Incorporating Regional Information to Enhance MAP-based Stochastic Feature Compensation for Robust Speech Recognition

Yu Tsao, Paul R. Dixon, Chiori Hori, and Hisashi Kawai

Spoken Language Communication Group
National Institute of Information and Communications Technology, Kyoto, 619-0288, Japan
{yu.tsao, paul.dixon, chiori.hori, hisashi.kawai}@nict.go.jp

Abstract

In this study, we propose an environment structuring framework to facilitate suitable prior density preparation for MAP-based stochastic feature matching (SFM) for robust speech recognition. We use a two-stage hierarchical structure to construct the environment structuring framework to characterize the regional information of various speaker and speaking environments. With the regional information, we derive three types of prior densities, namely clustered prior, sequential prior, and hierarchical prior densities. We also designed an integrated prior density to combine the advantages of the above three prior densities. From our experimental results on the Aurora-2 task, we confirmed that with regional information, we can obtain more suitable prior densities and thus enhance the performance of MAP-based SFM. Moreover, we found that by using the integrated prior density, which integrates multiple knowledge sources from the other three, MAP-based SFM gives the best performance.

Index Terms: stochastic feature matching, SFM, hierarchical SFM, environment structuring, robust speech recognition.

1. Introduction

The mismatch between training and testing conditions is a critical issue for automatic speech recognition (ASR). Many approaches have been proposed to reduce this mismatch. Among them, feature compensation is a known effective way. It estimates a transformation function to compensate testing speech features to match the acoustic model based on some optimality criterion and with available statistics from the testing condition. Maximum likelihood (ML) is a successful criterion. Effective ML-based feature compensation examples include the ML-based stochastic feature matching (SFM) algorithm [1] and feature space maximum likelihood linear regression (feature space MLLR) [2, 3]. However in real world applications, testing statistics might be limited or contain erroneous information. Over-fittings might occur in the ML-based methods. Maximum a posteriori (MAP)-based solutions were proposed to address this issue. For MAP-based feature compensation, a prior density is incorporated to confine the parameter estimation of the transformation function to avoid over-fitting issues. Noted examples include MAP-based SFM [4] and feature space maximum a posteriori linear regression (feature space MAPLR) [5].

For the MAP estimate, it is crucial to use a suitable prior density for the particular testing condition. Learning from MAP-based model adaptation studies, using a hierarchical structure can be successful. Structural MAP (SMAP) [6] and structural MAPLR (SMAPLR) [7] are two good examples. Another way is to use the knowledge obtained from previous utterances. The quasi-Bayes linear regression (QBLR) algorithm is a noted approach [8]. Meanwhile, some recent studies collected regional information of various acoustic conditions by using an environment framework. Then, the regional information is incorporated to calculate suitable prior density for the MAP-based acoustic model adaptation [9, 10].

In this paper, we focus our discussion on MAP-based feature compensation for robust ASR. We first incorporate the regional information to derive three advanced prior densities—clustered prior (CP), sequential prior (SP), and hierarchical prior (HP) densities. We also develop an integrated prior (IP) density that combines the three prior densities. Because the three densities are estimated by different schemes and in different operation phases, IP can be considered an integration of multiple knowledge sources from the other three.

The rest of this paper is organized as follows. Section 2 presents the environment structuring framework and the ML- and MAP-based SFM algorithms. Section 3 introduces four types of prior density. Section 4 discusses the experimental setup and results. Finally, Section 5 summarizes our findings.

2. Environment structuring framework with ML- and MAP-based SFMs

In this section, we first present two algorithms that we use to construct the environment framework. Next, we briefly review the ML- and MAP-based SFM algorithms. Finally, we present the integration of ML- and MAP-based SFMs with the environment structuring framework.

2.1. Environment structuring framework

The two algorithms with which we construct the environment structuring framework are environment clustering (EC) and environment partitioning (EP) [11]. The EC algorithm clusters the entire set of training data into several subsets, each of which contains training data of similar acoustic characteristics. In this study, a hierarchical tree facilitates the EC process. The EP algorithm partitions the mean parameters in a set of HMMs into several classes. Each class consists of the mean parameters of close properties. We also use a hierarchical tree structure to perform the EP process.

Figure 1 illustrates the overall environment structuring framework that combines the EC and EP structures [9]. We first build an EC tree to cluster the training data into *C* groups. For each EC node, we establish a representative HMM set using the training data for that node. Next, we construct an EP tree to partition mean parameters in each representative HMM set. Finally, the environment structuring framework consists of one EC tree that comprises *C* EC nodes. Each EC node, along with its representative HMM set and EP tree, corresponds to specific regional information of the ensemble environments.

We prepare the environment structuring framework in the offline. In the online, given the testing utterance we perform a cluster selection (CS) procedure to choose one EC node that best matches the testing condition; the best matched HMM set and EP tree structure for that EC node are accordingly located.

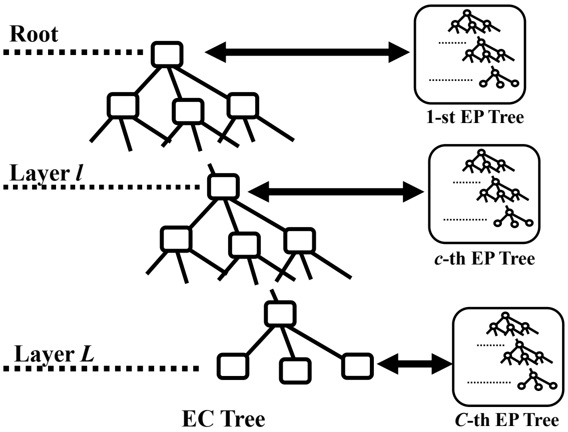

Figure 1. Environment clustering and partitioning framework

2.2. ML- and MAP-based SFMs

The SFM algorithm adopts a mapping function, F_v , to characterize the mismatch between the testing features and the acoustic model [1]. The goal of SFM is to compensate testing data Y to X that matches the training condition:

$$\mathrm{X} = \mathrm{F}_v(\mathrm{Y}). \tag{1}$$

For the ML-based SFM, we estimate nuisance parameters,v, of the mapping function, F_v, by:

$$\hat{v}_{ML} = \underset{v}{argmax}\ P(\mathrm{Y}|v,\Lambda_\mathrm{X}). \tag{2}$$

On the other hand, for the MAP-based SFM, we use:

$$\hat{v}_{MAP} = \underset{v}{argmax}\ P(\mathrm{Y}|v,\Lambda_\mathrm{X})[p(v)]^{\epsilon}, \tag{3}$$

where $p(v)$ is the prior density, and ϵ is a forgetting factor [4].

2.3. Environment structuring framework with SFM

With the environment structuring framework shown in Figure 1 and given a testing utterance, $\mathrm{Y} = [y^1\ y^2 \ldots y^T]$ with T frames, we perform SFM by the following five steps:

Step-1: With Y, perform the CS procedure to select one EC node (e.g., the c-th node) that best matches the testing condition. Representative HMMs, Λ_c, and EP tree, Ω_c, for that c-th node are accordingly located.

Step-2: Decode Y by Λ_c, and use the decoded results to obtain the alignment information of the testing frames and the corresponding nodes in Ω_c.

Step-3: Use the decoded results to estimate the mapping functions for every node in Ω_c. The transformation function in Eq. (1) now becomes F_v=$\{f_v^1\ f_v^2 \ldots f_v^N\}$, when $\Omega_c$has N nodes. For an EP node with fewer testing statistics than a predetermined threshold, we directly copy the transformation function from its parent node to this current node.

Step-4: Based on the alignment information from Step-2, perform SFM on Y. For the t-th frame, y^t, we first find its corresponding EP node (e.g., the n-th node) and perform:

$$x^t = f_v^n(y^t), \tag{4}$$

where x^t is the compensated feature, and $f_v^n(.)$ is the mapping function for the n-th EP node in Ω_c. Finally, we obtain the compensated feature $\mathrm{X} = [x^1\ x^2 \ldots x^T]$.

Step-5: Recognize X using Λ_c to obtain the final result.

In Step-3, the ML- and MAP-based SFMs use Eqs. (2) and (3), respectively, to estimate mapping functions, $f_v^n(.)$.

When we have sufficient adaptation statistics, we can use a complex parametric function for $f_v^n(.)$ to enhance the SFM accuracy. In this study, we focus on the testing situations that only have limited amounts of testing statistics. Therefore, a simple compensation bias is used for the mapping function, $f_v^n(.)$. Then, Eq. (4) becomes:

$$x^t = y^t - b_n. \tag{5}$$

Now, for the ML-based SFM, we have:

$$b_{n_{(i)}} = k_{n_{(i)}}/G_{n_{(i)}}\,, \tag{6}$$

where

$$G_{n_{(i)}} = \sum_{t=1}^{T}\sum_{s\in n} r_s(t)\left[\frac{1}{\sigma^2_{s_{(i)}}}\right], \tag{7}$$

$$k_{n_{(i)}} = \sum_{t=1}^{T}\sum_{s\in n} r_s(t)\left[\frac{y_{t_{(i)}} - \mu_{s_{(i)}}}{\sigma^2_{s_{(i)}}}\right], \tag{8}$$

$b_{n_{(i)}}$ is the i-th component of the compensation bias, and $r_s(t)$ is the posterior probability at the t-th observation. $\mu_{s_{(i)}}$and $\sigma^2_{s_{(i)}}$ are the i-th component of the mean and variance of the s-th Gaussian components belonging to the n-th EP node.

For the MAP-based SFM, we first define a prior density:

$$p(b_n) \sim \prod_{i=1}^{D} \exp\left[-\frac{1}{2\,\mathrm{V}_{n_{(i)}}}\left(b_{n_{(i)}} - \eta_{n_{(i)}}\right)^2\right], \tag{9}$$

where $\eta_{n_{(i)}}$ and $\mathrm{V}_{n_{(i)}}$ are the hyper-parameters, and D is the dimensionality of each feature vector.

Then, the MAP estimate of the compensation bias is:

$$b_{n_{(i)}} = k_{n_{(i)}}/G_{n_{(i)}}\,, \tag{10}$$

where

$$G_{n_{(i)}} = \frac{\epsilon}{\mathrm{V}_{n_{(i)}}} + \sum_{t=1}^{T}\sum_{s\in n} r_s(t)\left[\frac{1}{\sigma^2_{s_{(i)}}}\right], \tag{11}$$

$$k_{n_{(i)}} = \frac{\epsilon}{\mathrm{V}_{n_{(i)}}}\eta_{n_{(i)}} + \sum_{t=1}^{T}\sum_{s\in n} r_s(t)\left[\frac{y_{t_{(i)}} - \mu_{s_{(i)}}}{\sigma^2_{s_{(i)}}}\right]. \tag{12}$$

3. Four types of prior densities

The main focus of this study is to derive more suitable prior densities in Eq. (9) by exploiting the environment structuring framework for the MAP-based SFM. This section presents four prior densities—clustered prior (CP), sequential prior (SP), hierarchical prior (HP), and integrated prior (IP).

3.1. Clustered priors (CP)

We prepare a CP density for every EP node in the environment structuring framework (Figure 1). Here, we use the c-th EC node as an example, and our goal is to calculate a CP density for every node in the EP tree of that c-th EC node. Assume that the c-th EC node includes training data from K different environments. For the n-th EP node, we first calculate K sets of compensation bias $\{b_n^{(1)}, b_n^{(2)}, \ldots b_n^{(K)}\}$. Then, we estimate the hyper-parameters of the CP density, $\{\eta_n^{CP},\ \mathrm{V}_n^{CP}\}$ by:

$$\eta^{CP}_{n_{(i)}} = \frac{1}{K}\sum_{k=1}^{K} b^{(k)}_{n_{(i)}}\,, \tag{13}$$

$$\mathrm{V}^{CP}_{n_{(i)}} = \frac{1}{K}\sum_{k=1}^{K}\,(b^{(k)}_{n_{(i)}} - \eta^{CP}_{n_{(i)}})^2\,, \tag{14}$$

where $\eta^{CP}_{n_{(i)}}$ and $\mathrm{V}^{CP}_{n_{(i)}}$ are the i-th component of η_n^{CP} and V_n^{CP}.

With the same procedure, we can estimate the CP densities for every node in all the C EP trees. Because each CP density corresponds to a certain group of mean parameters for a particular cluster of environments, it provides regional information of the ensemble environments. With the online CS process, we can directly locate the CP density that best matches the testing condition for the MAP-based SFM.

3.2. Sequential priors (SP)

The sequential prior (SP) density is based on the sequential Bayesian learning. With SP density, the MAP-based SFM incorporates the information seen before for compensating the current testing utterances. In a previous study, an MAP estimate using SP density showed satisfactory performance [4]. In this study, we propose an environment dependent (ED) SP using the regional information provided by the EC algorithm. In this way, for any test utterance, Y, we prepare *C* sets of SP density. By performing the CS procedure, only the SP density that best matches the testing condition is selected for performing the MAP-based SFM. After that, only the selected set of SP density is updated and will be used as the new SP density for the next utterance. Here, we only sequentially update η_n^{SP} and use a fixed V_n^{SP} to simplify online computation.

3.3. Hierarchical priors (HP)

The estimate of HP densities for the MAP-based SFM resembles that performed in SMAP [6] and SMAPLR [7]. As mentioned earlier, we prepare *C* EP trees in the environment structuring framework, and each EP tree characterizes particular regional information. With the CS procedure, we locate the EP tree that matches the testing condition and estimate the HP density using the located EP tree. When computing the HP density, we first estimate a compensation bias at the top node of the selected EP tree. The estimated compensation bias is propagated to the child nodes and used as the HP density in the next layer. The estimation and propagation processes iterate and finally stop at the desired layer of the EP tree. Similar to SP, we only online estimate mean hyper-parameters, η_n^{HP}, and fix hyper-parameters, V_n^{HP}. The major advantage of HP is that it efficiently uses the information from the current utterance using the EP structure.

3.4. Integrated priors (IP)

We propose an IP density that combines the above three prior densities using a function, $\Gamma(.)$:

$$\eta_n^{IP} = \Gamma(\eta_n^{CP}, \eta_n^{SP}, \eta_n^{HP}) \,. \tag{15}$$

In this study, we simply used a linear combination function for $\Gamma(.)$ to estimate the integrated prior, η_n^{IP} by:

$$\eta_n^{IP} = w_{CP}\eta_n^{CP} + w_{SP}\eta_n^{SP} + w_{HP}\eta_n^{HP} \,, \tag{16}$$

with

$$w_{CP} + w_{SP} + w_{HP} = 1 \,, \tag{17}$$

where w_{CP}, w_{SP}, and w_{HP} are weighting coefficients. We can use a particular criterion to optimize the weighting coefficients. In this paper, we simply optimize the coefficients using a subset of training data. Similar to HP, we first locate one EP tree, and then based on Eqs. (15) and (16), we iteratively estimate and propagate the IP densities. Finally, the estimation and propagation stop at the desired layer of the EP tree. Note that CP, SP, and HP are calculated using the information from the training data, statistics seen from the previous utterances, and the current testing data with the EP tree, respectively. Therefore, we believe that the IP density incorporates multiple knowledge sources. Similar to SP and HP, we only online update η_n^{IP} and use a fixed V_n^{IP} for the IP density.

4. Experiments

In this section, we briefly introduce the experimental setup and present and discuss our experimental results of the ML- and MAP-based SFMs with four types of prior densities.

4.1. Experimental setup

We conducted speech recognition experiments on the Aurora-2 task [12]. The multi-condition training set was used for estimating acoustic models and for preparing an environment structuring framework (Figure 1). This training set includes 17 different speaking environments from four types of noise at 5dB, 10dB, 15dB, 20dB SNRs and a clean condition. By further dividing the speakers' genders, we obtained 34 sets of training data for 34 speaker and speaking environments.

To construct the environment structuring framework, we first built a two-layer EC tree to cluster the 34 environments into seven groups (one root, two intermediate, and four leaf nodes). The first layer classified the 34 environments into two groups of two genders, and the second layer separated 17 environments into two groups roughly according to high/low SNR levels. Next, we used the training data for each EC node to build a representative HMM set. Then, we built an EP tree for each EC node to partition mean parameters. Each EP tree has one root, three intermediate, and six leaf nodes. For the partitioning procedure, the weighted Euclidean distance was used as the distance measure for each pair of mean vectors.

A modified ETSI advanced front end (AFE) was adopted for feature extraction [13]. Every feature vector consisted of 13 static plus their first- and second-order time derivatives. To train the representative HMM sets, we followed the complex back-end topology [12]. Each digit was characterized with 20 mixtures per state, and the silence and short pause were modeled with 36 mixtures per state. To perform the ML- and MAP-based SFMs, we followed Eqs. (5)-(12) and the five steps introduced in Section 2.3. All the SFM evaluations were conducted in an unsupervised self-compensation mode.

4.2. Experimental results

In this paper, we report the results of 50 testing conditions (ten noise types, SNR 0dB to 20dB) from the Aurora-2 test set. The 50 testing conditions were divided into SetA, SetB and SetC. SetA includes the same four types of noise as that are used in the multi-condition training set, SetB contains four unseen types of noise, and SetC has an additional channel distortion.

4.2.1. SFM without regional information

First, we present the performance of the ML- and MAP-based SFMs without using the regional information. In this set of experiments, only the root node in the EC tree (accordingly only one representative HMM set and one EP tree) was used for feature compensation. The representative HMM set was trained from the entire multi-condition training set in Aurora-2, and the EP tree was constructed based on this HMM set. Table 1 lists the baseline and the ML- and MAP-based SFM results. For the baseline, we recognized the testing utterances without performing feature compensation. For the MAP-based SFM, we used the HP density as a representative. Table 1 results show that the MAP-based SFM clearly provides better performance than the baseline and the ML-based SFM among SetA, SetB, SetC and the overall testing conditions.

Table 1. WER (%) of baseline and SFM without regional information

Test Condition	SetA	SetB	SetC	All
Baseline	5.92	6.69	7.11	6.46
ML-SFM	5.88	6.57	6.97	6.37
MAP-SFM	5.71	6.09	6.74	6.07

4.2.2. SFM with regional information

Table 2 lists the results of the baseline and the ML- and MAP-based SFMs using the regional information. For this set of experiments, a seven-node EC structure was used. To obtain the baseline, each testing utterance was first used to locate one EC node and then recognized using the located representative HMM set. To get the ML- and MAP-based SFM results, we followed the five steps introduced in Section 2.3. Here, the HP density was again used for the MAP-based SFM.

To further investigate the effect of regional information, we compared two HP densities—environment independent HP

(EI-HP) and environment dependent HP (ED-HP). EI-HP is estimated using the EP tree of the root EC node, and ED-HP is calculated using the EP tree of the EC node that is located by the CS procedure. Therefore, ED-HP contains the regional information, but EI-HP doesn't. Table 2 lists the MAP-based SFM results using EI-HP and ED-HP as MAP-SFM (EI-HP) and MAP-SFM (ED-HP). Their implementations are the same; the only difference is the use of two different HP densities.

Table 2. WER (%) of baseline and SFM with regional information

Test Condition	SetA	SetB	SetC	All
Baseline	5.11	5.51	6.42	5.53
ML-SFM	5.10	5.48	6.40	5.51
MAP-SFM (EI-HP)	5.09	5.32	6.21	5.41
MAP-SFM (ED-HP)	5.05	5.29	6.14	5.36

From Table 2, MAP-based SFM outperforms the baseline and the ML-based SFM. Next by comparing Tables 1 and 2, clear performance improvements are achieved for baseline and ML- and MAP-based SFMs by using the regional information. Finally, note that MAP-based SFM using ED-HP can provide better performance than using EI-HP. This result verifies that by incorporating the regional information, we can prepare more suitable prior density for the MAP-based SFM.

4.2.3. MAP-based SFM using four types of prior density

Next, we compare the performances using different types of prior density. Table 3 lists the results of MAP-based SFM using CP, SP, HP, and IP. For this set of experiments, since the seven-node EC tree was used, the regional information was incorporated in the prior density estimates. MAP-SFM (HP) in Table 3 is the same as MAP-SFM (ED-HP) in Table 2. From Table 3, we can see that HP outperforms CP and SP. Moreover, with IP density, which combines multiple knowledge sources, the MAP-based SFM achieves the best performance.

Table 3. WER (%) of MAP-based SFM using different prior densities

Test Condition	SetA	SetB	SetC	All
MAP-SFM (CP)	5.10	5.49	6.33	5.50
MAP-SFM (SP)	5.09	5.48	6.32	5.49
MAP-SFM (HP)	5.05	5.29	6.14	5.36
MAP-SFM (IP)	5.05	5.26	6.09	5.34

4.2.4. MAP-SFM with SMAPLR acoustic model adaptation

In this section, we investigate the compatibility of MAP-based SFM with acoustic model adaptation. Here, we used SMAPLR [7, 9] as an example. We also used an environment structuring framework to enhance the SMAPLR performance. More details about this SMAPLR implementation can be found in our previous study [9]. Table 4 lists the results of SMAPLR alone and MAP-based SFM followed by SMAPLR. By comparing Tables 3 and 4, note that MAP-SFM (IP) provides better performance than SMAPLR in SetB, showing that MAP-based SFM can handle unseen types of noise more effectively than SMAPLR. Meanwhile, we find that SMAPLR has better capability of handling speaker and channel mismatches and achieves better performance in SetA and SetC than MAP-based SFM. However, note that SMAPLR requires higher online computational cost than MAP-based SFM. Finally, the integration of MAP-based SFM and SMAPLR gives better performance than the individual cases, verifying the positive property of MAP-based SFM being compatible with model adaptation to achieve further improvements.

Table 4. WER (%) of integration of MAP-based SFM and SMAPLR

Test Condition	SetA	SetB	SetC	All
SMAPLR	4.96	5.46	5.75	5.32
SFM+SMAPLR	4.95	5.29	5.74	5.24

5. Conclusions

We incorporated the regional information of all training conditions to facilitate suitable prior density estimation for the MAP-based SFM. Three types of prior density were designed: clustered prior (CP), sequential prior (SP), and hierarchical prior (HP) densities; they use knowledge from the training set, the previous testing statistics, and the current testing utterance with the prepared tree structure, respectively. We also derived an integrated prior (IP) density that integrates the three prior densities by a linear combination function. The experimental results indicate that the MAP-based SFM outperforms the baseline and the ML-based SFM. Moreover among all the prior densities in this study, IP density, which incorporates multiple knowledge sources, achieves the best performance. Finally, we found that by integrating SMAPLR model adaptation, MAP-based SFM can obtain further improvements. In this study, the linear combination weights for estimating the IP density are tuned by a set of development data. In future work, we will explore methods that can directly optimize the combination coefficients based on the testing statistics.

6. References

[1] Sankar, A. and Lee, C.-H., "A maximum-likelihood approach to stochastic matching for robust speech recognition," IEEE Trans. Speech Audio Proc., vol. 4, pp. 190-202, 1996.

[2] Gales, M. J. F., "Maximum likelihood linear transformations for HMM-based speech recognition," Technical Report, Cambridge University, 1997.

[3] Li, Y., Erdogan, H., Gao, Y., and Marcheret, E., "Incremental online feature space mllr adaptation for telephony speech recognition," in Proc. ICSLP, pp. 1417-1420, 2002.

[4] Jiang, H., Soong, F. and Lee, C.-H., "Hierarchical stochastic matching for robust speech recognition," in Proc. ICASSP, pp. 217 – 220, 2001.

[5] Lei, X., Hamaker, J. and He, X., "Robust feature space adaptation for telephony speech recognition," in Proc. ICSLP, pp. 773-776, 2006.

[6] Shinoda, K. and Lee, C.-H., "A structural Bayes approach to speaker adaptation," IEEE Trans. Speech Audio Proc., vol. 9, pp. 276-287, 2001.

[7] Siohan, O., Myrvoll, T. A. and Lee, C.-H., "Structural maximum a posteriori linear regression for fast HMM adaptation," Computer Speech and Language, vol. 16, pp. 5-24, 2002.

[8] Chien, J.-T., "Quasi-Bayes linear regression for sequential learning of hidden Markov models," IEEE Trans. Speech Audio Proc., vol. 10, pp. 268-278, 2002.

[9] Tsao, Yu, Isotani, R., Kawai, H. and Nakamura, S., "An environment structuring framework to facilitating suitable prior density estimation for MAPLR on robust speech recognition," in Proc. ISCSLP, pp. 29-32, 2010.

[10] Tsao, Yu, Matsuda, S., Nakamura, S., and Lee, C.-H., "MAP estimation of online mapping parameters in ensemble speaker and speaking environment modeling," in Proc. ASRU, pp. 271-275, 2009.

[11] Tsao, Yu and Lee, C.-H., "An ensemble speaker and speaking environment modeling approach to robust speech recognition," IEEE Trans. on Audio, Speech, and Language Proc., vol. 17, pp. 1025-1037, 2009.

[12] Macho, D., Mauuary, L., Noe, B., Cheng, Y. M., Ealey, D., Jouver, D., Kelleher, H., Pearce, D. and Saadoun, F., "Evaluation of a noise-robust DSR front-end on Aurora databases," in Proc. ICSLP, pp. 17-20, 2002.

[13] Wu, J. and Huo, Q., "Several HKU approaches for robust speech recognition and their evaluation on Aurora connected digit recognition tasks," in Proc. Eurospeech, pp. 21-24, 2003.