



Joint Environment and Speaker Normalization using Factored Front-End CMLLR

Shakti Rath, Sunil Sivasdas and Bin Ma

Human Language Technology Department
Institute for Infocomm Research (I2R), Singapore

{shaktir, sivadass, mabin}@i2r.a-star.edu.sg

Abstract

The problem of joint compensation of environment and speaker variabilities is addressed. A factored feature-space transform, named factored front-end CMLLR (F-FE-CMLLR), is investigated, which comprises of the cascade of two transforms – front-end CMLLR for environment normalization and CMLLR for speaker normalization. In this paper, we propose an iterative estimation algorithm for F-FE-CMLLR. We believe that the iterative estimation helps to decouple the effect of the two acoustic factors, allowing each transform to learn the effect of only factor, thereby yielding an improvement in speech recognition performance compared to sequential estimation. However, it is noted that the estimation of environment transform yields full co-variance Gaussians in the GMM-HMM, which makes direct estimation computationally expensive. An efficient training algorithm is presented that helps to reduce the computational cost considerably. Further, it is shown that a row-by-row optimization procedure can be employed, which makes the algorithm more efficient and attractive. On the multi-condition Aurora 4 task and discriminatively trained GMM-HMM, it is shown that F-FE-CMLLR yields 11.6% and 8.7% relative improvements on two evaluation sets over the baseline features that is processed only by CMLLR for speaker normalization.

Index Terms: Front-End CMLLR, Acoustic Factorization

1. Introduction

Speech recorded in real life automatic speech recognition (ASR) scenario is subject to distortion caused by the influence of different acoustic conditions, which leads to degradation in the performance. Two dominating sources responsible for such distortion are the speaker and environment factors. In situations when the recognizer is expected to be used under diverse environment conditions and by a large number of users, it becomes essential to make the recognizer adaptive to above two factors.

Amongst several existing methods, many consider adaptation to either of the above two factors separately, while ignoring the effect of the other factor [1, 2, 3]. In recent years, increasing research effort is devoted to finding ways that consider both factors simultaneously. One of the approaches in this direction is to combine various schemes developed for speaker and environment adaptation, and to tune each using selective data from the individual factor. This yields additional performance gain by enabling compensation of both factors. In a recent work Gales et al. [4] investigate a factored feature-space transform for communication link and speaker normalization.

A part of the work reported in this paper was done when the first author was at Cambridge University Engineering Department, UK. He would like thank Prof. M. J. F. Gales for his advise.

It uses front-end CMLLR (FE-CMLLR)¹ [7, 8, 9] for link and CMLLR [10] for speaker normalization. The observations vectors are processed by FE-CMLLR first, followed by CMLLR. Both transforms are estimated sequentially, that is one after the other. Xiao et al. [11] propose to use additive correction vectors to compensate the features. The correction terms are estimated for each speaker and noise combination.

A more principled strategy for joint adaptation is *acoustic factorization*, first introduced in [12]. The general idea in this scheme is to devise a mechanism to separate the effect of acoustic factors affecting speech and model them separately using a set of transforms, using only one transform for a factor. The advantage with such factorization is that the transform associated with a factor would remain relatively free from the influence of the other factors. This allows reuse of an environment transform, trained across one set of speaker transforms, in conjunction with a different set of speakers, and vice versa. With this objective, an explicit orthogonality condition between the transforms is formulated [13]. In the same spirit [14] proposes a joint factor analysis framework to enforce the orthogonality among the transforms. Combining adaptation strategies of different types have also been investigated [15, 16]. The other work that extends acoustic factorization are [17, 18, 19].

In this paper we present an *iterative* estimation algorithm for the factored transform proposed in [4] (for brevity, it is referred to as Factored FE-CMLLR, F-FE-CMLLR, in this paper). Specifically, the environment and speaker transforms are estimated iteratively in an interleaved manner. We believe that the iterative estimation may provide a mechanism to approximate acoustic factorization. However, this is not the case with sequential estimation, which is non-iterative over the two transforms. The algorithm is evaluated on the multi-condition Aurora 4 task, applying joint environment and speaker adaptive training. In this setup, environment normalization takes place in the supervised mode, i.e., the transform is estimated from the training set in conjunction with the speakers appearing in the set, and is reused over those in the test set. In this scenario, therefore, if acoustic factorization holds the impact of environment normalization is expected to be more effective than sequential estimation, yielding a better recognition performance.

Further, it is noted that the proposed iterative estimation yields full co-variance Gaussians in the GMM-HMM during the estimation of front-end CMLLR. This makes the direct training computationally very expensive. An efficient alternative is presented that helps to reduce the computational cost considerably. Moreover, it is shown that a row-by-row optimization procedure (similar to the one used in standard CMLLR) can be employed,

¹The transforms used in SPLICE [5] and FMPE [6] are similar to FE-CMLLR, but use only bias terms for feature compensation.

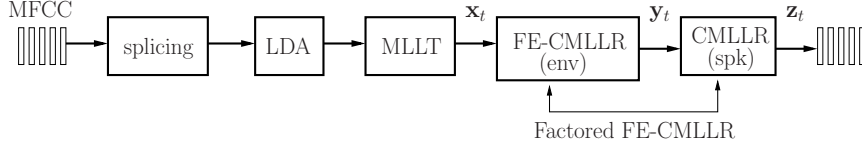


Figure 1: Block diagram of Factored Front-End CMLLR

which makes the algorithm more efficient and attractive.

The rest of the paper is organized as follows. In Section 2 the F-FE-CMLLR transformation is introduced. The estimation method is presented in Section 3, outlining the proposed iterative training procedure. The experimental results are presented in Section 4. Finally, we conclude in Section 5.

2. Factored Front-End CMLLR

The factored transform, F-FE-CMLLR, investigated in this paper is given by [4]:

$$\mathbf{z}_t = \mathcal{F}(\mathbf{x}_t) = \mathcal{F}^{(s)}(\mathcal{F}^{(e)}(\mathbf{x}_t)) \quad (1)$$

where $\mathcal{F}^{(e)}$ and $\mathcal{F}^{(s)}$ denote the environment and speaker transforms respectively, \mathbf{x}_t denotes the speaker and environment independent observation vector and \mathbf{z}_t denotes the corresponding normalized vector. The feature pipeline is shown in Figure 1. The environment FE-CMLLR is defined as:

$$\mathbf{y}_t = \mathcal{F}^{(e)}(\mathbf{x}_t) = \sum_{c=1}^C p(g_c | \mathbf{x}_t) \mathbf{W}_c^{(e)} \mathbf{x}_t^+ \quad (2)$$

The superscripted \mathbf{x}_t^+ indicates the extended observation vector: $\mathbf{x}_t^{+T} = [\mathbf{x}_t^T; 1]^T$. The parameters of $\mathcal{F}^{(e)}$ comprises of C affine transforms:

$$\mathbf{T}^{(e)} = \{\mathbf{W}_c^{(e)} = [\mathbf{A}_c^{(e)}; \mathbf{b}_c^{(e)}]\}_{c=1}^C, e = 1 \dots E, \quad (3)$$

and a front-end GMM. $p(g_c | \mathbf{x}_t^{(s)})$ denotes the posterior probability of the Gaussian component g_c in the front-end GMM. For every observation vector, the affine transforms are interpolated using the Gaussian posteriors and the resulting transform is applied on the observation vector to yield environment normalized features. Afterwards speaker transform, in this work global CMLLR, is applied to the above feature to produce the environment and speaker normalized features:

$$\mathbf{z}_t = \mathcal{F}^{(s)}(\mathbf{y}_t) = \mathbf{W}^{(s)} \mathbf{y}_t^+ \quad (4)$$

The parameters of the speaker transforms are $\mathbf{T}^{(s)} = \mathbf{W}^{(s)} = [\mathbf{A}^{(s)}; \mathbf{b}^{(s)}]$, $s = 1 \dots S$.

F-FE-CMLLR has several useful attributes. Firstly, the constituent FE-CMLLR is a highly non-linear operation on the observations that comprises of a large number of affine transforms. Hence, it may lend adept normalization capability to environment normalization. In addition, FE-CMLLR yields a consistent feature-space that helps to make speaker normalization more effective [4]. Moreover, being a feature-space operation, adaptive training becomes straightforward.

3. Iterative Estimation

There are three sets of parameters to be estimated - the speaker specific global CMLLRs, the environment FE-CMLLR and the

GMM-HMM parameters (i.e., adaptive training). The joint expectation maximization (EM) auxiliary function for maximum likelihood (ML) training is:

$$\mathcal{Q}(\mathbf{T}^{(e)}, \mathbf{T}^{(s)}, \mathbf{M}) = \sum_{m,t} \gamma_m(t) \log p(\mathbf{z}_t; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (5)$$

where $\log p(\mathbf{z}_t; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ is the log-likelihood for component m in the GMM-HMM. The estimation of the three set of parameters is achieved iteratively, estimating one set of parameter while fixing the other two. In the following sections, a bar over the parameter indicates the value yielded in previous iteration.

3.1. Speaker Transform

The auxiliary function for speaker transform for a fixed environment transform and model is:

$$\mathcal{Q}(\mathbf{T}^{(s)} | \bar{\mathbf{T}}^{(e)}, \bar{\mathbf{M}}) = \sum_{m,t} \gamma_m(t) \log p(\mathbf{z}_t^{(s)}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (6)$$

where $\mathbf{z}_t^{(s)} = \mathcal{F}(\mathbf{x}_t^{(s)})$. The superscript in $\mathbf{x}_t^{(s)}$ indicates the data from speaker s and the accumulation is done over all observations from the speaker. The transform parameters can be optimized using the formulae used for standard CMLLR [10].

3.2. Environment Transform

In this section the formulae for estimation of environment transform are presented; the general case is considered where the training data for an environment covers multiple speakers. The auxiliary function for fixed speaker transforms and model is:

$$\mathcal{Q}(\mathbf{T}^{(e)} | \bar{\mathbf{T}}^{(s)}, \bar{\mathbf{M}}) = \sum_{s=1}^S \sum_{m,t} \gamma_m(t) \log p(\mathbf{z}_t^{(s)}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

Note that the accumulation is done over all speakers (S) appearing in the environment. Substituting Eq. 4 into the above equation, the auxiliary function becomes equivalent to

$$\mathcal{Q}(\mathbf{T}^{(e)} | \bar{\mathbf{T}}^{(s)}, \bar{\mathbf{M}}) = \sum_{s=1}^S \sum_{m,t} \gamma_m(t) \log p(\mathbf{y}_t^{(s)}; \boldsymbol{\mu}_m^{(s)}, \boldsymbol{\Sigma}_m^{(s)}) \quad (7)$$

where $\boldsymbol{\mu}_m^{(s)}$ and $\boldsymbol{\Sigma}_m^{(s)}$ are the mean and co-variance of Gaussian component m in the GMM-HMM transformed using inverse of global CMLLR for speaker s :

$$\boldsymbol{\mu}_m^{(s)} = \mathbf{A}^{(s)-1} (\boldsymbol{\mu}_m - \mathbf{b}^{(s)}) \quad (8)$$

$$\boldsymbol{\Sigma}_m^{(s)} = \mathbf{A}^{(s)-1} \boldsymbol{\Sigma}_m \mathbf{A}^{(s)-T} \quad (9)$$

The likelihood in Eq. 7 is

$$\log p(\mathbf{y}_t^{(s)}; \boldsymbol{\mu}_m^{(s)}, \boldsymbol{\Sigma}_m^{(s)}) = \log \left| \frac{\partial \mathcal{F}^{(e)}(\mathbf{x}_t^{(s)})}{\partial \mathbf{T}^{(e)}} \right| - \frac{1}{2} (\mathcal{F}^{(e)}(\mathbf{x}_t^{(s)}) - \boldsymbol{\mu}_m^{(s)})^T \boldsymbol{\Sigma}_m^{(s)-1} (\mathcal{F}^{(e)}(\mathbf{x}_t^{(s)}) - \boldsymbol{\mu}_m^{(s)}) \quad (10)$$

The Jacobian term appears due to feature-space transformation. It is noted from Eq. 9 that the co-variance matrices, even though diagonal originally, are forced to become full after being transformed by the speaker global CMLLR. Optimization using this form of auxiliary function would require conversion of all Gaussians in the GMM-HMM to every speaker separately, and collection of associated statistics over the resulting *full-covariance* components, which are computationally inefficient.

However, with some rearrangement of terms, the auxiliary function can be reduced to the following form:

$$\mathcal{Q}(\mathbf{w}_{ci}^{(e)}) = \beta_c \log |\mathbf{A}_c^{(e)}| + \mathbf{w}_{ci}^{(e)} \left(\mathbf{k}_c^{iT} - \sum_{j=1, j \neq i}^d \mathbf{G}_{c,ij}^{(e)} \mathbf{w}_{cj}^{(e)T} \right) - \frac{1}{2} \mathbf{w}_{ci}^{(e)} \mathbf{G}_{c,ii}^{(e)} \mathbf{w}_{ci}^{(e)T} \quad (11)$$

where $\mathbf{w}_{ci}^{(e)}$ is row i of $\mathbf{W}_c^{(e)}$. \mathbf{k}_c^i is the row i of the first order statistic $\mathbf{K}_c^{(e)}$ and $\mathbf{G}_{c,ij}^{(e)}$ is the $(i, j)^{th}$ block of second order statistic $\mathbf{G}_c^{(e)}$, both for transform c . The occupation count is

$$\beta_c = \sum_{s=1}^S \sum_{m,t} \gamma_m(t) p(g_c | \mathbf{x}_t^{(s)}). \quad (12)$$

The first and second order statistics appearing in Eq. 11 are

$$\begin{aligned} \mathbf{K}_c^{(e)} &= \sum_{s=1}^S \mathbf{A}^{(s)T} \mathbf{K}_c^{(es)} \\ \mathbf{G}_{c,ij}^{(e)} &= \sum_{s=1}^S \sum_{u=1}^d a_{ui}^{(s)} a_{uj}^{(s)} \mathbf{G}_{c,u}^{(es)} \end{aligned} \quad (13)$$

$i, j = 1, \dots, d,$

where d is the dimension of observation vectors, $a_{ij}^{(s)}$ is $(i, j)^{th}$ element of speaker transform $\mathbf{A}^{(s)}$ and

$$\begin{aligned} \mathbf{K}_c^{(es)} &= \sum_{m,t} \gamma_m(t) p(g_c | \mathbf{x}_t^{(s)}) \Sigma_m^{-1} (\boldsymbol{\mu}_m - \mathbf{b}^{(s)}) \mathbf{x}_t^{(s)+T} \\ \mathbf{G}_{c,i}^{(es)} &= \sum_{m,t} \frac{\gamma_m(t) p(g_c | \mathbf{x}_t^{(s)})}{\sigma_{mi}^2} \mathbf{x}_t^{(s)+} \mathbf{x}_t^{(s)+T}. \end{aligned} \quad (14)$$

σ_{mi}^2 denotes the i^{th} element of the diagonal co-variance of component m in the GMM-HMM.

As prescribed by Eq. 14, now the $\mathbf{K}_c^{(es)}$ and $\mathbf{G}_{c,i}^{(es)}$ statistics are collected over the *diagonal* co-variance Gaussians for each speaker separately. Afterwards, as shown in Eq. 13, these statistics are combined over the speakers after being transformed with the CMLLR to yield the statistics that pertain to the full-covariance components and are used for the optimization. This way direct accumulation over the GMM-HMM, by converting it to each speaker, is avoided. Moreover, it is noted from Eq. 11 that row-by-row optimization is applicable, however, with appropriate modification of the \mathbf{k}_c^i statistics. The above modifications help to improve computational efficiency considerably.

3.3. Training procedure – iterative estimation

The iterative training proposed in this paper is outlined here, which is based on the theory developed in this section.

1. Initialization: The GMM-HMM system is trained in the un-normalized feature space (LDA+MLLT features in this case), followed by speaker adaptive training (SAT) using global CMLLR. The speaker transforms and the model are initialized to those yielded by the SAT system.

The environment transform is initialized to the identity FE-CMLLR, by setting all constituent affine transforms to identity matrices. The front-end GMM is trained by clustering all Gaussians in the GMM-HMM in the un-normalized space to the required number of components. It is kept fixed throughout the training.

2. Fix speaker transforms and model, and update environment transforms $\mathcal{F}^{(e)}$ using un-normalized features.
3. Fix environmental transforms and the model, and update the speaker transforms $\mathcal{F}^{(s)}$ using environment normalized features.
4. Fix environment and speaker transforms, update model \mathcal{M} using environment and speaker normalized features.
5. Iterate step 2 to 4 several times.
6. Discriminative training of GMM-HMM with model-space boosted Maximum Mutual Information (BMML) [20] on the final features.

3.4. Training procedure – sequential estimation

The sequential training outlined here is similar in spirit to [4] and is used as one of the baseline systems for the experiments.

1. Environment Adaptive Training:
 - (a) The model is initialized to the GMM-HMM trained in the un-normalized feature space. The environment transform is initialized to the identity FE-CMLLR. The front-end GMM is trained in the same way as in iterative training.
 - (b) Fix the model, and update environment transforms $\mathcal{F}^{(e)}$ using un-normalized features.
 - (c) Fix environmental transforms and update model \mathcal{M} using environment normalized features.
 - (d) Iterate step (b) and (c) several times.
2. Speaker Adaptive Training (SAT):
 - (a) The model and the environment transforms are initialized to the final GMM-HMM and the FE-CMLLR obtained in Step-1. The speaker CMLLRs are initialized to identity affine matrices.
 - (b) Fix the model, and update speaker transform $\mathcal{F}^{(e)}$ using environment normalized features.
 - (c) Fix speaker transform and update model \mathcal{M} using environment and speaker normalized features.
 - (d) Iterate step (b) and (c) several times.
3. BMML training of GMM-HMM on the final features.

4. Experimental Setup and Results

The proposed joint normalization method is evaluated on the Aurora-4 task [21], which is derived from the Wall Street Journal (WSJ0) 5k-word dictation task by digitally adding noise. The 16KHz data was used in the experiments. The training set consists of 12 hours of audio, containing 7137 utterances from 83 speakers. Recording is done under two microphone conditions – Mic-1, consisting of a close-talking Sennheiser microphone, and Mic-2, consisting of multiple desk-mounted secondary microphones. Each of the two sets are equally divided into 7 subsets, out of which 6 are further corrupted by adding different types of noise (car, babble, restaurant, street,

Table 1: WER (%) with F-FE-CMLLR iterative and sequential training (Mic-1 eval set).

method	clean	airport	babble	car	restaurant	street	train	Avg
multi-condition SI	5.6	10.7	10.2	6.0	15.4	12.7	13.8	10.6
CMLLR-speaker	5.2	8.6	7.7	4.7	11.7	11.4	10.7	8.6
F-FE-CMLLR-sequential	4.3	7.5	7.3	4.4	10.8	9.5	9.6	7.6
F-FE-CMLLR-iterative	4.1	7.7	7.7	4.5	10.8	9.5	8.9	7.6

Table 2: WER (%) with F-FE-CMLLR iterative and sequential training (Mic-2 eval set).

method	clean	airport	babble	car	restaurant	street	train	Avg
multi-condition SI	13.8	22.6	23.0	17.1	28.5	27.5	27.5	22.9
CMLLR-speaker	9.1	19.2	19.3	13.4	22.6	22.4	22.3	18.3
F-FE-CMLLR-sequential	8.6	17.9	17.9	12.1	20.3	21.7	21.6	17.2
F-FE-CMLLR-iterative	8.3	17.6	17.1	12.6	19.7	20.5	20.8	16.7

airport and train) at randomly selected SNR between 10 and 20 dB. Noise is not added to the remaining subset, which remains “clean”. The evaluation set comprises of 14 subsets, each containing 330 utterances from 8 speakers. 2 of them are recorded under Mic-1 and Mic-2 without further addition of noise, and the remaining 12 are produced by adding one of the above 6 noise types at randomly chosen SNR between 5 and 15 dB to each microphone type. The speakers in the evaluation set are different from those in the training set.

All experiments are conducted using Kaldi speech recognition toolkit [22]. The feature processing pipeline is shown in Figure 1. The 13 dimensional Mel-frequency cepstral coefficients (MFCC), comprising of 12 cepstral coefficients and C_0 , are mean normalized on a per speaker basis, and spliced taking a context length of 3 frames on both side of the central frame. Subsequently Linear Discriminant Analysis (LDA) [23] is applied to reduce the dimensionality to 40. The resulting features are further de-correlated using Maximum Likelihood Linear Transform (MLLT) [24], also known as global semi-tied covariance (STC) [25]. The features are further processed by different feature normalization schemes studied in this paper. The context-dependent tri-phone GMM-HMMs contain about 2500 states with a total of approximately 15000 components. The WSJ tri-gram language model with word pruning was used.

Three systems are developed for the experiments. The baseline system comprises the standard SAT [10] GMM-HMM using global CMLLR, which is further trained using BMMI in the speaker normalized space. The other two are the F-FE-CMLLR systems, built using either iterative or sequential training, followed by BMMI training. Environment and speaker adaptive training is applied as discussed in Section 3. Since environment labels were not available in the training set, “global” environment transform consisting of an FE-CMLLR with 128 affine-transforms is used. Environment normalization takes place in the supervised mode, i.e., the transform is estimated from the training set and is used to normalize the test data. Speaker normalization is done in the unsupervised mode; the CMLLRs are estimated using the final ML GMM-HMM and environment normalized features from the respective systems. The hypothesis for supervision were obtained from the ML multi-condition speaker independent (SI) system. Finally, adaptive decoding is done using the normalized features and BMMI models.

Percentage of word error rate (% WER) on Mic-1 and Mic-2 evaluation sets are presented in Table 1 and Table 2 respectively, showing the WER for all noise types separately. The results are indicated by CMLLR-speaker, F-FE-CMLLR-iterative and F-FE-CMLLR-sequential for the above three sys-

tems. Results with BMMI multi-condition SI GMM-HMM are also shown for both evaluation sets; a lower overall WER on the Mic-1 eval set compared to Mic-2 eval set indicates that the distortion by the primary microphone is less significant than the Mic-2 microphones. Only speaker normalization provides a significant improvement on both eval sets for all types of noise, whereas the improvement on Mic-2 eval set is much larger.

F-FE-CMLLR provides an additive improvement over CMLLR only normalization. Both the sequential and iterative training yield similar performance gain on the Mic-1 eval set. The effectiveness of iterative F-FE-CMLLR is apparent on Mic-2 eval set, where the influence of microphone variability is more significant. On this set it outperforms sequential training under 6 out of 7 types of noise – the improvement is maximum under street noise, i.e., 1.2% absolute, while there is a drop in the performance for car noise. The average WER improved from 17.2% with sequential training to 16.7% with iterative training. The relative improvements with the later over the standard SAT system is 11.6% on Mic-1 eval set and 8.7% on Mic-2 set.

Additionally it is noted from Mic-2 eval set that car noise and restaurant noise yield the lowest and highest WERs among all noise types in the task, indicating the former being the least difficult noise type, while the later the most difficult one. Iterative training outperforms sequential training on the harder noise conditions (restaurant, street and train noises), does moderately better on clean condition, and underperforms under the relatively weak (car) noise. We might expect a more consistent result if environment-specific transforms were used.

5. Conclusions and Future Direction

In this paper an iterative training algorithm for F-FE-CMLLR is presented. On the Aurora 4 task, it is noted that this yields better normalization performance than sequential training under most noise conditions. The results make the argument stronger that iterative training of F-FE-CMLLR is better equipped to approximate acoustic factorization than sequential training, i.e., it helps to separate the environment and speaker effects, allowing each transform to model only one factor. As a consequence, the environment transform (trained in the supervised mode) generalizes better when used in conjunction with the (test) speakers not seen during the training, leading to a better ASR accuracy. However, due to the absence of environment labels in the training set, a global environment transform was used in the experiments. Our future direction includes further investigation of the scheme on other tasks where the environment information is known, which may reveal its effectiveness more clearly.

6. References

- [1] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Comp. Speech and Lang.*, vol. 9, no. 2, pp. 171–185, 1995.
- [2] M. J. F. Gales, "Model-based techniques for noise robust speech recognition," in *Ph.D. thesis, Cambridge University*, 1995.
- [3] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. of ICASSP*, 1996.
- [4] M. J. F. Gales and F. Flego, "Model-based approaches for degraded channel modelling in robust ASR," in *Proc. of InterSpeech*, 2012.
- [5] Jasha Droppo, Alex Acero, and Li Deng, "Evaluation of the SPLICE algorithm on the Aurora 2 database," in *Proc. of EuroSpeech*, Sydney, 2001.
- [6] D. Povey, B. Kingsbury, et al., "fMPE: Discriminatively trained features for speech recognition," in *Proc. of IEEE ICASSP*, 2005.
- [7] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Proc. of Interspeech*, 2005.
- [8] S. P. Rath, L. Burget, M. Karafiat, O. Glembek, and J. Cernocky, "A region-specific feature-space transformation for speaker adaptation and singularity analysis of Jacobian matrix," in *Proc. of Interspeech*, 2013.
- [9] S. S. Kozat, K. Visweswariah, and R. Gopinath, "Feature adaptation based on Gaussian posteriors," in *Proc. of IEEE ICASSP*, Toulouse, 2006.
- [10] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.
- [11] Xiong Xiao, Jinyu Li, Eng Siong Chng, and Haizhou Li, "Feature compensation using linear combination of speaker and environment dependent correction vectors," in *Proc. of IEEE ICASSP*, 2014.
- [12] M. J. F. Gales, "Acoustic factorisation," in *Proc. of ASRU*, 2001.
- [13] Y. Q. Wang and M. J. F. Gales, "An explicit independence constraint for factorised adaptation in speech recognition," in *Proc. of InterSpeech*, 2013.
- [14] Hyunson Seo, Hong-Goo Kang, and Michael L. Seltzer, "Factored adaptation of speaker and environment using orthogonal subspace transforms," in *Proc. of IEEE ICASSP*, 2014.
- [15] Y. Q. Wang and M. J. F. Gales, "Speaker and noise factorisation on the Aurora 4 task," in *Proc. of ICASSP*, 2011.
- [16] M.L. Seltzer and A. Acero, "Factored adaptation using a combination of feature-space and model-space transforms," in *Proc. of InterSpeech*, 2012.
- [17] L. Rigazio, P. Nguyen, D. Kryze, and J. C. Junqua, "Separating speaker and environmental variabilities for improved recognition in non-stationary conditions," in *Proc. of EuroSpeech*, 2001.
- [18] M.L. Seltzer and A. Acero, "Separating speaker and environmental variability using factored transforms," in *Proc. of InterSpeech*, 2011.
- [19] 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.
- [20] D. Povey, D. Kanevsky, et al., "Boosted MMI for model and feature-space discriminative training," in *Proc. of IEEE ICASSP*, 2008.
- [21] N. Parihar and J. Picone, "Aurora working group: Dsr frontend LVCSR evaluation AU/384/02," in *Tech. Rep., Inst. for Signal and Information Process, Mississippi State University*.
- [22] D. Povey, A. Ghoshal, et al., "The Kaldi Speech Recognition Toolkit," in *Proc. of IEEE ASRU*, 2011.
- [23] R. O. Duda, P. E. Hart, and David G. Stork, *Pattern Classification (2nd Edition)*, Wiley-Interscience, 2000.
- [24] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. IEEE ICASSP*, 1998, vol. 2, pp. 661–664.
- [25] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech and Audio Proc.*, vol. 7, no. 3, pp. 272–281, May 1999.