Feature Normalization Using Structured Full Transforms for Robust Speech Recognition

Xiong Xiao[1], *Jinyu Li*[2], *Eng Siong Chng*[1,3], *Haizhou Li*[1,3,4]

[1]Temasek Laboratories @ NTU, Nanyang Technological University, Singapore
[2]Microsoft Corporation, USA
[3]School of Computer Engineering, Nanyang Technological University, Singapore
[4]Department of Human Language Technology, Institute for Infocomm Research, Singapore

xiaoxiong@ntu.edu.sg, jinyli@microsoft.com, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg

Abstract

Classical mean and variance normalization (MVN) uses a diagonal transform and a bias vector to normalize the mean and variance of noisy features to reference values. As MVN uses diagonal transform, it ignores correlation between feature dimensions. Although full transform is able to make use of feature correlation, its large amount of parameters may not be estimated reliably from a short observation, e.g. 1 utterance. We propose a novel structured full transform that has the same amount of free parameters as diagonal transform while being able to capture correlation between feature dimensions. The proposed structured transform can be estimated reliably from one utterance by maximizing the likelihood of the normalized features on a reference Gaussian mixture model. Experimental results on Aurora-4 task show that the structured transform produces consistently better speech recognition results than diagonal transform and also outperforms advanced frontend (AFE) feature extractor.

Index Terms: robust speech recognition, feature normalization, maximum likelihood, eigen-decomposition.

1. Introduction

Speech recognition performance on noisy speech data is poor if the acoustic model is trained from clean speech data. This is due to the mismatch between the distributions of the clean and noisy speech data. Many techniques have been proposed to reduce the mismatch and can be grouped into two approaches, the model adaptation approach and feature compensation approach.

Model adaptation approach adapts clean acoustic model towards noisy test features. For example, maximum *a posteriori* (MAP) [1] and maximum likelihood linear regression (MLLR) [2] adapt acoustic model by using noisy speech data. Parallel model combination (PMC) [3] and vector Taylor series (VTS)-based adaptation [4] predict noisy acoustic model based on noise estimate and a physical model that characterizes the relationship between clean and noisy features. Although model adaptation approach is powerful, they generally require much higher computational load than feature compensation approach.

Feature compensation approach estimates clean features from noisy observations. For example, minimum mean square error (MMSE) estimators of clean speech are proposed in spectral domain [5] and cepstral domain (e.g. [6]). The success of these techniques heavily depends on accurate noise estimation which itself is a difficult problem. A group of feature space techniques, called feature normalization, do not require noise estimation. Feature normalization methods normalize the distribution of noisy features (typically over an utterance) to that of clean features. For example, the cepstral mean normalization (CMN) [7] normalizes the mean of noisy features; mean and variance normalization (MVN) [8] normalizes both mean and variance of noisy features; and histogram equalization (HEQ) [9] generalizes MVN by normalizing the histograms of noisy features. These feature normalization methods are also extended to multi-class normalization for better performance. In augmented CMN [10], speech and silence frames are normalized to their own reference means rather than a global mean. Similar two-class extension is also applied to MVN in [11] and it is shown that two-class MVN produces similar performance as the advanced feature extraction (AFE) [12] on Aurora-4 task [13]. In [14, 15], multi-class HEQ is proposed and good performance was reported on Aurora-2 task.

A limitation of feature normalization techniques is that they ignore the correlation between feature dimensions and process each dimension independently. Although cepstral features are only weakly correlated, the correlation between feature dimensions can be used to improve speech recognition performance. For example, semi-tied covariance model [16] shows that it is beneficial to model the cross-covariance between feature dimensions for speech recognition.

In this paper, we propose to incorporate feature correlation information in feature normalization. As we will show later, MVN and its multi-class extension are two special cases of constrained MLLR (CMLLR) [17] and therefore belong to the maximum likelihood (ML) feature adaptation framework. MVN uses a diagonal transform to scale the feature dimensions independently. We proposed to use full transform to allow interactions between dimensions. To keep the number of free parameters low, we use a novel structured full transform that has the the same number of free parameters as diagonal transform. The new transform is estimated in the CMLLR framework.

The organization of this paper is as follows. In section 2, we review MVN in the CMLLR framework and introduce the proposed structured full transform. In section 3, the proposed method is evaluated on the Aurora-4 task. In section 4, conclusion is presented.

2. Feature Normalization with Full Structured Transform

CMLLR [17] is a popular model adaptation method. Due to its constrained form of transform, CMLLR can also be implemented in feature space, also known as feature space MLLR (fMLLR) [18]. As CMLLR provides a general maximum like-

lihood framework for linear feature transformation, we will use the CMLLR formulation to derive our proposed method. We will first show that MVN is a special case of CMLLR, and then describe the proposed structured full transform.

2.1. MVN as A Special Case of CMLLR

Assume that the features are linearly transformed as follows:

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

where $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are the original and transformed feature vectors at frame t, respectively, $\mathbf{A}$ is a positive-definite matrix and $\mathbf{b}$ is a bias vector. The auxiliary function of CMLLR is

$$\begin{aligned} Q(\lambda, \hat{\lambda}) &= T \log |\mathbf{A}| - \frac{1}{2} \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_m(t) (\mathbf{A}\mathbf{x}(t) + \mathbf{b} \\ &\quad -\mu_m)^T \Sigma_m^{-1} (\mathbf{A}\mathbf{x}(t) + \mathbf{b} - \mu_m) \end{aligned} \tag{2}$$

where $\lambda = \{\mathbf{A}, \mathbf{b}\}$ is the set of parameters to be estimated, $\hat{\lambda} = \{\hat{\mathbf{A}}, \hat{\mathbf{b}}\}$ is the current estimate of the parameters, μ_m and Σ_m are the mean and covariance matrix of the m^{th} mixture in the model, M is the number mixtures, T is the number of frames in the noisy data to be processed, $\gamma_m(t)$ is the posterior probability of mixture m at frame t after the noisy features are observed. The solution of $\mathbf{A}$ and $\mathbf{b}$ are given in [19] for diagonal transform case and in [17] for full transform case.

If there is only one Gaussian in the reference model, and assume that the transform $\mathbf{A}$ is diagonal, the following closed-form solution can be derived:

$$\hat{\mathbf{A}} = \Sigma_r^{1/2} \Sigma_x^{-1/2} \tag{3}$$

$$\hat{\mathbf{b}} = \mu_r - \hat{\mathbf{A}} \mu_x \tag{4}$$

where Σ_r is the diagonal covariance matrix of the only Gaussian in the model, Σ_x is the diagonal covariance matrix of the data, μ_r and μ_x are the reference and data mean vectors, respectively. With this solution, we have:

$$\mathbf{y}(t) = \Sigma_r^{1/2} \Sigma_x^{-1/2} (\mathbf{x}(t) - \mu_x) + \mu_r \tag{5}$$

As this solution coincides with MVN, MVN is a special case of CMLLR for single Gaussian model and diagonal transform.

If there are multiple Gaussians in the model and each associated with a diagonal transform, the auxiliary function becomes

$$\begin{aligned} Q(\lambda, \hat{\lambda}) &= \sum_{m=1}^{M} \Bigg[\gamma_m \log |\mathbf{A}_m| - \frac{1}{2} \sum_{t=1}^{T} \gamma_m(t) (\mathbf{A}_m \mathbf{x}(t) \\ &\quad +\mathbf{b}_m - \mu_m)^T \Sigma_m^{-1} (\mathbf{A}_m \mathbf{x}(t) + \mathbf{b}_m - \mu_m) \Bigg] \end{aligned} \tag{6}$$

where $\gamma_m = \sum_{t=1}^{T} \gamma_m(t)$, $\mathbf{A}_m$ and $\mathbf{b}_m$ are the diagonal transform and bias vector for mixture m, respectively. The M transforms and bias vectors can be solved independently:

$$\hat{\mathbf{A}}_m = \Sigma_m^{1/2} \Sigma_{x,m}^{-1/2} \tag{7}$$

$$\hat{\mathbf{b}}_m = \mu_m - \hat{\mathbf{A}}_m \mu_{x,m} \tag{8}$$

where

$$\mu_{x,m} = \frac{1}{\gamma_m} \sum_{t=1}^{T} \gamma_m(t) \mathbf{x}(t) \tag{9}$$

$$\Sigma_{x,m} = \text{diag} \left[\frac{1}{\gamma_m} \sum_{t=1}^{T} \gamma_m(t) (\mathbf{x}(t) - \mu_{x,m})(\mathbf{x}(t) - \mu_{x,m})^T \right] \tag{10}$$

The final transformed feature vector is a linear combination of the mixture-dependent transformed features:

$$\mathbf{y}_c(t) = \sum_{m=1}^{M} \gamma_m(t) \left[\Sigma_m^{1/2} \Sigma_{x,m}^{-1/2} (\mathbf{x}(t) - \mu_{x,m}) + \mu_m \right] \tag{11}$$

The two-Gaussian MVN in [11] is a special case of the multi-class MVN, where one Gaussian is used to represent speech frames and the other for silence frames. When implementing multi-class MVN, the noisy features are first preprocessed by MVN with one Gaussian, and then used to find the posterior probability $\gamma_m(t)$. This is because the quality of $\gamma_m(t)$ from original noisy features could be too bad and will lead the normalization to wrong directions. Similar preprocessing is also used in multi-class HEQ in [14].

2.2. MVN with Structured Full Transform

The diagonal transforms in MVN do not allow interaction between feature dimensions. Although full transforms will be more flexible, they have a large amount of parameters and cannot be reliably estimated from a small amount of data, e.g. 1 utterance. In this section, we propose a structured full transform that is more powerful than diagonal transform, but with the same amount of free parameters as diagonal transforms.

The proposed transform has following structure:

$$\mathbf{A} = \mathbf{E}\mathbf{S}\mathbf{E}^{-1} \tag{12}$$

where $\mathbf{E}$ is a nonsingular (invertible) matrix and $\mathbf{S}$ is a diagonal matrix. With this structure, $\mathbf{A}$ can be seen as a linear combination of D rank-1 matrix:

$$\mathbf{A} = \sum_{i=1}^{D} s_i \mathbf{e}_i \mathbf{f}_i^T \tag{13}$$

where D is the dimension of the feature vectors, s_i is the i^{th} diagonal element of $\mathbf{S}$, and $\mathbf{e}_i$ and $\mathbf{f}_i$ are the i^{th} column vectors of $\mathbf{E}$ and $\mathbf{E}^{-1}$, respectively. $\mathbf{E}$ is pretrained and only $\mathbf{S}$ needs to be estimated during feature normalization. Hence, there are only D free parameters in the transform, the same as a diagonal transform. Similar structured matrix has been used for modeling the precision matrix of Gaussian in [20]. With the structured transform and a meaningful $\mathbf{E}$, it is possible to find $\mathbf{A}$ that is more powerful than a diagonal transform without increasing the number of free parameters.

2.2.1. Solution for S

Let's first assume that we already know $\mathbf{E}$ and derive the solution for $\mathbf{S}$. Substitute (12) into the auxiliary function (2) we get

$$\begin{aligned} Q(\lambda, \hat{\lambda}) &= T \log |\mathbf{E}\mathbf{S}\mathbf{E}^{-1}| - \frac{1}{2} \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_m(t) (\mathbf{E}\mathbf{S}\mathbf{E}^{-1}\mathbf{x}(t) \\ &\quad +\mathbf{b} - \mu_m)^T \Sigma_m^{-1} (\mathbf{E}\mathbf{S}\mathbf{E}^{-1}\mathbf{x}(t) + \mathbf{b} - \mu_m) \\ &= T + T \log |\mathbf{S}| - \frac{1}{2} \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_m(t) \\ &\quad (\mathbf{S}\mathbf{E}^{-1}\mathbf{x}(t) + \mathbf{E}^{-1}\mathbf{b} - \mathbf{E}^{-1}\mu_m)^T (\mathbf{E}^T \Sigma_m^{-1} \mathbf{E}) \\ &\quad (\mathbf{S}\mathbf{E}^{-1}\mathbf{x}(t) + \mathbf{E}^{-1}\mathbf{b} - \mathbf{E}^{-1}\mu_m) \end{aligned} \tag{14}$$

Unlike standard CMLLR, the covariance matrices of the Gaussians are not assumed to be diagonal in case of structured transform.

Let's make the following projections:

$$\mathbf{x}_p(t) = \mathbf{E}^{-1}\mathbf{x}(t) \quad (15)$$

$$\mathbf{b}_p = \mathbf{E}^{-1}\mathbf{b} \quad (16)$$

$$\mu_{m,p} = \mathbf{E}^{-1}\mu_m \quad (17)$$

$$\Sigma_{m,p} = \mathbf{E}^{-1}\Sigma_m\mathbf{E}^{-T} \quad (18)$$

Then the auxiliary function can be rewritten as follows:

$$Q(\lambda,\hat{\lambda}) = T + T\log|\mathbf{S}| - \frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)(\mathbf{S}\mathbf{x}_p(t)+\mathbf{b}_p - \mu_{m,p})^T\Sigma_{m,p}^{-1}(\mathbf{S}\mathbf{x}_p(t)+\mathbf{b}_p-\mu_{m,p}) \quad (19)$$

This is exactly the CMLLR problem with diagonal transform $\mathbf{S}$, but in the space projected by $\mathbf{E}^{-1}$ rather than in the original feature space. Closed-form solution for $\mathbf{S}$ exists if $\Sigma_{m,p}$ is diagonal [19].

*2.2.2. Solution for **E***

We now discuss how to obtain $\mathbf{E}$ and how to guarantee that the projected covariance matrix $\Sigma_{m,p}$ is diagonal. We can solve both problems by choosing $\mathbf{E}$ properly. In the simplest case, if we have just one Gaussian in the model, one option is to choose $\mathbf{E}$ as the eigenvectors matrix of the Gaussian covariance matrix, i.e. $\Sigma = \mathbf{E}\Lambda\mathbf{E}^T$, where Σ is the global full covariance matrix of the clean feature space. Then, the projected covariance matrix will be diagonal: $\Sigma_p = \mathbf{E}^{-1}\Sigma\mathbf{E} = \mathbf{E}^T\Sigma\mathbf{E} = \Lambda$, where $\mathbf{E}^{-1} = \mathbf{E}^T$ as eigenvector matrix is orthonormal. The resulting problem is the same as MVN except that the normalization takes place in the projected space of $\mathbf{E}^{-1}$ rather than in the original space. If we set $\mathbf{E} = \mathbf{I}$, the algorithm becomes MVN.

In a more general case, there are multiple Gaussians in the model. If we use one transform with each Gaussian, i.e. we have $\mathbf{A}_m = \mathbf{E}_m\mathbf{S}_m\mathbf{E}_m^{-1}$ for Gaussian mixture m, then $\mathbf{E}_m$ can be set to the eigenvector matrix of Σ_m, which is now full covariance matrix. The resulting normalization is similar to multi-class MVN, except that the normalization is performed in projected space by $\mathbf{E}_m^{-1}$ for mixture m rather than in the original feature space. If $\mathbf{E}_m = \mathbf{I}$ for all m, the normalization degenerates to multi-class MVN. In this paper, we study this case and the single Gaussian case in the experiments.

In the most general case, there are multiple transforms $\mathbf{A}_m$ and each transform is shared by a set of Gaussian mixtures. In this case, one possible solution is to adopt semi-tied covariance modeling [16] to build the reference GMM and associate $\mathbf{E}_m$ with the semi-tied transforms. With this selection, $\mathbf{E}_m$ will decorrelate the Gaussians belonging to transform $\mathbf{A}_m$ (although not perfectly). Due to page limit, we will not discuss this general case in this paper.

3. Experiments

3.1. Experimental Settings

The proposed feature normalization algorithm is evaluated on the large vocabulary Aurora-4 task [13] that is widely used as benchmarking for different noise robust techniques. The MFCC features are extracted using the standard WI007 feature extraction program [21]. In total, 39 features, including the 13 static cepstral features and their delta and acceleration features, are used as raw features. The cepstral energy feature c0 is used instead of the log energy. In all the experiments, the training and testing features are always processed by the same feature processing method. A triphone-based acoustic model is trained using the clean condition training scheme and 8kHz. There are about 2,800 tied states in the model, each with 8 Gaussian mixtures. The acoustic model and a bigram language model are tested on the 14 test cases of Aurora-4 task.

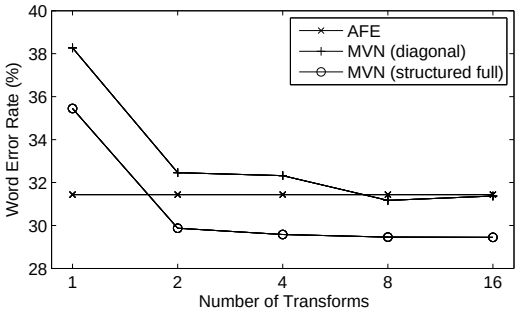

Figure 1: Performance of MVN with diagonal and structured full transforms on Aurora-4 task with different number of transforms. AFE result is also shown for comparison.

For feature normalization, reference GMMs are trained from the same clean features used to train the acoustic model. For MVN with diagonal transforms, GMM with diagonal covariance matrices. For MVN with structured full transforms, GMM with full covariance matrices are trained. The projection matrices for each Gaussian $\mathbf{E}_m$ are obtained as the eigenvectors matrices of the corresponding covariance matrices.

3.2. Experimental Results

We first examine the recognition performance with different number of Gaussians in the reference model as shown in Fig. 1. It is observed that the results obtained by using structured full transforms is consistently better than that with diagonal transforms. This shows that by allowing interaction between feature dimensions, the structured full transforms are able to use the correlation information between feature dimensions and this leads to better robustness of the normalized features. We also tried to use full covariance GMM with diagonal transforms, but this leads to worse results than using diagonal covariance GMM with diagonal transform. This shows that the advantage of structured full transform over diagonal transform is not due to the use of better full covariance matrix in structured full transform.

Fig. 1 also shows that WER obtained by diagonal transforms and structure transforms are both reduced when the number of classes are increased. The biggest improvement is from 1 to 2 mixtures. However, from 2 mixtures to 16 mixtures, there is only marginal improvement for both kinds of transforms. This suggests that the biggest improvement probably comes from using different mixtures to represent speech and silence as was suggested in [10] and [11]. The benefit of using more mixtures for better modeling of the speech frames is perhaps offsetted by less accurate posterior probability $\gamma_m(t)$ (the more mixtures, the less accurate posteriors). The result obtained by advanced front end (AFE) [12] is also shown in the figure for comparison. AFE is a state-of-the-art feature compensation technique and produces good results on Aurora-4 task. Our results show that multi-class MVN with diagonal transforms performs similarly as AFE (consistent with results in [11]) and multi-class MVN with structured transforms performs better than AFE. This shows that the proposed method is a competitive feature

Table 1: Performance on Aurora-4 task using clean condition training. MVNd and MVNf represent MVN with 1 diagonal and 1 structured full transform, respectively. MVNd8 and MVNf8 denote MVN with 8 diagonal and 8 structured full transforms, respectively. Avg. refers to the averaged WER over all 14 test cases. R.R. is the relative reduction of WER achieved by structured full transform over diagonal transform. AFE results are also shown for comparison.

Test Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Avg.
AFE	12.04	20.63	30.53	35.69	30.20	34.22	31.12	19.08	28.43	39.82	42.03	40.88	38.78	36.72	31.44
MVNd	13.44	31.79	37.09	38.23	37.20	36.24	40.26	21.33	40.11	45.52	49.28	52.56	46.22	50.17	38.53
MVNf	13.33	21.62	32.71	38.01	37.61	34.95	38.86	19.82	29.94	42.95	46.26	50.79	41.73	47.66	35.45
R.R	0.8	32.0	11.8	0.6	-1.1	3.6	3.5	7.1	25.3	5.7	6.1	3.4	9.7	5.0	**8.0**
MVNd8	12.56	16.50	31.05	36.61	33.55	32.30	34.51	16.61	23.35	36.35	43.57	42.36	37.16	39.85	31.17
MVNf8	11.57	15.65	28.58	35.36	31.38	30.02	33.96	15.80	21.10	33.19	40.66	41.10	35.95	38.05	29.46
R.R	7.9	5.1	7.9	3.4	6.5	7.1	1.6	4.9	9.6	8.7	6.7	3.0	3.3	4.5	**5.5**

space technique for improving robustness of features.

The detailed results of selected number of transforms are shown in Table 1. From the table, it is observed using structured transforms produces consistently better results than using diagonal transforms.

4. Conclusions

In this paper, we proposed to use structured full transform to replace diagonal transform of MVN feature normalization. The proposed transforms are estimated by maximizing the likelihood of the normalized features on a clean GMM reference model. Experimental results on Aurora-4 task show that the proposed structured full transform is able to use feature correlation information to improve robustness of features and improve speech recognition performance in noisy environments, while using the same number of free parameters as diagonal transforms. In the future, we will investigate structured full transform with projection matrices other than eigenvectors matrix, e.g. semi-tied transforms [16] and discriminative projections.

5. References

[1] J. L. Gauvain and C. H. Lee, "Maximum *a posterirori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[2] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[3] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition," *Speech Communication*, vol. 12, no. 3, pp. 231–239, Jul. 1993.

[4] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. ICSLP '00*, Beijing, China, Oct. 2000, pp. 869–872.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[6] L. Deng, J. Droppo, and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 3, pp. 218–223, May 2004.

[7] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[8] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[9] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[10] A. Acero and X. Huang, "Augmented cepstral normalization for robust speech recognition," in *Proc. of the IEEE Workshop on Automatic Speech Recognition*, Dec. 1995.

[11] L. Garca, J. C. Segura, J. Ramrez, A. de la Torre, and C. Bentez, "Parametric nonlinear feature equalization for robust speech recognition," in *Proc. ICASSP '06*, vol. I, Toulouse, France, May 2006, pp. 529–532.

[12] D. Macho, L. Mauuary, B. Noé, Y. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on aurora databases," in *Proc. ICSLP '02*, Denver, USA, Sep. 2002, pp. 17–20.

[13] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Institute for Signal and Infomation Processing, Mississippi State Univ., MS, Tech. Rep., Dec. 2002.

[14] Y. Suh, M. Ji, and H. Kim, "Probabilistic class histogram equalization for robust speech recognition," *IEEE Signal Processing letters*, vol. 14, no. 4, pp. 287–290, 2007.

[15] S.-H. Lin, B. Chen, and Y.-M. Yeh, "Exploring the use of speech features and their corresponding distribution characteristics for robust speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 84 –94, jan. 2009.

[16] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, may 1999.

[17] ——, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[18] Y. Li, H. Erdogan, Y. Gao, and E. Marcheret, "Incremental online feature space MLLR adaptation for telephony speech recognition," in *Proc. ICSLP '02*, Denver, USA, Sep. 2002, pp. 1417–1420.

[19] V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, 1995.

[20] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proc. ICASSP '98*, Seattle, WA, May 1998, pp. 661–664.

[21] D. Pearce and H.-G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recogntion systems under noisy conditions," in *Proc. ICSLP '00*, vol. 4, Beijing, China, Oct. 2000, pp. 29–32.