COMBINING EIGENVOICE SPEAKER MODELING AND VTS-BASED ENVIRONMENT COMPENSATION FOR ROBUST SPEECH RECOGNITION[1]

Zhijian Ou, Kan Deng

Department of Electronic Engineering, Tsinghua University, Beijing, China
Emails: ozj@tsinghua.edu.cn, dengk11@mails.tsinghua.edu.cn

ABSTRACT

Eigenvoice and vector Taylor series (VTS) are good models for speaker differences and environmental variations separately. However, speaker and environmental variation always coexist in real-world speech. In this paper, we propose to combine eigenvoice and VTS. Specifically, we introduce eigenvoice speaker modeling for the clean speech into VTS's nonlinear mismatch function. In contrast, the standard VTS uses speaker-independent modeling to represent the clean speech, regardless of speaker differences. The eigenvoice coefficients and the noise model parameters are jointly estimated in the new approach. Experimental results on the Aurora2 task show the improved performances of combining eigenvoice and VTS and demonstrate its ability for speaker and noise factorization.

Index Terms— robust speech recognition, vector Taylor series, speaker adaptation, eigenvoice

1. INTRODUCTION

The performances of speech recognition systems have been improved greatly, yet the robustness to various random interferences in speech signals still remains a challenging problem. Speaker differences and environmental variations are two major random factors in speech signals, which interfere in speech recognition.

Model-based approaches have been shown to be promising to deal with these two random factors separately. Regarding the robustness to speaker differences, speaker adaptation techniques, such as MLLR [1] and eigenvoice [2, 3], can adapt acoustic model to a new speaker. It is worthwhile to comment on these two approaches. We can find physical interpretations for eigenvoices, while it is hard to do so for MLLR transformation matrices. The advantage of the eigenvoice approach is that the a priori information about the inter-speaker variation is explicitly modeled and utilized to derive constraints for rapid speaker adaptation. The benefit of the MLLR approach is that it is flexible and can be used as a general adaptation method. Although this is practically useful, speaker differences and environmental variations are treated in a confused way. Regarding the robustness to environmental variations, model-based approaches based on vector Taylor series (VTS) expansion of the nonlinear mismatch function in the cepstral domain can effectively compensate for additive and convolutional distortions [4, 5, 6].

Note that speaker and environmental variation always coexist in real-world speech. There are several studies that consider joint handling of these two random factors. Acoustic factorization [7, 8] uses MLLR as speaker transform and cluster adaptive training as noise transform, both of which are linear transforms. This is not optimal if considering the nonlinear nature of the mismatch function relating the clean speech and the noisy speech. In a recent work [9], two combination schemes of MLLR and VTS are considered. One combination called "VTS+MLLR" conducts MLLR on top of the standard VTS. Using VTS is physically clear for noise compensation, but we can hardly interpret the MLLR used in this scheme as modeling the speaker variation. The "Joint" scheme replaces the clean speech model used in the VTS with a speaker-adapted clean speech model by MLLR transform. It is discovered that the speaker's MLLR transform estimated from the noisy speech using the "Joint" scheme still models some of the limitations of the VTS mismatch function [9], i.e. carries information about current noise characteristics. This hurts the performance when the estimated speaker transform from one noisy condition is used to recognize the speech from the same speaker under clean conditions.

In this paper, we consider how to do better speaker and noise factorization. Speaker and environmental variations have different characteristics. For speaker variation, the a priori information could be obtained by analyzing the training data with various speakers. The correlation analysis of the speaker supervectors leads to the eigenvoices, along which the speaker variation are significant [2, 3]. Using eigenvoice is statistically clear for speaker variation modeling. On the other hand, noise is hard to be modeled a priori. It is beneficial to perform adaptive noise compensation using online noise estimation based on the physical model relating the clean speech and the noisy speech – the mismatch function. With the above analysis, we propose to combine eigenvoice speaker modeling and VTS-based environment compensation so as to do better speaker and noise factorization. Intuitively, compared to MLLR, the eigenvoice speaker modeling puts strong restrictions on the speaker model; thus can help to better distinguish speaker variability from environment distortions.

Experiments are carried out using the Aurora2 database. First, the standard experimental setup of Aurora2 is used. Improved performances over the baseline VTS are obtained by combining eigenvoice and VTS. Next, we design an experiment where the estimated speaker model from a noisy utterance is used to recognize the corresponding clean utterance. The obtained performance is close to that of using the estimated speaker model from the clean utterance itself to do the recognition. This is a demonstration of speaker and noise factorization.

This paper is organized as follows. In Section 2, we review eigenvoice speaker modeling and VTS noise adaptation method. In

[1] This work is supported by National Natural Science Foundation of China (61075020) and China 863 (2006AA01Z149).

Section 3, we present the proposed joint adaptation scheme which combines eigenvoice and VTS. Section 4 gives the experimental results on the Aurora2 database. Finally, the conclusions are made in Section 5.

2. REVIEW OF EIGENVOICE SPEAKER MODELING AND VTS-BASED NOISE COMPENSATION

In this section, we review two adaptation methods, namely eigenvoice and VTS. Both of them rely heavily on prior knowledge, either obtained from statistical analysis of training data or governed by physical modeling of the signal transmission process.

2.1. Eigenvoice

In eigenvoice speaker modeling, the HMM's Gaussian means in any speaker-dependent (SD) model are concatenated to form a speaker supervector. Once the supervector's covariance matrix is estimated from the training data, we apply principle component analysis (PCA) to obtain the dominant eigenvectors, namely eigenvoices. The speaker supervector μ_x (as the clean speech model for a speaker) is assumed to be a weighted linear combination of the speaker average model e_0 and R eigenvoices e_r (r=1, 2, …, R) :

$$\mu_x = e_0 + \sum_{r=1}^{R} w_r e_r \tag{1}$$

The maximum likelihood estimation of the eigenvoice coefficients $w = (w_1, w_2, ..., w_R)^T$ is derived using the EM algorithm.

The auxiliary function of the parameters $\Theta = \{w\}$ is expressed as:

$$Q(\hat{\Theta} \mid \Theta) = \sum_t \sum_{j,k} \gamma_{jk}(t) \log p\left(x_t \middle| j,k, e_{0,jk} + \sum_{r=1}^{R} \hat{w}_r e_{r,jk}, \Sigma_{x,jk} \right) \tag{2}$$

where we use the subscript j,k of e_0 , e_r to denote the elements corresponding to state j and component k; $\gamma_{jk}(t)$ is the posteriori probability of state j and component k at frame t. $\mu_{x,jk}, \Sigma_{x,jk}$ are the Gaussian mean and covariance matrix of the HMM's state output distribution for state j and component k. Solving $\partial Q / \partial \hat{\Theta} = 0$ gives the linear equations to estimate the eigenvoice coefficients $\hat{w}$:

$$A\hat{w} = b \tag{3}$$

where, for $1 \le r, l \le R$,

$$\begin{cases} A(r,l) = \sum_{jk} \sum_t \gamma_{jk}(t) e_{r,jk}^T \Sigma_{jk}^{-1} e_{l,jk} \\ b(r) = \sum_{jk} \sum_t \gamma_{jk}(t) e_{r,jk}^T \Sigma_{jk}^{-1} (x_t - e_{0,jk}) \end{cases} \tag{4}$$

2.2. VTS

Noise is hard to be modeled a priori. However, we can exploit the physical model relating the clean speech and the noisy speech. Consider the effect of the additive noise n and the convolutional distortion h on the clean speech x . In the mel-cepstral domain, we have the following nolinear mismatch function relating the clean speech and the noisy speech [5]:

$$y = x + h + C \ln\left(1 + \exp\left(C^{-1}(n - x - h)\right)\right) \triangleq g(x,n,h) \tag{5}$$

where C is the DCT matrix. Here x is modeled by the standard acoustic HMM model with Gaussian means $\{\mu_{x,jk}\}$ and covariance matrices $\{\Sigma_{x,jk}\}$. For each utterance, n is assumed to be Gaussian distributed as $N(\mu_n, \Sigma_n)$, and $h = \mu_h$ is an unknown constant. There are two issues in order to apply the above model to noise compensation. First, given a clean acoustic model $\{\mu_{x,jk}, \Sigma_{x,jk}\}$ and an estimate of the noise model parameters $\Theta = \{\mu_n, \Sigma_n, \mu_h\}$, we need to obtain the noisy speech parameters $\{\mu_{y,jk}, \Sigma_{y,jk}\}$ for each Gaussian component of the acoustic HMM. Second, given the noisy speech, the noise model parameters Θ need to be estimated.

For the first issue, the basic idea is to approximate the nonlinear model of Equ. (5) by the first-order VTS expansion around μ_n, μ_h and $\mu_{x,jk}$ for each Gaussian component :

$$y \approx y\big|_{(\mu_{x,jk}, \mu_n, \mu_h)} + G_{x,jk}(x - \mu_{x,jk}) + G_{n,jk}(n - \mu_n) \tag{6}$$

where $G_{x,jk} = \left.\frac{\partial y}{\partial x}\right|_{(\mu_{x,jk}, \mu_n, \mu_h)}$ and $G_{n,jk} = \left.\frac{\partial y}{\partial n}\right|_{(\mu_{x,jk}, \mu_n, \mu_h)}$ are the Jacobian matrices. This yields the following distribution of the noisy speech y [5][2]:

$$\begin{cases} \mu_{y,jk} = g(\mu_{x,jk}, \mu_n, \mu_h) \\ \Sigma_{y,jk} = G_{x,jk} \Sigma_{x,jk} G_{x,jk}^T + (I - G_{x,jk}) \Sigma_n (I - G_{x,jk})^T \end{cases} \tag{7}$$

For the second issue, the noise model parameters Θ are re-estimated using the EM algorithm. The auxiliary function of the parameters $\Theta = \{\mu_n, \Sigma_n, \mu_h\}$ is expressed as:

$$Q(\hat{\Theta} \mid \Theta) = \sum_t \sum_{j,k} \gamma_{jk}(t) \log p\left(y_t \middle| j,k, \hat{\mu}_{y,jk}(\hat{\Theta}), \hat{\Sigma}_{y,jk}(\hat{\Theta}) \right) \tag{8}$$

Since the derivative of Q is a non-linear function of $\hat{\Theta}$, we approximate the root of the derivative through iterative refinements. Note that the nonlinear relationship between $\hat{\mu}_{y,jk}$ and $\hat{\Theta}$, as shown in Equ. (7), can again be approximated using the first-order VTS expansion around the old parameters Θ :

$$\hat{\mu}_{y,jk} \approx \mu_{y,jk} + G_{n,jk}(\hat{\mu}_n - \mu_n) + G_{h,jk}(\hat{\mu}_h - \mu_h) \tag{9}$$

where $G_{n,jk}, G_{h,jk}$ are the Jacobian matrices evaluated at the old parameter Θ . Based on this approximation, solving $\partial Q / \partial \hat{\mu}_n = 0$ gives the following re-estimation formula for μ_n :

$$\hat{\mu}_n = \mu_n + \left[\sum_{j,k} \gamma_{jk} G_{n,jk}^T \Sigma_{y,jk}^{-1} G_{n,jk} \right]^{-1} \sum_{jk} G_{n,jk}^T \Sigma_{y,jk}^{-1} c_{y,jk} \tag{10}$$

where we define the following sufficient statistics

$$\gamma_{jk} = \sum_t \gamma_{jk}(t) \tag{11}$$

$$c_{y,jk} = \sum_t \gamma_{jk}(t) \left(y_t - \mu_{y,jk} \right) \tag{12}$$

The derivation of the re-estimation formula for μ_h is derived in [5]. The re-estimation formula for Σ_n based on Gauss-Newton method is derived in [6].

[2] The compensation for derivative features is slightly different and will be omitted in the following of this paper to save space.

3. COMBINING EIGENVOICE AND VTS

Eigenvoice and VTS are good models for speaker differences and environmental variations separately. However, speaker and environmental variation always coexist in real-world speech. It is beneficial to combine eigenvoice and VTS. Specifically, we introduce eigenvoice speaker modeling for the clean speech into VTS's nonlinear mismatch function. In contrast, the standard VTS uses a speaker-independent HMM to represent the clean speech.

3.1. Parameter estimation for joint adaptation

The parameters to be estimated in the joint adaptation scheme, $\Theta = \left\{ \mu_n^{(u)}, \Sigma_n^{(u)}, \mu_h^{(u)}, u = 1, \cdots, U;\ w \right\}$, consists of two parts: the eigenvoice coefficients w for the target speaker and the noise model parameters $\Lambda = \left\{ \mu_n^{(u)}, \Sigma_n^{(u)}, \mu_h^{(u)}, u = 1, \cdots, U \right\}$ for a total of U utterances from that speaker. Each utterance u has its own noise parameters (indexed with the superscript u). Introducing eigenvoice speaker modeling for the clean speech into VTS's nonlinear mismatch function yields the following distribution for the u-th noisy speech y:

$$\begin{cases} \mu_{y,jk}^{(u)} = g\left(e_{0,jk} + \sum_{r=1}^{R} w_r e_{r,jk}, \mu_n^{(u)}, \mu_h^{(u)} \right) \\ \Sigma_{y,jk}^{(u)} = G_{x,jk}^{(u)} \Sigma_{x,jk} \left(G_{x,jk}^{(u)} \right)^T + (I - G_{x,jk}^{(u)}) \Sigma_n^{(u)} (I - G_{x,jk}^{(u)})^T \end{cases} \tag{13}$$

This is obtained by plugging Equ. (1) into Equ. (7) and noting that the covariance matrix $\Sigma_{x,jk}$ for the clean speech is still modeled globally (i.e. utterance-independent).

The auxiliary function of the parameters Θ is expressed as:

$$Q(\hat{\Theta} \mid \Theta) = \sum_{u,t} \sum_{j,k} \gamma_{jk}^{(u)}(t) \log p\left(y_t \middle| j, k, \hat{\mu}_{y,jk}^{(u)}\left(\hat{\Theta}\right), \hat{\Sigma}_{y,jk}^{(u)}\left(\hat{\Theta}\right) \right) \tag{14}$$

where $\hat{\mu}_{y,jk}^{(u)}, \hat{\Sigma}_{y,jk}^{(u)}$ are the compensated mean and covariance matrix for the u-th noisy utterance y as in Equ. (13). A block coordinate ascent strategy is used to optimize w and Λ iteratively. Λ is optimized while keeping w fixed, and vice versa.

Estimating the noise model parameters Λ while keeping w fixed is a simple extension of the VTS-based noise estimation as described in Section 2.2. The only difference is that in noise estimation, we use the speaker-dependent clean speech mean μ_x as in Equ. (1), instead of using the speaker-independent mean.

The estimation of the speaker's eigenvoice coefficients w given the current noise estimation $\hat{\Lambda}$ is derived as follows. The key is that the nonlinear relationship between $\hat{\mu}_{y,jk}^{(u)}$ and $\hat{w}$, as shown in Equ. (13), can again be approximated using the first-order VTS expansion around the old parameter w:

$$\hat{\mu}_{y,jk}^{(u)} \approx \mu_{y,jk}^{(u)} + G_{x,jk}^{(u)} \sum_{r=1}^{R} \left(\hat{w}_r - w_r \right) e_{r,jk} \tag{15}$$

Based on this approximation, we can compute $\partial Q / \partial \hat{w}$ as follows:

$$\frac{\partial Q(\hat{\Theta} \mid \Theta)}{\partial \hat{w}_r} = \sum_{u,t} \sum_{j,k} \gamma_{jk}^{(u)}(t) e_{r,jk}^T \left(G_{x,jk}^{(u)} \right)^T \left(\hat{\Sigma}_{y,jk}^{(u)} \right)^{-1} \left(y_t - \left\{ \mu_{y,jk}^{(u)} + G_{x,jk}^{(u)} \sum_{r=1}^{R} \left(\hat{w}_r - w_r \right) e_{r,jk} \right\} \right) \tag{16}$$

Solving $\partial Q / \partial \hat{w} = 0$ gives the linear equations to estimate $\hat{w}$:

$$A\hat{w} = b \tag{17}$$

where, for $1 \le r, l \le R$,

$$\begin{cases} A(r,l) = \sum_{jk} \sum_{u,t} \gamma_{jk}^{(u)}(t) e_{r,jk}^T \left(G_{x,jk}^{(u)} \right)^T \left(\hat{\Sigma}_{y,jk}^{(u)} \right)^{-1} G_{x,jk}^{(u)} e_{l,jk} \\ b(r) = \sum_{jk} \sum_{u,t} \gamma_{jk}^{(u)}(t) e_{r,jk}^T \left(G_{x,jk}^{(u)} \right)^T \left(\hat{\Sigma}_{y,jk}^{(u)} \right)^{-1} \left(y_t - \mu_{y,jk}^{(u)} + G_{x,jk}^{(u)} \sum_{r=1}^{R} w_r e_{r,jk} \right) \end{cases} \tag{18}$$

3.2. The proposed joint adaptation scheme

The implementation of the above iterative parameter estimation is described step by step in the following.

1. For each utterance, initialize the noise model mean and variance $\mu_n^{(u)}, \Sigma_n^{(u)}$ using the first and last several frames that are assumed to be speech-free, and set $\mu_h^{(u)} = 0$.
2. Update the noisy speech model according to Equ. (13) with w being initialized as 0, and do one pass recognition.
3. Based on current estimates of noise model parameters and eigenvoice coefficients, re-estimate the eigenvoice coefficients according to Equ. (18) and update the speaker adapted mean.
4. Based on current speaker adapted mean and noise model parameters, re-estimate the noise model parameters according to Equ. (10) and update the noisy speech model according to Equ. (13).

After step 3, the speaker-dependent clean speech model is estimated. After step 4, the noisy speech model is estimated. Steps 3 and 4 can be iterated for several times.

4. EXPERIMENTAL RESULTS

The proposed adaptation scheme is evaluated on the standard Aurora2 task of recognizing digit strings in noisy environment [10]. Three test sets called SetA, SetB and SetC are designed to evaluate recognition accuracies under different noise conditions. SetA and SetB each contain 4 types of additive noise. SetC contains 2 types of additive noise as well as channel distortion. The *average recognition accuracy* is calculated over 5 SNR levels between 0~20dB.

The clean training set, containing 8440 utterances from 110 speakers, is used to train the speaker independent (SI) model and eigenvoices. The feature is 39-dimensional MFCCs computed based on the power spectrum with the 0-th cepstral coefficient. The HMM configuration is consistent with the standard system described in [10]. The baseline system using speaker independent model achieves the average accuracy of 59.46%.

To obtain eigenvoices with insufficient data from each training speaker, we use supervised MLLR transforms to generate speaker-adapted models. Then PCA is performed to obtain the first 9 eigenvoices with the largest eigenvalues. To evaluate eigenvoice adaptation for clean speech, we perform unsupervised adaptation for each utterance in the clean test set, increasing the number of eigenvoice from 0 (i.e. using SI model) to 9. It can be seen from Table 1 that eigenvoice adaptation can further improve recognition accuracy even in the per-utterance mode, which means limited amount of adaptation data.

Various noise compensation methods are conducted to recognize in the noisy conditions, which are all applied in the per-utterance mode. The results are shown in Table 2. First, we initialize the mean and covariance of the additive noise by averaging the first and last 20 frames of an utterance. Using this initialized noise estimate, the average recognition accuracy is

Table 1: Recognition accuracies for per-utterance unsupervised eigenvoice adaptation under the clean condition

Eigenvoice Num	SI	1	2	3	4
Clean Acc (%)	99.00	98.98	99.10	99.10	99.10
Eigenvoice Num	5	6	7	8	9
Clean Acc (%)	99.11	99.09	99.08	99.10	99.09

88.04%, denoted as "VTS Init". Based on this initial hypothesis, the standard VTS, which uses the SI clean speech model, is performed to re-estimate the noise parameters for one iteration. This results in the accuracy of 90.23%, denoted as "VTS with SI model". Also based on the initial hypothesis produced by "VTS Init", the eigenvoice coefficients and noise parameters are jointly estimated for one iteration. It is shown in Fig. 1 how the average accuracies change as we use different number of eigenvoices in the joint adaptation scheme. As the number of eigenvoice increases to 5 and more, the performances of the joint scheme degrade due to the unreliable estimation of the eigenvoice ceofficients using limited adaptation data. The best average accuracy of 90.78% is obtained using 4 eigenvoices, as shown in the last row of Table 2. This is a significant improvement over the accuracy of 90.23% obtained by the standard VTS. The benefit of combining eigenvoice and VTS for robust speech recognition is clear.

In addition to this overall recognition accuracy evaluation, we design another experiment, again in the per-utterance mode, to investigate whether the clean speaker model estimated under noisy condition is affected by environmental factors. In this experiment, the clean speaker model estimated from the noisy utterance under the "VTS with 4 eigenvoices" scheme is used to recognize the corresponding clean utterance. This is compared to the standard unsupervised eigenvoice adaptation, which estimates the clean speaker model from the clean utterance itself (also using 4 eigenvoices) and do the recognition. It is can be seen from Table 3 that the performances of using the clean speaker models estimated from noisy data (across various SNRs from 0dB to 20 dB) are very close to that of using the clean speaker models estimated from clean data. This is a good demonstration of speaker and noise factorization.

Table 2: Average recognition accuracies for per-utterance unsupervised adaptation under noisy conditions by various schemes

Scheme	SetA	SetB	SetC	Avg. Acc.
Baseline	59.33	56.19	66.26	59.46
VTS Init	87.65	88.38	88.11	88.04
VTS with SI model	90.03	90.39	90.30	90.23
VTS with 4 eigenvoices	90.58	91.15	90.43	90.78

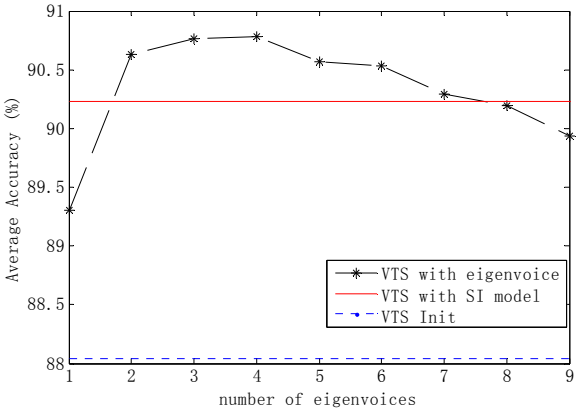

Fig. 1: Average recognition accuracies for per-utterance unsupervised adaptation under noisy conditions by various schemes

Table 3: Per-utterance unsupervised adaptation experimental results for recognizing the clean utterance, using the clean speaker model estimated from the noisy utterance under the "VTS with 4 eigenvoices" scheme. The "clean" represents the standard unsupervised eigenvoice adaptation scheme (i.e. using the clean speaker model estimated from the clean utterance itself)

SNR	clean	20dB	15dB	10dB	5dB	0dB
Acc (%)	99.10	99.10	99.11	99.12	99.08	99.00

5. CONCLUSION

Eigenvoice and VTS are good models for speaker differences and environmental variations separately. However, speaker and environmental variation always coexist in real-world speech. In this paper, we propose to combine eigenvoice and VTS. The eigenvoice coefficients and the noise model parameters are jointly estimated. Experimental results on the Aurora2 task show the improved performances of combining eigenvoice and VTS in the per-utterance mode that means limited amount of adaptation data and demonstrate its ability for speaker and noise factorization.

6. ACKNOWLEDGEMENT

We would like to thank Dr. Yong Zhao at Georgia Institute of Technology for valuable discussions.

7. REFERENCES

[1] C.J. Leggetter and P.C. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density HMMs, *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[2] 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, Nov.2000.

[3] Z. Ou, J. Luo, Latent correlation analysis of HMM parameters for speech recognition, *Proc. ICASSP*, 2007.

[4] P. J. Moreno, Speech Recognition in Noisy Environments, Ph.D. thesis, Carnegie Mellon Univ., 1996.

[5] J. Li, *et al.*, A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions, *Computer Speech and Language*, no. 3, vol. 23, 2009.

[6] Y. Zhao, B.H. Juang, Non-linear noise compensation for robust speech recognition using Gauss-Newton method, *Proc. ICASSP*, 2011.

[7] M.J.F Gales, Acoustic factorisation, *Proc, ASRU*, 2001

[8] K. Yu, M.J.F Gales, Adaptive training using structured transforms, *Proc. ICASSP*, 2004

[9] Y.Q Wang, M.J.F Gales, Speaker and noise factorization on the Aurora4 Task, *Proc. ICASSP*, 2011

[10] H.G. Hirsch, D. Pearce, The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, *ISCA ITRW ASR*, 2000.