transform (*CMLLR*) was estimated in a noisy *restaurant* condition and applied to remaining 6 environments.

scheme	clean	car	babble	street	airport	train	Avg.
Baseline	92.5	90.4	84.0	79.7	85.9	79.5	85.3
CMLLR	81.5	85.2	82.5	80.8	84.0	77.0	81.8
Batch	94.5	92.7	87.8	83.2	88.6	83.9	88.5

utterances was clustered as one of 7 environment classes depending on the noise type in advance. In this work, the constrained maximum linear regression (CMLLR) method [14] was used to evaluate the effectiveness of the proposed subspace model. For subspace models, $r_s = 30$ for $\mathbf{U}$ and $r_e = 6$ for $\mathbf{V}$ were used for evaluation.

The experiments simulated a practical enrollment scenario in which adaptation utterances were collected in a single noise condition (restaurant) while the test environment varied over all different 7 noise conditions. For speaker enrollment, 10 utterances were used for each speaker to adapt the speaker independent (SI) model to the speaker in a unsupervised mode. A global CMLLR transform was used, and unadapted decoding was first performed using the SI model trained on the multi-condition training data. Table 1 shows the word accuracy for 6 other environment conditions. The results were compared to that of system *Baseline*, which was decoded without speaker adaptation. As expected, the speaker transform estimated in a specific environment, and applied to different environments, degraded the accuracy on all 6 environments. The average accuracy dropped to 81.8% from the baseline accuracy of 85.3% . This reflects the fact that the speaker transforms estimated in noisy environments model both of speaker and the environment variability and thus, their effectiveness is not guaranteed when the environment changes. As an upper bound on performance, we aevaluated the *batch* mode CMLLR adaptation in which each test utterances were decoded using the exact speaker/environment transform. It assumed that we had adaptation data for all operating conditions and learned transforms for each speaker/environemnt combination. By applying the proposed factored adaptation method, we expect to achieve a competitive result to that of the *batch* scheme while not requiring the complete set of adaptation data and transforms.

The same experiments were repeated using the proposed subspace transforms in an acoustic factorization framework. First, we demonstrated the effectiveness of the speaker subspace transform, *SSCMLLR-spk*. It was estimated by projecting the *CMLLR* transforms onto the speaker subspace so that it excluded the impact of environment while keeping the most of speaker information intact. The results are shown in table 2. *SSCMLLR-spk* achieved an accuracy of 86.4% on average, which is a 7.5% reduction of word error rate compared to the baseline (85.3%). Notice that, this improvement in accuracy required neither additional adaptation data nor estimating a new transform. The speaker information which was successfully decoupled from the enrollment data was recycled to adapt the model. Compared to the conventional scheme of *CMLLR*, in the proposed *SSCMLLR-spk* framework, the subspace expanded by the speaker factors was orthogonal to the subspace expanded by the environment factors. It allowed speaker transforms learned on noisy adaptation data to explain only the speaker characteristics, making it applicable regardless of environment. A lowered WER in the clean condition indicates that, although the speaker and environment factors can be decoupled to some extent, the speaker factor derived from noisy environments may have some limitations to fully reflect the speaker characteristics in clean condition.

Table 2. Word accuracy (%) of the proposed acoustic *factorization framework*. The speaker transform *SSCMLLR-spkr* was estimated in a noisy (restaurant) condition and applied to remaining 6 environments. The environment transforms in *SSCMLLR-(spkr+envr)* were obtained independently of speaker identity.

scheme	clean	car	babble	street	airport	train	Avg.
SSCMLLR-spkr	91.4	91.2	85.7	82.1	86.6	81.4	86.4
SSCMLLR-(spkr+envr)	93.4	91.7	86.7	82.6	87.3	82.8	87.4

To jointly compensate the speaker and environment variabilities in the proposed acoustic factorization framework, the speaker transform estimated in the previous experiment was used in conjunction with an environment transform estimated during the training phase, independently of speaker identities. Table 2 shows the results. The recognition accuracy in the unseen environments improved to 87.4%, which is a 14.3% relative improvement from that of the baseline scheme, and fairly close to the upper bound two-pass (*batch*) performance of 88.5%. Notice that without acquiring adaptation data from all combinations of speaker and environment operating conditions, the proposed factored adaptation framework obtained performance competitive with the *batch* mode system. The benefit of this *SSCMLLR-(spkr+envr)* system is more significant across complex acoustic conditions. It provides a more efficient adaptation process by controlling each source of variability rather than covering all the possible combinations of such factors. The system can recycle the speaker transform obtained in one noise condition even if the speaker's environment later changes. Moreover, it can import an environment transform which was obtained by different users in the same environment, which allows the system to synthesize target transforms for many possible operating conditions.

5. CONCLUSION

In this paper, we proposed a subspace-based *acoustic factorization* framework, which enables the speech recognizer to adapt to specific speaker or/and environment variability separately. The technique to approximate transforms in factor-dependent low-rank subspaces enabled each of transforms lying on different subspaces to relate to different sources of variability. It allowed efficient and fast speaker adaptation in noisy environments by reusing the speaker transform estimated in one environment even if the speaker's environment changed. Environment transforms, estimated from data from a number of speakers, were combined with any speaker transforms to enable the recognizer to operate beyond the environment seen in the adaptation data. This was possible as a result of the independence among the estimated factor-dependent subspace transforms.

6. REFERENCES

[1] M. J. F. Gales, "Acoustic factorisation," in *Proc. ASRU*, 2001.

[2] L. Rigazio, P. Nguyen, D. Kryze, and J.-C. Junqua, "Separating speaker and environmental variabilities for improved recognition in non-stationary conditions," in *Proc. Eurospeech*, 2001, pp. 1792–1795.

[3] M.L. Seltzer and A. Acero, "Separating speaker and environmental variability using factored transforms," in *Proc. Interspeech*, 2011, pp. 1097–1100.

[4] M.L. Seltzer and A. Acero, "Factored adaptation for separable compensation of speaker and environmental variability," in *Proc. ASRU*, 2011.

[5] M.L. Seltzer and A. Acero, "Factored adaptation using a combination of feature-space and model-space transforms," in *Proc. Interspeech*, 2012, pp. 1792–1795.

[6] Y. Q. Wang and M. J. F. Gales, "Speaker and noise factorisation on the Aurora 4 task," in *Proc. ICASSP*, 2011, pp. 4584–4587.

[7] Y. Q. Wang and M. J. F. Gales, "Speaker and noise factorization for robust speech recognition," *IEEE Trans. on Audio Speech and Language Processing*, vol. 20, pp. 2149–2158, July 2012.

[8] YQ Wang and MJF Gales, "An explicit independence constraint for factorised adaptation in speech recognition," 2013.

[9] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector taylor series for noisy speech recognition," in *Proc. ICSLP*, 2000.

[10] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Audio Speech and Language Processing*, vol. 15, 2007.

[11] Kuan-ting Chen, Wen-wei Liau, Hsin-min Wang, and Lin-shan Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression.," in *Proc. Interspeech*, 2000, pp. 742–745.

[12] Roland Kuhn, J-C Junqua, Patrick Nguyen, and Nancy Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[13] X. D. Cui, J. Xue, and B. Zhou, "Improving online incremental speaker adaptation with eigen feature space MLLR," in *Proc. ASRU*, 2009, pp. 136–140.

[14] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov modelsmarkov models," in *Computer Speech and Language*, 1985, vol. 9, pp. 171–185.

[15] Simon Haykin et al., "Adaptive filtering theory," *Englewood Cliffs, NJ: Prentice-Hall*, 1996.

[16] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006.

[17] Hyunson Seo, Chi-Sang Jung, and Hong-Goo Kang, "Robust session variability compensation for SVM speaker verification," *IEEE Trans. on Audio Speech and Language Processing*, vol. 19, pp. 1631–1641, 2011.

[18] Alex Solomonoff, Carl Quillen, and William M Campbell, "Channel compensation for svm speaker recognition," in *Proc. Odyssey, Speaker and Language Recognition Workshop*, 2004, pp. 57–62.

[19] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," in *Inst. for Signal and Information Process*. Mississippi State Univ.