Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 47 (2005) 243–264
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Maximum likelihood sub-band adaptation for robust speech recognition

Donglai Zhu [a,b,*], Satoshi Nakamura [a], Kuldip K. Paliwal [a,c], Renhua Wang [b]

[a] *ATR Spoken Language Translation Research Labs, Japan*
[b] *University of Science and Technology of China, China*
[c] *School of Microelectronic Engineering, Griffith University, Australia*

Received 5 January 2004; received in revised form 15 February 2005; accepted 15 February 2005

Abstract

Noise-robust speech recognition has become an important area of research in recent years. In current speech recognition systems, the Mel-frequency cepstrum coefficients (MFCCs) are used as recognition features. When the speech signal is corrupted by narrow-band noise, the entire MFCC feature vector gets corrupted and it is not possible to exploit the frequency-selective property of the noise signal to make the recognition system robust. Recently, a number of sub-band speech recognition approaches have been proposed in the literature, where the full-band power spectrum is divided into several sub-bands and then the sub-bands are combined depending on their reliability. In conventional sub-band approaches the reliability can only be set experimentally or estimated during training procedures, which may not match the observed data and often causes degradation of performance. We propose a novel sub-band approach, where frequency sub-bands are multiplied with weighting factors and then combined and converted to cepstra, which have proven to be more robust than both full-band and conventional sub-band cepstra in our experiments. Furthermore, the weighting factors can be estimated by using maximum likelihood adaptation approaches in order to minimize the mismatch between trained models and observed features. We evaluated our methods on AURORA2 and Resource Management tasks and obtained consistent performance improvement on both tasks.

Keywords: Robust speech recognition; Sub-band; Adaptation

1. Introduction

It is well known that current ASR systems don't work as well as humans. Existing recognizers are extremely sensitive to channel variability and

* Corresponding author. Address: Dept. Computer Science, The University of Hong Kong, Pokfulam road, Hong Kong.
E-mail addresses: dlzhu@cs.hku.hk, donglai.zhu@ustc.edu (D. Zhu), satoshi.nakamura@atr.co.jp (S. Nakamura), k.paliwal@me.gu.edu.au (K.K. Paliwal), rhw@ustc.edu.cn (R. Wang).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.02.006

additive background noise and require careful preprocessing. However, humans are able to achieve excellent recognition accuracy in these cases. Fletcher and his colleagues (Fletcher, 1953; Allen, 1994) suggested that in human auditory perception, the linguistic message gets decoded independently in different frequency sub-bands and the final decoding decision is based on merging the decisions from the sub-bands. Some other experiments studied human performance on filtered (low-pass, high-pass, band-pass and band-reject filtered) speech, in order to gain a better understanding of human speech perception or to find the contribution of different parts of the speech spectrum for perception. Miller and Niely (1955) showed that humans can achieve high recognition rates for narrow-band speech signals. Kryter (1960) showed that humans can combine the perception coming from narrow sub-bands to enhance intelligibility. Riener et al. (1992) and Warren et al. (1995) showed that high intelligibility can be maintained for band-pass filtered speech signals. They also concluded that humans possess processing mechanisms that are able to employ limited spectral regions that can enhance comprehension under difficult listening conditions. Lippmann (1996) showed that humans can recognize speech signals produced by severe band-reject filtering. The results discussed above show that humans can recognize speech signals with limited spectral cues and can easily integrate acoustic cues from different frequency regions for speech perception.

Noise-robust speech recognition has become an important area of research in recent years. In current speech recognition systems, the Mel-frequency cepstrum coefficients (MFCCs) are used as recognition features. When the speech signal is corrupted by narrow-band noise, the entire MFCC feature vector gets corrupted and it is not possible to exploit the frequency-selective property of the noise signal to make the recognition system robust. Recently, a number of sub-band speech recognition approaches have been proposed in the literature, where the full-band power spectrum is divided into several sub-bands and then the sub-bands are combined depending on their reliability. In conventional sub-band approaches the reliability can only be set experimentally or estimated during training procedures, which may not match the observed data and often causes degradation of performance.

Two modes of sub-band approaches have been proposed: parallel sub-band (PSB) and concatenating sub-band (CSB) (Hermansky et al., 1996; Bourlard and Dupont, 1996, 1997; Tibrewala and Hermansky, 1997; Cerisara et al., 1998, 2000; Okawa et al., 1998; Paliwal and Chen, 2000). In both modes, features are extracted from the sub-band spectra independently. If the cepstrum is used as the feature, the extraction procedure is as follows: firstly, the frequency spectrum of the speech signal is divided into sub-bands; secondly, for each sub-band, the spectrum is converted to a cepstrum. There are several free parameters in the procedure, e.g., the number of sub-bands and the frequency boundaries of the sub-bands. They can be adjusted to appropriate values for given tasks (Tibrewala and Hermansky, 1997). For the PSB, the sub-band features are modeled independently, and the likelihood scores of the sub-bands are combined at some segmental levels (phonemes, words and sentences, etc.). Results showed that the PSB can improve the recognition performance for speech signals corrupted by band-limited noises (Bourlard and Dupont, 1997), but may exhibit poorer performance when the additive background noise has wide bands (Tibrewala and Hermansky, 1997). An issue for the PSB is the sub-band recombination. Mainly, three methods have been studied for it. The first is the weighted average method (Bourlard and Dupont, 1996, 1997; Cerisara et al., 1998; Okawa et al., 1998). It produces the overall probability based on an arithmetic or geometric average of the sub-band probabilities. The contribution of each sub-band is weighted by its local signal-to-noise ratio (SNR), or by its reliability. The second is the neural network approach (Hermansky et al., 1996; Tibrewala and Hermansky, 1997). Multi-layer perceptions (MLPs) may be trained to merge the sub-band probabilities to estimate the probability of all possible combinations of the sub-bands. The third is the full combination. The probabilities of different combinations of different sub-bands are combined using a linear method, with each weight

proportional to the relative reliability of a specific set of sub-bands (Morris et al., 1999). An extension of the full combination approach is HMM2 (Weber et al., 2003), where a secondary HMM is inserted at the level of each state of the primary HMM. Paths are generated by the secondary HMM to allocate weights to the sub-bands dynamically in order to extract formant-like structures. Comparatively, our weighting procedure is to attenuate noise corruption and be able to model sub-band dependence. Another issue for the PSB is sub-band synchronization. The recognition procedure is asynchronous within the combining level except that the sub-bands are combined on the state level. In continuous speech recognition, two ways have been studied to implement asynchronous sub-band recognition: two-level dynamic programming (Cerisara et al., 2000) and HMM state composition (Tam and Mak, 2001; Mak and Tam, 2000; Tomlinson et al., 1997). For the CSB, the feature vector is obtained by concatenating the sub-band cepstral vectors. Results showed that the CSB can improve the performance when the additive background noise was band-limited, but may degrade the baseline performance for clean speech signals because the sub-band features lose the correlation between sub-bands (Okawa et al., 1998; Paliwal and Chen, 2000).

We propose in this paper a novel sub-band approach that can consider sub-band dependence. In our approach, bins of log filter-band energy (FBE) in each sub-band are multiplied with a weighting factor depending on the reliability of the sub-band. For each sub-band, zero padding is performed on the log FBE vector lengthening it to the size of the full-band vector, and then it is converted to a cepstral vector by discrete cosine transformation (DCT). Finally, a feature vector is obtained by adding all the sub-band cepstral vectors. Since the DCT has the size of the fullband FBE vector, the feature vector consists of the correlations across the sub-bands. After the weighting factors have been estimated, they should be multiplied on both feature space and model space simultaneously to keep the two spaces consistent. For model space weighting, mean vectors in Gaussian components in the hidden Markov models (HMMs) are converted into log FBE vectors via inverse DCT (IDCT) and then multiplied by the weighting factors. Because DCT is a linear transformation, the weighting factors can be equivalently multiplied to sub-band cepstra and embedded into the framework of HMM. Hence it becomes possible to adopt optimization algorithms to obtain the weighting factors. Since there are only small amounts of weighting factors, an adaptation technique can be used to estimate them. In this paper we used the maximum likelihood adaptation theory, which has been studied on both the feature space and the model space (Sankar and Lee, 1996). We deduced estimation algorithms for the weighting factors on both spaces.

In the weighting procedure on the feature space, sub-band log power values containing more noise tend to be degraded while those containing less noise tend to be enhanced. Therefore, the procedure keeps power conservation on the full band. The weighting procedure on the model space is similar to that on the feature space. Compared with traditional spectral attenuation approaches, where power spectrum of noise is reduced from the whole power spectrum, the weighting procedure not only degrades noise but also enhances clean speech signal. The spectral attenuation approaches have commonly been used for noise suppression. In spectral subtraction (Boll, 1979), additive noise spectrum is estimated and then subtracted from the noisy speech spectrum to recover the clean speech spectrum. In RASTA processing (Hermansky and Morgan, 1994), a band-pass filter is applied to the log spectrum to alleviate the effect of convolutional noise. In the ETSI standard of the advanced front end for distributed speech recognition (Macho et al., 2002), the Wiener filter has been adopted for noise reduction.

In literature, a previous work that applies weights to sub-bands is HMM2 (Weber et al., 2003). In this approach, a secondary HMM is inserted at the level of each state of the primary HMM, estimating local emission probabilities of acoustic feature vectors. Each path through the secondary HMM corresponds to a kind of sub-band segmentation and recombination. Therefore, each band can be considered to be weighted by the corresponding transition probability in the second

HMM dynamically. However, in our approach, the weighting process is implemented by multiplying each sub-band by a weighting factor. Parameters in both approaches can be estimated with maximum likelihood criteria.

This paper is organized as follows. In Section 2 we introduce the weighting procedures on both the feature space and the model space, and also present the experimental results of the weighting procedures. In Section 3, first, we describe the weighting adaptation algorithms, including the feature space adaptation and the model space adaptation. Secondly, we present the experimental results of the adaptation algorithms. Finally, a conclusion is described in Section 4.

2. Sub-band weighting

Sub-band weighting may be performed on both the feature space and the model space. For the weighting procedure on the feature space, weighting factors are multiplied on log FBEs of sub-bands depending on the reliability of the sub-bands; for the procedure on the model space, the mean vectors of Gaussian mixture components in the HMM set are converted back to log FBEs via IDCT, and then weighted in the same way as those of the feature space. In this section we present the two weighting procedures and their experimental results.

2.1. Weighting procedures

In this section we will first review the general MFCC procedure, and then present sub-band weighting procedures on feature space and model space respectively.

2.1.1. General MFCC procedure

In a general feature extraction procedure for the MFCC, the speech signal is converted to spectra via discrete Fourier transformation (DFT), the spectra are passed through a Mel-frequency filter bank to get Mel-frequency FBEs, a logarithm is applied, and finally the MFCC is obtained from the log FBEs via a DCT. The procedure is shown in Fig. 1.

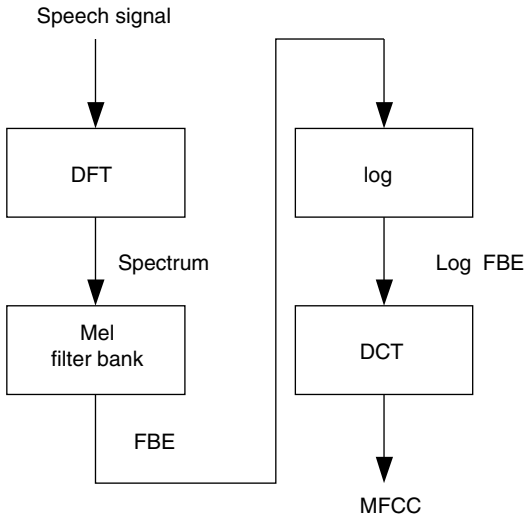

Fig. 1. Standard feature extraction procedure for MFCC.

2.1.2. Weighting procedure on feature space

In the weighting procedure on feature space, filter-bank bins are separated into several sub-bands. A weighting factor is used to multiply the filter-bank bins in each sub-band. For each sub-band, cepstra can be obtained by using DCT. After adding all of the sub-band cepstra we may get the full-band cepstra, which will be the same as the standard MFCCs if the weighting factors are equal. The procedure is as follows (Fig. 2):

1. The log FBE vector $\boldsymbol{f} = \{f_1, f_2, \ldots, f_D\}$ is passed through a sub-band filter, separating it into K sub-bands $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_K\}$, where D is the dimension of the vector. Each sub-band filter has a unit rectangular impulse response, which reserves the log FBE bins within the current sub-band while setting zero for those outside.
2. The log FBE vector for each sub-band is a D-dimension vector: $\boldsymbol{f}_i = \{f_1^i, f_2^i, \ldots, f_D^i\}$, where the elements beyond the frequency domain of current sub-band are zeros. The relationship between the original log FBE vector and the sub-band log FBE vectors is

$$\boldsymbol{f} = \sum_{i=1}^{K} \boldsymbol{f}_i. \tag{1}$$

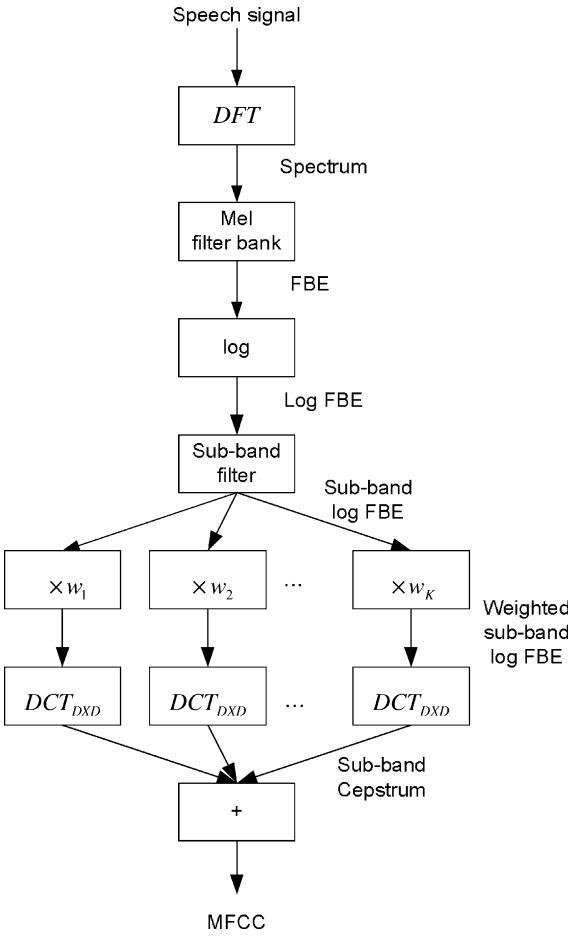

Fig. 2. Feature extraction procedure with sub-band weighting.

3. For each sub-band, the log FBE vector is multiplied by a weight

$$\widehat{\boldsymbol{f}}_i = w_i \boldsymbol{f}_i, \quad 1 \leqslant i \leqslant K, \tag{2}$$

where w_i is the weight for the ith sub-band.

4. For each sub-band, the log FBE vector is converted to a cepstrum vector via DCT

$$\widehat{\boldsymbol{c}}_i = \mathrm{DCT}(\widehat{\boldsymbol{f}}_i), \quad 1 \leqslant i \leqslant K. \tag{3}$$

5. Adding all of the sub-band cepstrum vectors, we get the last MFCC

$$\widehat{\boldsymbol{c}} = \sum_{i=1}^{K} = \widehat{\boldsymbol{c}}_i. \tag{4}$$

If all of the weights are set as 1, the result is the same as the standard MFCC because DCT is a linear transformation. It is proven as follows

$$\begin{aligned}\widehat{\boldsymbol{c}} &= \sum_{i=1}^{K} \widehat{\boldsymbol{c}}_i = \sum_{i=1}^{K} \mathrm{DCT}(\widehat{\boldsymbol{f}}_i) = \sum_{i=1}^{K} \mathrm{DCT}(\boldsymbol{f}_i) \\ &= \mathrm{DCT} \sum_{i=1}^{K} \boldsymbol{f}_i = \mathrm{DCT}(\boldsymbol{f}) = \boldsymbol{c}, \end{aligned} \tag{5}$$

where $\boldsymbol{c}$ is the MFCC vector obtained by the standard procedure.

Because DCT is a linear transformation, weighting on the sub-band log FBE vector is identical with that on the sub-band cepstrum vector

$$\begin{aligned}\widehat{\boldsymbol{c}} &= \sum_{i=1}^{K} \widehat{\boldsymbol{c}}_i = \sum_{i=1}^{K} \mathrm{DCT}(w_i \boldsymbol{f}_i) \\ &= \sum_{i=1}^{K} w_i \mathrm{DCT}(\boldsymbol{f}_i) = \sum_{i=1}^{K} w_i \boldsymbol{c}_i, \end{aligned} \tag{6}$$

where $\boldsymbol{c}_i$ is the sub-band cepstrum vector without weight.

Because w_i is the weight coefficient, it has the following constraint

$$\sum_{i=1}^{K} w_i = K. \tag{7}$$

2.1.3. Weighting procedure on model space

In the weighting procedure on model space, weights are performed on the mean vectors of Gaussian mixture components in HMMs. If we adopt MFCC as the feature vector in a speech recognition system, the mean vectors are values in the cepstrum domain. We may convert them back to log FBEs via IDCT. Then we may divide them into sub-bands and weight them. Fig. 3 illustrates the procedure as follows,

1. The mean vector $\boldsymbol{\mu} = \{\mu_0, \mu_1, \ldots, \mu_{N-1}\}$ is converted back to log FBE vector $\boldsymbol{m} = \{m_1, m_2, \ldots, m_N\}$ via IDCT

$$\boldsymbol{m} = \mathrm{IDCT}(\boldsymbol{\mu}). \tag{8}$$

2. Log FBE vector $\boldsymbol{m}$ is divided into sub-bands $\{\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_K\}$, and they are converted into cepstrum vectors $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K\}$ via DCT

$$\boldsymbol{\mu}_i = \mathrm{DCT}(\boldsymbol{m}_i), \quad 1 \leqslant i \leqslant K. \tag{9}$$

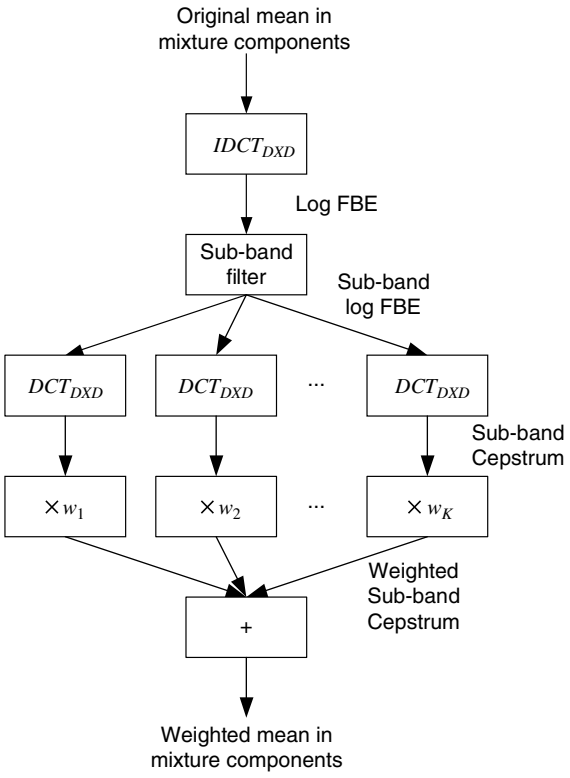

Fig. 3. Sub-band weighting on mean vectors.

3. The sub-band cepstrum vectors are multiplied by weighting factors

$$\widehat{\boldsymbol{\mu}}_i = w_i \boldsymbol{\mu}_i, \quad 1 \leqslant i \leqslant K. \tag{10}$$

4. All of the sub-band cepstrum vectors are added and the adapted mean vector is

$$\widehat{\boldsymbol{\mu}} = \sum_{i=1}^{K} \widehat{\boldsymbol{\mu}}_i. \tag{11}$$

2.2. Experimental results of sub-band weighting

In this experiment we investigate the performance of the weighting approaches. Speech data with additive noise signals were adopted as test data. It is assumed that the noise signals are band-limited in the frequency domain and that the scale of the band is known. Hence, we may artificially set the weights to restrict the noisy parts. One setting method is to let the weights for this band equal zero and the others equal one.

2.2.1. Experimental setup

We investigated the performance of the approach on the DARPA Resource Management (RM) database. It is a collection of speaker-dependent and speaker independent continuous speech data for both training and testing speech recognition systems (Price et al., 1988). The speech comes from a variety of speakers of North American English collected by Texas Instruments in quiet environments using close talking and head mounted microphones. The subject of the material is the management of naval resources. The vocabulary includes approximately 1000 words. The corpus is split into speaker-independent and speaker-dependent sections. In this paper we used its speaker-independent section. The speaker-independent training set consists of 2880 sentences from 72 speakers. The test set is the Feb89 set, consisting of 300 sentences from 10 new speakers. The word pair grammar was used in the decoder. In our experiment there are 49 monophone HMMs. Each HMM consists of three states, four mixtures per state except for the one-state "sp" model. The features were composed of 13 cepstra including a zero-order cepstrum, and their delta and accelerator coefficients.

We added a band-limited noise signal to the test set. The amplitude of the noise signal was set to 1.0E5 (100 dB). The cepstra are converted to log filter-banks via IDCT, so there are 13 bins in the log filter-banks. We assumed that the band scale of the noise signal consisted of 3 bins and added the noise signal in 4 ways: 1–3 bins, 4–6 bins, 7–9 bins and 10–12 bins. On both clean test data and noisy test data, we compared four features, such as full band cepstra (FB), concatenated sub-band cepstra (CSB), concatenated clean sub-band cepstra (CCSB), and weighted sub-band cepstra (WSB). FB is the general full band MFCC. For CSB, we divided the 13 bins in the log filter-banks into 4 sub-bands (1–3 bins, 4–6 bins, 7–9 bins and 10–13 bins). Hence, the noise signal was added to only one sub-band in each noisy case. Independent DCTs were performed on the sub-bands to obtain

the sub-band cepstra. No truncation was performed on these sub-band cepstral vectors, i.e., the sub-band cepstrum vectors have the same dimension as the sub-band log FBE vectors. The dimension of the first three DCTs was 3, while that of the last was 4. Lastly the sub-band cepstra were concatenated into the cesptral vector. For CCSB, only the clean sub-bands in CSB were concatenated into the full cepstrum. WSB was our weighted sub-band. Here, we set 13 sub-bands, i.e., each bin of the log filter-bank was regarded as one sub-band. We set the weighting factors for the noise-added bins to zero, while the others were set to one. The weighting factors were multiplied within both feature space and model space to keep consistency.

2.2.2. Results

Table 1 shows the accuracy of different features in different conditions. For clean test data, FB achieved better performance than CSB. This is consistent with the results in (Paliwal and Chen, 2000). For band-limited noisy test data, FB degraded the performance. The noise signal added to lower bins had more degradation because it confused the formants of the speech signal. CSB improved the performance over FB when the noise signal was impaired severely but degraded the performance when the impairment became weak. Its average accuracy was lower than that of FB. CCSB was a little worse than CSB. This suggested that losing the information of the noisy sub-band might be more harmful than keeping the noisy sub-band there. However, it was suggested in missing feature approaches (Cooke et al., 2001) that data marginalization is superior over data imputation. The reason for the contradictive results is as follows. Cooke used a 64-channel filter-bank while we used a 13-channel filter-bank, i.e., data resolution in Cooke's experiments is higher than that in ours. Therefore, CCSB is inferior to CSB due to data sparsity in our experiments while data marginalization was superior to data imputation due to data reliability in Cooke's experiments. WSB achieved much better performance than both CSB and FB in the noisy case, which was even similar to that in the clean case.

3. Sub-band weighting adaptation

One problem in the sub-band weighting approach is the estimation of the weights. Because there are only a small number of weights, they may be estimated from a small amount of adaptation data. Hence, we adopt an adaptation technique to solve the problem in this paper. Widely used adaptation methods like MLLR (Legetter and Woodland, 1995) and MAP (Gauvain and Lee, 1991) have proved to be capable of effectively improving recognition performance. We estimate the weights by using maximum likelihood criteria (Sankar and Lee, 1996). In this section, we will first review the maximum likelihood adaptation framework, and then present the maximum likelihood estimation for sub-band weighting factors on both feature space and model space.

3.1. Maximum likelihood adaptation framework

The recognition problem in ASR is: Given a trained model set $\boldsymbol{\Phi}_X = \{\phi_{x_i}\}$, where ϕ_{x_i} is the ith

Table 1
Accuracy (%) of different features for clean test data and noisy test data

Clean	FB	88.76				
	CSB	82.24				
Noisy	Noisybins	1–3	4–6	7–9	10–12	Average
	FB	71.73	84.86	86.67	87.31	82.64
	CSB	80.19	81.55	81.39	82.60	81.43
	CCSB	77.93	79.82	80.71	80.75	79.80
	WSB	86.51	86.47	86.55	86.75	86.57

FB: full band; CSB: concatenated sub-band; CCSB: concatenated clean sub-band; WSB: weighted sub-band.

model, and test data $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$, we need to recognize a corresponding sequence of events $\boldsymbol{W} = \{W_1, W_2, \ldots, W_L\}$. When there's a mismatch between $\boldsymbol{\Phi}_X$ and $\boldsymbol{Y}$, errors are caused in the recognized sequence $\boldsymbol{W}$. In most ASR systems, we use a maximum a posterior (MAP) decoder

$$\begin{aligned}\widehat{\mathbf{W}} &= \arg\max_{w} p(\mathbf{W}|\mathbf{Y}, \boldsymbol{\Phi}_X) \\ &= \arg\max_{w} p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\Phi}_X) p(\mathbf{W}|\boldsymbol{\Phi}_X). \end{aligned} \tag{12}$$

The motivation of the adaptation is to minimize the mismatch in order to improve the performance. It can be performed on either feature space or model space.

In feature space, we assume that training data $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ is mapped into the sequence of observations $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$. If the distortion is invertible, we may map $\boldsymbol{Y}$ back to $\boldsymbol{X}$ with an inverse function F_v

$$\boldsymbol{X} = F_v(\boldsymbol{Y}), \tag{13}$$

where v are the parameters of the inverse distortion function.

Correspondingly, in model space, we may map the trained model $\boldsymbol{\Phi}_X$ into the model $\boldsymbol{\Phi}_Y$ that matches test data with a transformation G_η

$$\boldsymbol{\Phi}_Y = G_\eta(\boldsymbol{\Phi}_X), \tag{14}$$

where η are the parameters of the transformation function.

One method to minimize the mismatch is to estimate parameters v or η and $\boldsymbol{W}$, to maximize the joint likelihood.

In feature space,

$$(\widehat{v}, \widehat{\mathbf{W}}) = \arg\max_{v, \mathbf{W}} p(\mathbf{Y}|\mathbf{W}, v, \boldsymbol{\Phi}_X) p(\mathbf{W}|\boldsymbol{\Phi}_X). \tag{15}$$

In model space,

$$(\widehat{\eta}, \widehat{\boldsymbol{W}}) = \arg\max_{\eta, \boldsymbol{W}} p(\boldsymbol{Y}|\boldsymbol{W}, \eta, \boldsymbol{\Phi}_X) p(\boldsymbol{W}|\boldsymbol{\Phi}_X). \tag{16}$$

The joint maximization over variables v or η and $\boldsymbol{W}$ may be done iteratively by keeping v or η fixed and maximizing over $\boldsymbol{W}$, and then keeping $\boldsymbol{W}$ fixed and maximizing over v or η.

The process of estimating $\boldsymbol{W}$ is similar to the standard training procedure in ASR. Here, we consider the estimation of v or η. Because $\boldsymbol{W}$ is a fixed value, we remove the dependence on $\boldsymbol{W}$, and write (15) and (16) as

$$\widehat{v} = \arg\max_{v} p(\boldsymbol{Y}|v, \boldsymbol{\Phi}_X), \tag{17}$$

and

$$\widehat{\eta} = \arg\max_{\eta} p(\boldsymbol{Y}|\eta, \boldsymbol{\Phi}_X). \tag{18}$$

Let $\boldsymbol{S} = \{S_1, S_2, \ldots, S_T\}$ be the set of all possible state sequences, and $\boldsymbol{K} = \{k_1, k_2, \ldots, k_T\}$ be the set of all mixture component sequences. Then (17) and (18) can be written as

$$\begin{aligned}\widehat{v} &= \arg\max_{v} p(\boldsymbol{Y}|v, \boldsymbol{\Phi}_X) \\ &= \arg\max_{v} \sum_{\boldsymbol{S}} \sum_{\boldsymbol{K}} p(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{K}|v, \boldsymbol{\Phi}_X), \end{aligned} \tag{19}$$

and

$$\begin{aligned}\widehat{\eta} &= \arg\max_{\eta} p(\boldsymbol{Y}|\eta, \boldsymbol{\Phi}_X) \\ &= \arg\max_{\eta} \sum_{\boldsymbol{S}} \sum_{\boldsymbol{K}} p(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{K}|\eta, \boldsymbol{\Phi}_X). \end{aligned} \tag{20}$$

3.1.1. Feature space adaptation

The EM algorithm is used to estimate $\widehat{v}$. The EM algorithm is a two-step iterative procedure (Dempster et al., 1977). In the first step, called the expectation step (E step), we compute the auxiliary function

$$\begin{aligned}Q(v, v') &= E\{\log p(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{K}|v', \boldsymbol{\Phi}_X)|\boldsymbol{Y}, v, \boldsymbol{\Phi}_X\} \\ &= \sum_{\boldsymbol{S}} \sum_{\boldsymbol{K}} p(\boldsymbol{S}, \boldsymbol{K}|\boldsymbol{Y}, v, \boldsymbol{\Phi}_X) \\ &\quad \times \log p(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{K}|v', \boldsymbol{\Phi}_X). \end{aligned} \tag{21}$$

In the second step, called the maximization step (M step), we find the value of $\widehat{v}$ that maximizes the auxiliary function

$$\widehat{v} = \arg\max_{v} Q(v, v'). \tag{22}$$

It is proved that if, $Q(v, v') \geqslant Q(v, v)$ then, $p(\boldsymbol{Y}|v', \boldsymbol{\Phi}_X) \geqslant p(\boldsymbol{Y}|v, \boldsymbol{\Phi}_X)$. Thus, iteratively applying the E and M steps guarantees that the likelihood is nondecreasing. The iterations are continued until the increase of the likelihood is less than some predetermined threshold.

We assume that the inverse distortion function maps each frame of $\boldsymbol{Y}$ onto the corresponding frame of $\boldsymbol{X}$

$$\boldsymbol{x}_t = f_v(\boldsymbol{y}_t). \tag{23}$$

The auxiliary function may be written as

$$Q(v, v') = \sum_{\boldsymbol{S}} \sum_{\boldsymbol{K}} p(\boldsymbol{S}, \boldsymbol{K}|\boldsymbol{Y}, v, \boldsymbol{\Phi}_X) \times \log \prod_{t=1}^{T} a_{S_{t-1},S_t} C_{S_t,k_t} p_y(\boldsymbol{y}_t|S_t, k_t, v', \boldsymbol{\Phi}_X), \tag{24}$$

where $p_y(\boldsymbol{y}_t|s_t, k_t, v', \boldsymbol{\Phi}_X)$ is the probability density function of the random variable $\boldsymbol{y}_t$, whose relationship with the probability density function of the random variable $\boldsymbol{x}_t$ is

$$p_{\boldsymbol{y}}(\boldsymbol{y}_t|S_t, k_t, v', \boldsymbol{\Phi}_X) = \frac{p_x(f_{v'}(\boldsymbol{y}_t)|S_t, k_t, v', \boldsymbol{\Phi}_X)}{|J_{v'}(\boldsymbol{y}_t)|}, \tag{25}$$

where $J_{v'}(\boldsymbol{y}_t)$ is the Jacobian matrix whose, (i,j)th element is

$$J_{v'}(\boldsymbol{y}_t)_{i,j} = \frac{\partial y_{t,i}}{\partial f_{v',j}(\boldsymbol{y}_i)}, \tag{26}$$

where $y_{t,i}$, is the ith element of $\boldsymbol{y}_t$, and $f_{v',j}(\boldsymbol{y}_t)$, is the jth element of $f_{v'}(\boldsymbol{y}_t)$.

In general, $p_{\boldsymbol{x}}(\boldsymbol{x}_t|s_t, k_t, v', \boldsymbol{\Phi}_X)$ is defined as a Gaussian distribution, so the auxiliary function may be written as

$$Q(v, v') = \sum_{\boldsymbol{S}} \sum_{\boldsymbol{K}} p(\boldsymbol{S}, \boldsymbol{K}|\boldsymbol{Y}, v, \boldsymbol{\Phi}_X) \times \log \prod_{t=1}^{T} a_{S_{t-1},S_t} C_{S_t,k_t} \times \frac{N(f_{v'}(\boldsymbol{y}_t); \boldsymbol{v}_{S_t,k_t}, \boldsymbol{\Sigma}_{S_t,k_t})}{|J_{v'}(\boldsymbol{y}_t)|}. \tag{27}$$

It may also be written as

$$Q(v, v') = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} p(S_{t-1} = i, S_t = j|\boldsymbol{Y}, v, \boldsymbol{\Phi}_X) \times \log a_{ij} + \sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} p(S_t = j, c_t = k|\boldsymbol{Y}, v, \boldsymbol{\Phi}_X) \times \log c_{jk} + \sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} p(S_t = j, c_t = k|\boldsymbol{Y}, v, \boldsymbol{\Phi}_X) \times \log N(f_{v'}(\boldsymbol{y}_t); \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) - \sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} p(S_t = j, c_t = k|\boldsymbol{Y}, v, \boldsymbol{\Phi}_X) \times \log |J_{v'}(\boldsymbol{y}_t)|. \tag{28}$$

We are only interested in the terms involving v', so the auxiliary function may be written as

$$Q(v, v') = \sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} p(S_t = i, c_t = k|\boldsymbol{Y}, v, \boldsymbol{\Phi}_X) \times \left\{ -\frac{1}{2} [f_{v'}(\boldsymbol{y}_t) - \boldsymbol{\mu}_{jk}]^{\mathrm{T}} \boldsymbol{\Sigma}_{jk}^{-1} [f_{v'}(\boldsymbol{y}_t) - \boldsymbol{\mu}_{jk}] - \log |\boldsymbol{J}_{v'}(\boldsymbol{y}_t)| \right\}. \tag{29}$$

In order to find the maximum of the auxiliary function, we may differentiate it with respect to v' and solve for its zeros

$$\frac{\partial}{\partial v'} Q(v, v') = 0. \tag{30}$$

In the sub-band weighting algorithm, the inverse distortion function from the observed data to the training data is as follows

$$\boldsymbol{c}' = f_w(\boldsymbol{c}). \tag{31}$$

According to Eq. (6), the function may be represented as

$$\boldsymbol{c} = \boldsymbol{C}\boldsymbol{i} \tag{32}$$

$$\boldsymbol{c}' = \boldsymbol{C}\boldsymbol{w}, \tag{33}$$

where $\boldsymbol{i}$ is a unit vector. $\boldsymbol{C}$ is the cepstrum matrix whose column vectors are the sub-band cepstrums $\boldsymbol{C} = \{\boldsymbol{c}_1, c_2, \ldots, \boldsymbol{c}_K\}$. K is the number of sub-bands. $\boldsymbol{w}$ is the weight vector $\boldsymbol{w} = \{w_1, w_2, \ldots, w_K\}$.

Because f_w is not a direct function, it is difficult to calculate the following Jacobian matrix

$$J_w(\boldsymbol{c})_{ij} = \frac{\partial c_i}{\partial c'_j}. \tag{34}$$

Hence, in order to make the algorithm feasible, we make two assumptions: (1) the estimated values of the weights are close to 1, and (2) the cepstral coefficients are statistically independent. The reason for the first assumption is that the weights used to multiply log FBEs are intended to restore the noise-corrupted shapes. As a result, in order to

avoid unexpected distortion of the log FBEs, the values of the weights should be close to 1. This will be illustrated in later experiments. The second assumption has been implicitly made due to the diagonal covariance matrices in Gaussian components. Then we may get

$$\frac{\partial c_i}{\partial c'_j} \approx \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases} \tag{35}$$

Then the determinant of the Jacobian matrix is

$$|J_w(\boldsymbol{c})| = 1. \tag{36}$$

Then, Eq. (29) may be written as

$$Q(\boldsymbol{w}, \boldsymbol{w}') = \sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} p(S_t = j, c_t = k | \boldsymbol{c}_t, \boldsymbol{w}, \boldsymbol{\Phi}_X) \times \left\{ -\frac{1}{2} \left[\boldsymbol{C}_t \boldsymbol{w}' - \boldsymbol{\mu}_{jk} \right]^{\mathrm{T}} \boldsymbol{\Sigma}_{jk}^{-1} \left[\boldsymbol{C}_t \boldsymbol{w}' - \boldsymbol{\mu}_{jk} \right] \right\}. \tag{37}$$

Defining

$$p(S_t = j, c_t = k | \boldsymbol{c}_t, \boldsymbol{w}, \boldsymbol{\Phi}_X) = \zeta_t(t_j, k). \tag{38}$$

Then (37) is written as

$$Q\boldsymbol{w}, \boldsymbol{w}' = \sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} \zeta_t(j,k) \times \left\{ -\frac{1}{2} \left[\boldsymbol{C}_t \boldsymbol{w}' - \boldsymbol{\mu}_{jk} \right]^{\mathrm{T}} \boldsymbol{\Sigma}_{jk}^{-1} \left[\boldsymbol{C}_t \boldsymbol{w}' - \boldsymbol{\mu}_{jk} \right] \right\}. \tag{39}$$

Meanwhile, weights have the constraint of (7), which may be written as

$$\boldsymbol{i}^{\mathrm{T}} \boldsymbol{w} = K. \tag{40}$$

Defining the objective function as

$$F(\boldsymbol{w}, \boldsymbol{w}') = Q(\boldsymbol{w}, \boldsymbol{w}') + \lambda(\boldsymbol{i}^{\mathrm{T}} \boldsymbol{w}' - K) = \sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} \zeta_t(j,k) \times \left\{ -\frac{1}{2} \left[\boldsymbol{C}_t \boldsymbol{w}' - \boldsymbol{\mu}_{jk} \right]^{\mathrm{T}} \boldsymbol{\Sigma}_{jk}^{-1} \left[\boldsymbol{C}_t \boldsymbol{w}' - \boldsymbol{\mu}_{jk} \right] \right\} + \lambda(\boldsymbol{i}^{\mathrm{T}} \boldsymbol{w}' - K). \tag{41}$$

In order to find the maximum of the objective function, we may differentiate it with respect to $\boldsymbol{w}'$ and solve for its zeros using Lagrange's multiplier λ

$$\nabla_{w'} F(\boldsymbol{w}, \boldsymbol{w}') = 0. \tag{42}$$

Then we may get the following solution

$$-\sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{\mathrm{T}} \zeta_t(j,k) \boldsymbol{C}_t^T \boldsymbol{\Sigma}_{jk}^{-1} \left[\boldsymbol{C}_t \boldsymbol{w}' - \boldsymbol{\mu}_{jk} \right] + \lambda \boldsymbol{i} = 0. \tag{43}$$

The solution is derived as

$$\boldsymbol{w}' = \boldsymbol{X}^{-1}(\boldsymbol{y} + \lambda \boldsymbol{i}), \tag{44}$$

where

$$\boldsymbol{X} = \sum_{j} \sum_{k} \sum_{t} \zeta_t(j,k) \boldsymbol{C}_t^{\mathrm{T}} \boldsymbol{\Sigma}_{jk}^{-1} \boldsymbol{C}_t, \tag{45}$$

$$\boldsymbol{y} = \sum_{j} \sum_{k} \sum_{t} \zeta_t(j,k) \boldsymbol{C}_t^{\mathrm{T}} \boldsymbol{\Sigma}_{jk}^{-1} \boldsymbol{\mu}_{jk}. \tag{46}$$

λ may be obtained by combining (40) and (44)

$$\lambda = \frac{K - \boldsymbol{i}^{\mathrm{T}} \boldsymbol{X}^{-1} \boldsymbol{y}}{\boldsymbol{i}^{\mathrm{T}} \boldsymbol{X}^{-1} \boldsymbol{i}}. \tag{47}$$

3.1.2. Model space adaptation

In feature space adaptation, we assumed that the determinants of the Jacobian matrices equal one, which may cause incorrectness during the adaptation procedure. In this section we discuss the adaptation on model space, which will not require such an assumption. For model space adaptation, the required assumption is that in the features there are cepstrums including the zero-order cepstrum. Here, the EM algorithm is used to estimate $\widehat{\eta}$. In the E step, we calculate the following auxiliary function

$$Q(\eta, \eta') = E\{\log p(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{K} | \eta', \boldsymbol{\Phi}_X) | \boldsymbol{Y}, \eta, \boldsymbol{\Phi}_X\} = \sum_{S} \sum_{K} p(\boldsymbol{S}, \boldsymbol{K} | \boldsymbol{Y}, \eta, \boldsymbol{\Phi}_X) \times \log p(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{K} | \eta', \boldsymbol{\Phi}_X). \tag{48}$$

In the M step, we estimate $\widehat{\eta}$ to maximize the auxiliary function

$$\widehat{\eta} = \arg\max_{\eta} Q(\eta, \eta'). \tag{49}$$

We assume that adaptation is performed on the mean vectors and covariance matrices in the models trained from training data $\boldsymbol{X}$. The adaptation transformations are as follows

$$\boldsymbol{\mu}^Y = g_\eta^\mu(\boldsymbol{\mu}^X), \tag{50}$$

$$\boldsymbol{\Sigma}^Y = g_\eta^\Sigma(\boldsymbol{\Sigma}^X). \tag{51}$$

Then, the probability density function of the random variable $\boldsymbol{y}_t$ is

$$p_y(\boldsymbol{y}_t|S_t, k_t, \eta', \boldsymbol{\Phi}_X) = N(\boldsymbol{y}_t; g_{\eta'}^\mu(\boldsymbol{\mu}_{S_t,k_t}), g_{\eta'}^\Sigma(\Sigma_{S_t,k_t})). \tag{52}$$

We are only interested in the terms involving η', so the auxiliary function may be written as

$$\begin{aligned} Q(\eta, \eta') = \sum_{j=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T} p(S_t = j, c_t = k|\boldsymbol{Y}, \eta, \boldsymbol{\Phi}_X) \\ \times \left\{ -\frac{1}{2}\left[\boldsymbol{y}_t - g_{\eta'}^\mu(\boldsymbol{\mu}_{jk})\right]^{\mathrm{T}} g_{\eta'}^\Sigma(\boldsymbol{\Sigma}_{jk})^{-1} \right. \\ \left. \times \left[\boldsymbol{y}_t - g_{\eta'}^\mu(\boldsymbol{\mu}_{jk})\right] \right\}. \end{aligned} \tag{53}$$

In order to find the maximum of the auxiliary function, we may differentiate it with respect to η' and solve for its zeros

$$\frac{\partial}{\partial \eta'} Q(\eta, \eta') = 0. \tag{54}$$

According to the procedure in Section 2.1.3, the model space transformation procedure may be represented by the following equations

$$\boldsymbol{\mu} = \boldsymbol{U}\boldsymbol{i}, \tag{55}$$

$$\boldsymbol{\mu}' = \boldsymbol{U}\boldsymbol{w}, \tag{56}$$

where $\boldsymbol{U}$ is the cepstrum matrix whose columns are cepstrum vectors of sub-bands $\boldsymbol{U} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K]$. $\boldsymbol{w}$ is the weight vector $\boldsymbol{w} = [w_1, w_2, \ldots, w_K]^{\mathrm{T}}$.

Then, (53) may be written as

$$\begin{aligned} Q(w, w') = \sum_{j=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T} \zeta_t(j,k) \\ \times \left\{ -\frac{1}{2}[\boldsymbol{y}_t - \boldsymbol{U}_{jk}w']^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}[\boldsymbol{y}_t - \boldsymbol{U}_{jk}\boldsymbol{w}'] \right\}. \end{aligned} \tag{57}$$

Meanwhile, weights have the constraint in (40).

Defining the objective function as

$$\begin{aligned} F(\boldsymbol{w}, \boldsymbol{w}') &= Q(\boldsymbol{w}, \boldsymbol{w}') + \lambda(\boldsymbol{i}^{\mathrm{T}}\boldsymbol{w}' - K) \\ &= \sum_{j=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T} \zeta_t(j,k) \\ &\quad \times \left\{ -\frac{1}{2}[\boldsymbol{y}_t - \boldsymbol{U}_{jk}\boldsymbol{w}']^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}[\boldsymbol{y}_t - \boldsymbol{U}_{jk}\boldsymbol{w}'] \right\} \\ &\quad + \lambda(\boldsymbol{i}^{\mathrm{T}}\boldsymbol{w}' - K). \end{aligned} \tag{58}$$

In order to find the maximum of the objective function, we may differentiate it with respect to $\boldsymbol{w}'$ and solve for its zeros

$$\nabla_{w'} F(\boldsymbol{w}, \boldsymbol{w}') = 0. \tag{59}$$

Then we may get

$$\sum_{j=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T} \zeta_t(j,k)\boldsymbol{U}_{jk}^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}[\boldsymbol{y}_t - \boldsymbol{U}_{jk}\boldsymbol{w}'] + \lambda\boldsymbol{i} = 0. \tag{60}$$

And the solution is

$$\boldsymbol{w}' = \boldsymbol{X}^{-1}(y + \lambda\boldsymbol{i}), \tag{61}$$

where

$$\boldsymbol{X} = \sum_j\sum_k\sum_t \zeta_t(j,k)\boldsymbol{U}_{jk}^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}\boldsymbol{U}_{jk}, \tag{62}$$

$$\boldsymbol{y} = \sum_j\sum_k\sum_t \zeta_t(j,k)\boldsymbol{U}_{jk}^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}\boldsymbol{y}_t. \tag{63}$$

λ may be obtained by combining (40) and (61), and is equal to (47).

3.1.2.1. Multiple class case. If we separate the Gaussian components in the model set into L classes and multiply each different class with a different weighing factor, Eq. (56) will be

$$\boldsymbol{\mu}_c' = \boldsymbol{U}_c\boldsymbol{w}_c \qquad 1 \leqslant c \leqslant L. \tag{64}$$

The weighting factors have the same constraint as (40) for each class

$$\boldsymbol{i}^{\mathrm{T}}\boldsymbol{w}_c = K. \tag{65}$$

In this case, the objective function will be

$$F(\boldsymbol{w},\boldsymbol{w}') = \sum_{c=1}^{L}\left[\sum_{j,k\in c}\sum_{t=1}^{\mathrm{T}}\zeta_t(j,k) \times\left\{-\frac{1}{2}[\boldsymbol{y}_t-\boldsymbol{U}_{jk}\boldsymbol{w}'_c]^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}[\boldsymbol{y}_t-\boldsymbol{U}_{jk}\boldsymbol{w}'_c]\right\} +\lambda(\boldsymbol{i}^{\mathrm{T}}\boldsymbol{w}'_c-\boldsymbol{K})\right]. \quad (66)$$

Differentiating it with respect to $\boldsymbol{w}'_c$ and solving for its zeros

$$\nabla_{\boldsymbol{w}'_c}F(\boldsymbol{w},\boldsymbol{w}')=0. \quad (67)$$

We may obtain

$$\sum_{j,k\in c}\sum_{t=1}^{T}\zeta_t(j,k)\boldsymbol{U}_{jk}^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}[\boldsymbol{y}_t-\boldsymbol{U}_{jk}\boldsymbol{w}'_c]+\lambda\boldsymbol{i}=0. \quad (68)$$

The solution will be

$$\boldsymbol{w}'_c=\boldsymbol{X}_c^{-1}(\boldsymbol{y}_c+\lambda_c\boldsymbol{i}), \quad (69)$$

where

$$\boldsymbol{X}_c=\sum_{j,k\in c}\sum_{t=1}^{T}\zeta_t(j,k)\boldsymbol{U}_{jk}^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}\boldsymbol{U}_{jk}, \quad (70)$$

$$\boldsymbol{y}_c=\sum_{j,k\in c}\sum_{t=1}^{T}\zeta_t(j,k)\boldsymbol{U}_{jk}^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}\boldsymbol{y}_t, \quad (71)$$

λ_c may be obtained by combining (65) and (69)

$$\lambda_c=\frac{K-\boldsymbol{i}^{\mathrm{T}}\boldsymbol{X}_c^{-1}\boldsymbol{y}_c}{\boldsymbol{i}^{\mathrm{T}}\boldsymbol{X}_c^{-1}\boldsymbol{i}}. \quad (72)$$

3.1.2.2. Unconstrained weighting case. In model space adaptation, the weighting procedure on mean vectors may be regarded as a type of transformation on them. So it is possible to remove the constraint on the weights. The solution is slightly different from the previous one. Differentiating (57) with respect to $\boldsymbol{w}'$ and solving for its zeros

$$\nabla_{w'}Q(\boldsymbol{w},\boldsymbol{w}')=0, \quad (73)$$

we obtain

$$\sum_{j=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T}\zeta_t(j,k)\boldsymbol{U}_{jk}^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}[\boldsymbol{y}_t-\boldsymbol{U}_{jk}\boldsymbol{w}']=0. \quad (74)$$

The solution will then be

$$\boldsymbol{w}'=\boldsymbol{X}^{-1}\boldsymbol{y}, \quad (75)$$

where $\boldsymbol{X}$, $\boldsymbol{y}$, are equal to (62) and (63) respectively.

3.2. Experimental results of sub-band weighting adaptation

We evaluated the maximum likelihood sub-band adaptation algorithms on two tasks: the Aurora2 task and the Resource Management task. On both tasks, we investigated the convergent property of the adaptation procedures; and the relationship of accuracy with the number of sub-bands, adaptation data and model classes.

3.2.1. Aurora2 task evaluation

3.2.1.1. Experimental setup. The Aurora2 corpus was created for research on distributed speech recognition under noisy environments (Hirsch and Pearce, 2000). It is composed of connected digits from the clean TIDIGITS database (Leonard, 1984). The data were pre-filtered according to the frequency characteristics of common telecommunication channels (G.712 or MIRS) and were artificially supplied with realistic noises at six different signal-to-noise (SNR) ratios ranging from 20 to −5 dB at 5 dB steps. In its baseline, there are two training models: clean training and multi-condition training, and three test sets defined to evaluate recognition technologies under matched and unmatched noises, and matched and unmatched channel characteristics. Every test set was separated into several sub-sets. Each of them was added with a certain noise at a certain SNR, including 1001 utterances. In all of our adaptation experiments, we divided each sub-set into two parts: adaptation data and test data. The adaptation procedure was performed on the previous, and the performance was evaluated on the latter. The performance measure for one noise or one test set was defined as the average over SNRs between 0 and 20 dB in the corresponding cases. Feature vectors were composed of 13 cepstrum coefficients including the zero-order coefficient, and their delta and acceleration coefficients. The digit models have 16 states with 3 Gaussians per state. The silence model has 3 states with 6 Gaussians per state.

Also, a one-state short pause model is used and is tied with the middle state of the silence model.

3.2.1.2. Weighting on signals. This experiment investigated the performance of weighting on signals in feature space adaptation. A clean speech utterance (FAK_521Z9A.08, in the clean1 sub-set in AURORA2 Test set A) was artificially supplied with noise signals that had special spectrums and used for adaptation. We assumed that the spectrum of the added noise was a flat value A on high half frequency and was zero on low half frequency, which is shown in Fig. 4(a). Here A took three values: 1E3, 1E4, or 1E5 (60 dB, 80 dB, or 100 dB). We added the noisy spectrum amplitude to that of the speech signal and got the spectrum amplitude of the simulated noisy speech signal. Fig. 4(b) shows the spectrum amplitudes of a frame of the clean speech signal, and its noisy speech signals at different SNRs. Fig. 4(c) shows their log filter-bank amplitudes, where there are 23 bins. It is shown that the amplitudes of the high bins increased with the enhancement of the added noise. The range of change on frequency is less than half because it is on the Mel-frequency scale, which compressed low linear frequencies. The number of sub-bands is set to two, so there were two weights to be estimated. The initial models were the clean HMMs in the baseline. Ten iterations were performed in the feature space adaptation procedure. Fig. 4(d) shows the estimated weights in four noisy cases. $w(1)$ is the weight on the low-frequency band and $w(2)$ is that on the high-frequency band. $w(2)$ decreased and $w(1)$ increased with the enhancement of added noise, which indicated that the adaptation was suppressing the distortion on high frequencies. If we multiply the estimated weights on the log filter-bank amplitudes of the corresponding noisy signals, the resulting signals should be closer to the clean speech signal. Fig. 4(e) compares the log filterbank amplitude of the clean speech signal, the noisy speech signal with 100 dB noise, and its weighted signal. It is shown that the weighting decreased the distance of the distortion part in the noisy speech signal with the clean signal. Fig. 4(f) shows the log filter-bank amplitude distance between the clean speech signal and the noisy speech signals. It is shown that the weighted noisy speech gets closer to the clean speech than the original noisy speech signals.

3.2.1.3. Convergence of adaptation procedures. This experiment investigated the convergent property of the adaptation procedures. The test data come from the sub-set supplied with subway noise in SNR of 5 dB in test set A of AURORA2, with a total of 1001 utterances. Ten utterances are selected from the set as adaptation data and the remainder as test data. The number of sub-bands is set to 7. The feature vectors are composed of 13 cepstrum coefficients (including the zero-order coefficient) and their delta and acceleration. The original HMMs are the clean HMMs in the AURORA2 baseline. The iteration number in the adaptation procedures ranged from 1 to 10. Fig. 5 shows the convergent property of feature space adaptation. It is shown that all variables are in vibrating change. This may be because of the omission of the Jacobian matrix. Nevertheless, they all have a slow trend to convergence. Fig. 5(c) shows that accuracy increased after the adaptation was performed. Fig. 6 shows the convergent property of model space adaptation. Fig. 6(a) shows that the values of seven weights all converged to 1. The reason for this is that the models are updated with estimated weights after each epoch. This indicates that mean vectors are matched with adaptation data after the adaptation procedure. In Fig. 6(b), likelihood monotonously increased and was converged after 2 iterations, which corresponded to the property of the EM algorithm. In Fig. 6(c), it is shown that the accuracy reached convergence quickly, which attained high improvement in one step and didn't change much while iterations increased. Hence, in the following experiments, one-step iteration was performed in the adaptation procedures.

3.2.1.4. Number of sub-bands. This experiment investigates the relationship between accuracy and the number of sub-bands. The test data are test set A. Ten utterances were selected from each sub-set as adaptation data and the remainder as test data. Because there are 13 coefficients in the mean vectors, the number of sub-bands ranged

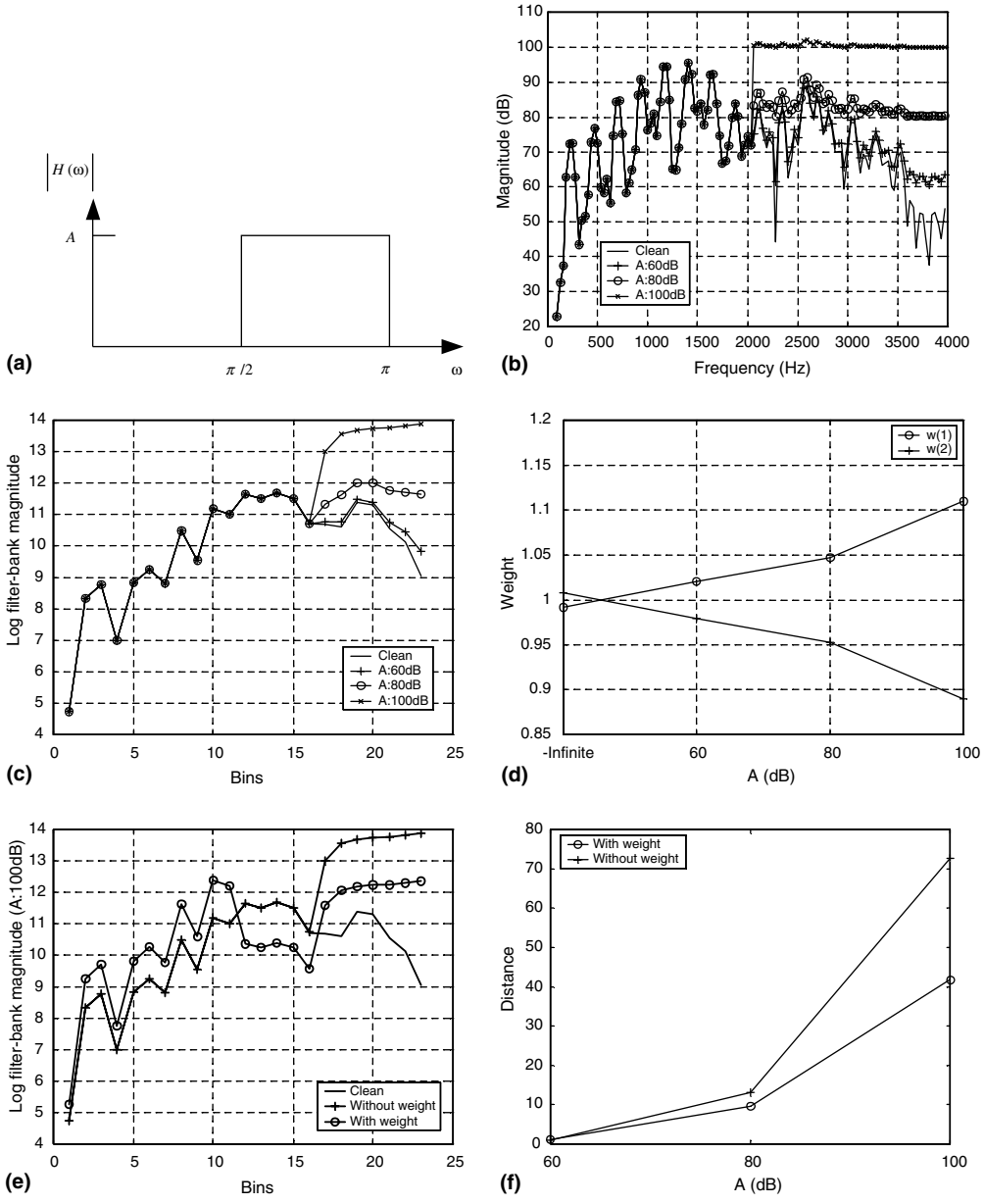

Fig. 4. Signals and weights. (a) Spectrum amplitude of added noise; (b) spectrum amplitude of a frame of speech signal (Clean speech signal; Noisy speech signals supplied with artificial noise signals whose high-frequency amplitude is at 60 dB, 80 dB and 100 dB, respectively); (c) log filter-bank amplitude in all cases; (d) estimated weights in all cases; (e) log filter-bank amplitude of clean speech signal, noisy speech signal with 100 dB noise and its weighted signal; (f) log filter-bank amplitude distances between noisy speech signals (with and without weight) and clean speech signal.

from 1 to 13. The log FBEs are divided into approximately uniform sub-bands. Table 2 shows the width of the sub-bands with the number of sub-bands.

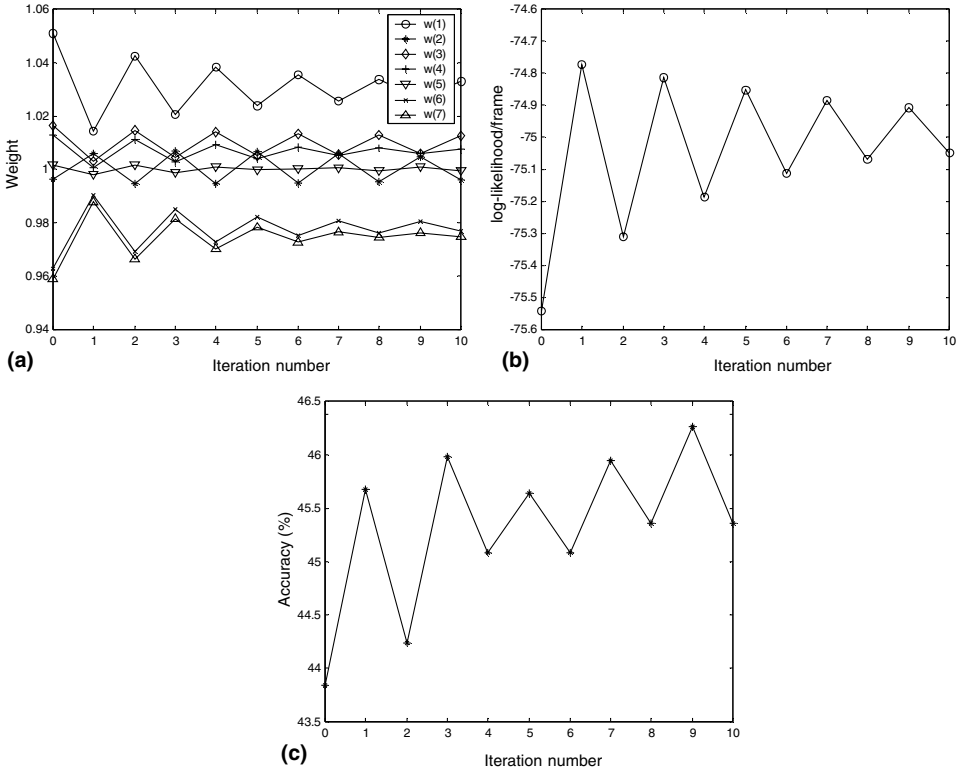

Fig. 5. Convergent property of feature space adaptation with the increase of iteration number. (a) Weights; (b) log likelihood for adaptation data; (c) accuracy for test data.

Fig. 7(a) and (b) show the performance with different numbers of sub-bands based on clean HMMs and multi-condition HMMs, respectively. It is shown that when the number of sub-bands is near seven, performance was best. For the case in which the adaptation is based on clean HMMs, the performance of unconstrained model space adaptation (UMA) is better than that of model space adaptation (MA). The reason for this is that UMA is the unconstrained case of MA and may obtain a more optimal estimation of the weights. The feature space adaptation (FA) exhibited the worst performance because its omission of Jacobian matrices lost some precision in estimation. The irregular behavior of curves for MA is probably due to the discontinuous allocation of sub bands. For the case in which the adaptation is based on multi-condition HMMs, the adaptation procedures showed a slight improvement.

3.2.1.5. Amount of adaptation data. This experiment investigates the relationship between accuracy and the amount of adaptation data. The test data are test set A. The number of adaptation utterances ranged from 1 to 10. The other 991 utterances were used as test data. The number of sub-bands was set to seven. Figs. 8(a) and (b) show the accuracy with different numbers of adaptation utterances based on both clean HMMs and multi-condition HMMs. This shows that the adaptation procedures require a small amount of adaptation data. One adaptation utterance is enough to improve the performance.

3.2.1.6. Number of model classes. This experiment also investigates the relationship between accuracy and the number of model classes in model space adaptation. The test data are the sub-set supplied with babble noise in test set A. One hundred

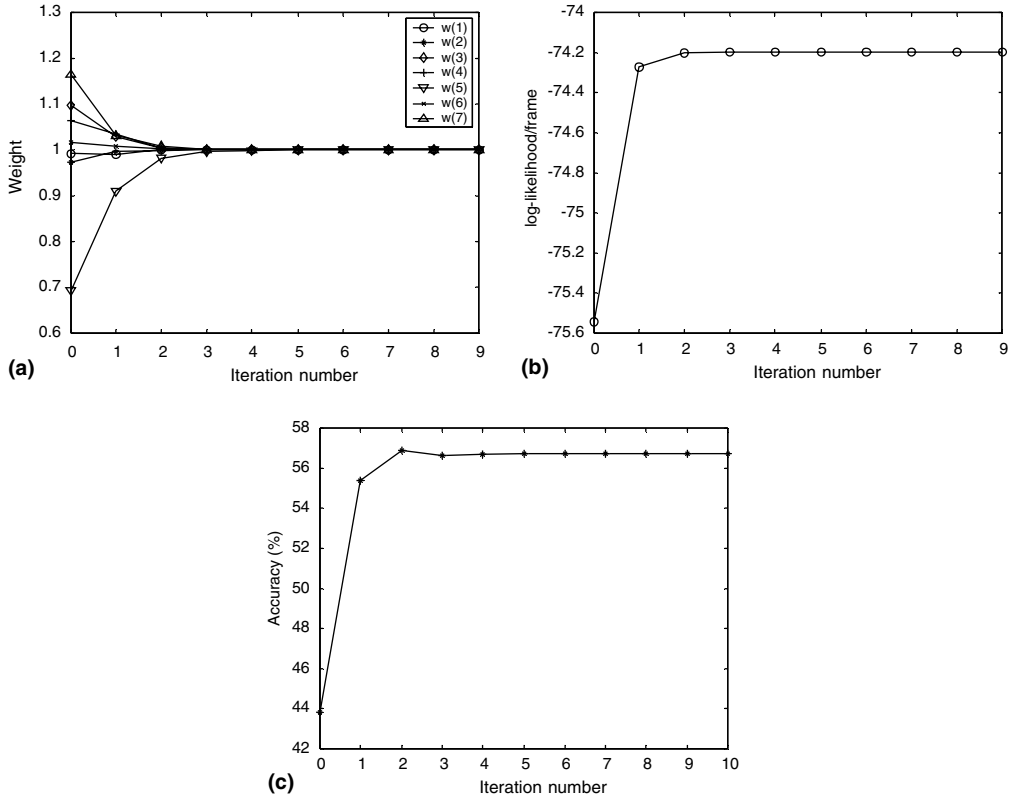

Fig. 6. Convergent property of model space adaptation with the increase of iteration number. (a) Weights; (b) log likelihood for adaptation data; (c) accuracy for test data.

Table 2
Width of sub-bands

Number of sub-bands	Width of sub-bands												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	13												
2	6	7											
3	4	4	5										
4	3	3	3	4									
5	2	2	2	2	5								
6	2	2	2	2	2	3							
7	2	2	2	2	2	2	1						
8	2	2	2	2	2	1	1	1					
9	1	1	1	1	1	1	1	1	5				
10	1	1	1	1	1	1	1	1	1	4			
11	1	1	1	1	1	1	1	1	1	1	3		
12	1	1	1	1	1	1	1	1	1	1	1	2	
13	1	1	1	1	1	1	1	1	1	1	1	1	1

utterances were selected from the set as adaptation data and the remainder as test data. The number of sub-bands was set to seven. We created a regression class tree for Gaussian components to dynam-

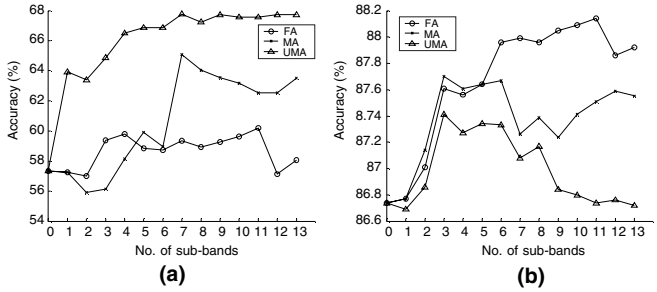

Fig. 7. Accuracy with different numbers of sub-bands based on (a) clean HMMs, (b) multi-condition HMMs. FA: Feature space adaptation; MA: Model space adaptation; UMA: Unconstrained model space adaptation.

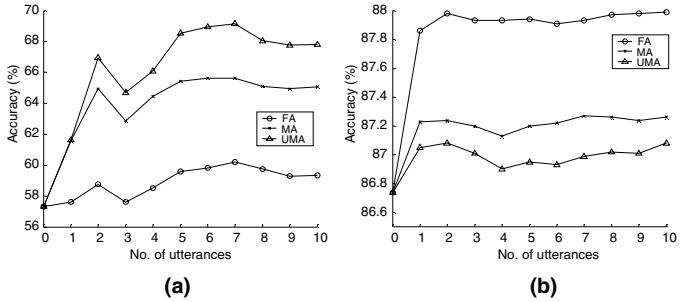

Fig. 8. Accuracy with different numbers of adaptation utterances based on (a) clean HMMs, (b) multi-condition HMMs.

ically set the adaptation classes, depending on the amount of adaptation data that are available. The tree was constructed with a centroid splitting algorithm, which used a Euclidean distance measure (Steve et al., 2001). Fig. 9 shows the performance with different numbers of model classes. This illustrates that increasing the model classes didn't obtain much improvement.

3.2.1.7. Results on advanced front-end. The above experiments were implemented based on the front-end and backend configuration defined in (Hirsch and Pearce, 2000). Recently, a noise-robust front-end has been published for the ETSI standardization of the Advanced Front-End (AFE) for Distributed Speech Recognition (DSR), where a two-stage Mel-warped Wiener filter was used for noise reduction. Complex models, where the digit models have 20 Gaussians per state and the silence model has 36 Gaussians per state, can help improve recognition performance (Macho et al., 2002). In this section, we investigated our approaches based on the AFE and complex models. Table 3 shows the accuracy (%) of four approaches (baseline, FA, MA and UMA) based on clean HMMs and multi-condition HMMs. The test data were test set A. The number of sub-bands was 7 and the number of adaptation utterances was 5. It is shown that FA slightly degraded the baseline performance, while MA and UMA improved the performance. In this experiment, our methods are combined with speech enhancement techniques. It is well known that speech enhancement relies on noisy power spectral density (PSD) estimations and some form of residual noise persists in the enhanced signals. The results show that our methods can further improve the performance of state-of-the-art robust feature extraction and compensation.

3.2.1.8. Results of online unsupervised adaptation. The adaptation approaches have been investigated in batch modes, i.e., the sub-band weights were optimized with a set of adaptation data and then

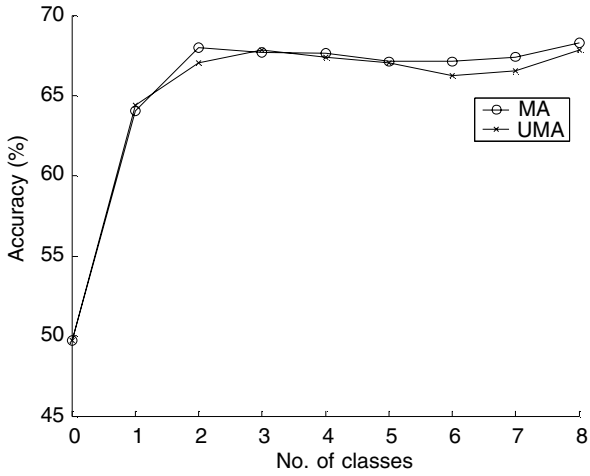

Fig. 9. Accuracy with different numbers of model classes based on clean HMMs.

Table 3
Accuracy (%) with advanced front-end and complex back-end, for four approaches (baseline, FA, MA and UMA) with clean and multi-condition HMMs

	Clean	Multi-condition
Baseline	85.86	94.03
FA	85.81	93.95
MA	86.58	94.10
UMA	87.93	94.04

were used to recognize utterances with the same noise type and level. In this section, we assume that the noise type and level of each test utterance are unknown. We may optimize the weights and word sequence for each utterance with online unsupervised adaptation. The recognition procedure is as follows:

1. Recognize the utterance with baseline models.
2. Given the recognized transcription, adapt the weights in FA or models in MA and UMA.
3. Re-recognize the utterance with the adapted weights or models.

Table 4 shows the accuracy (%) of the adaptation approaches (FA, MA and UMA) in the online unsupervised mode. The test data were test set A and the number of sub-bands was 7. Clean HMMs were used. It is shown that all three adaptation approaches can improve performance in clean and high-SNR conditions but degrade the performance in low-SNR conditions. When the SNR levels tend to be lower, the transcriptions become more easily misrecognized, which causes in accurate estimated weights and corresponding performance degradation.

3.2.2. Resource management evaluation

3.2.2.1. Experimental setup. As the RM task is for clean speech recognition, we need to artificially add noise to its test data in order to investigate its performance in noisy environments. In this experiment, we added the voice babble noise in the Noisex-92 database (Varga et al., 1992) to the Feb89 set at four SNRs from 5 to 20 dB at steps of 5 dB. Hence, we had four sub-test-sets at different SNRs. The performance is defined as the average accuracy over the four sets. We will investigate the performance with different numbers of sub-bands, adaptation data and model classes in the following sections.

3.2.2.2. Number of sub-bands. This experiment investigates the relationship between accuracy and the number of sub-bands. Ten utterances were selected from each sub-set as adaptation data and the remainder as test data. The number of sub-bands ranged from 1 to 13. Fig. 10 shows the performance with different numbers of sub-bands. It

Table 4
Accuracy (%) of adaptation approaches in online unsupervised mode

	SNR (dB)							
	Clean	20	15	10	5	0	−5	Average
Baseline	99.10	95.42	85.44	62.10	31.48	12.54	7.41	57.40
FA	99.12	95.89	85.94	61.21	28.30	10.56	6.79	56.38
MA	99.12	96.30	87.88	64.52	30.93	11.58	7.09	58.24
UMA	99.11	96.58	88.55	65.61	31.51	11.57	7.34	58.76

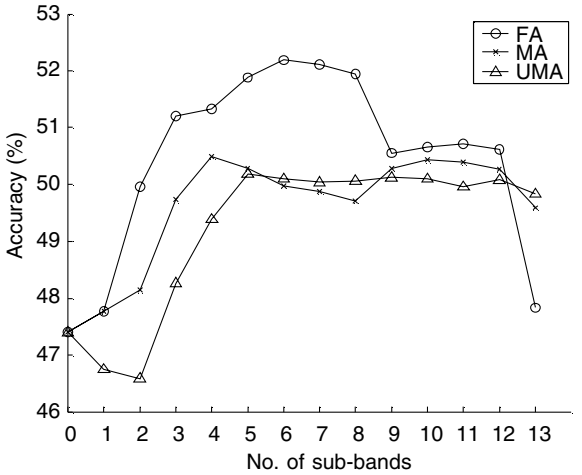

Fig. 10. Accuracy with different numbers of sub-bands. FA: Feature space adaptation; MA: Model space adaptation; UMA: Unconstrained model space adaptation.

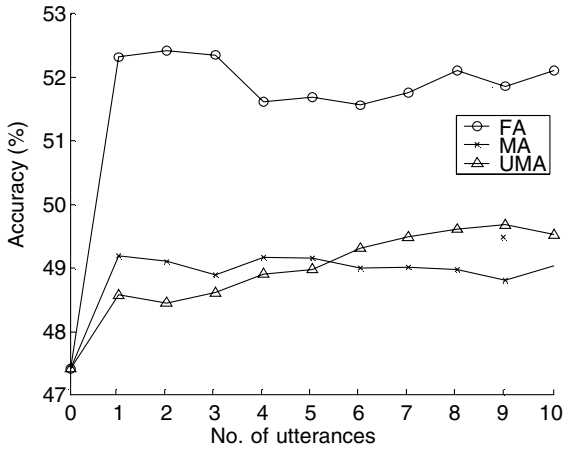

Fig. 11. Accuracy with different numbers of adaptation utterances.

is shown that the three methods all improved the performance and the improvements were similar.

3.2.2.3. Amount of adaptation data. This experiment investigates the relationship between accuracy and the amount of adaptation data. The number of adaptation utterances ranged from 1 to 10. The other utterances were used as test data. The number of sub-bands was set to seven. Fig. 11 shows the accuracy with different numbers of adaptation utterances. It shows that the adaptation procedures require a small amount of adaptation data. One adaptation utterance is enough to improve the performance. FA attained the best performance. FA also achieved the best performance for multi-condition HMMs in Aurora2 (Fig. 8(b)). In both cases, the HMMs are less discriminative due to large amounts of Gaussian components or multi-condition training. However, MA and UMA are better than FA in the following two cases: (1) clean HMMs in Aurora2 (Fig. 8(a)), and (2) Advanced front-end in Aurora2 (Table 3). In both cases, the HMMs are more discriminative due to the clean HMMs or the advanced front-end. Therefore, we may draw the following conclusion: FA is more effective when the HMMs are less discriminative while MA and UMA are more effective when the HMMs are more discriminative.

3.2.2.4. Number of model classes. This experiment investigates the relationship between accuracy and the number of model classes in model space adaptation. Ten utterances were selected from the set as adaptation data and the remainder as test data. The number of sub-bands is set to seven. A regression class tree for Gaussian components was used to dynamically set the adaptation classes.

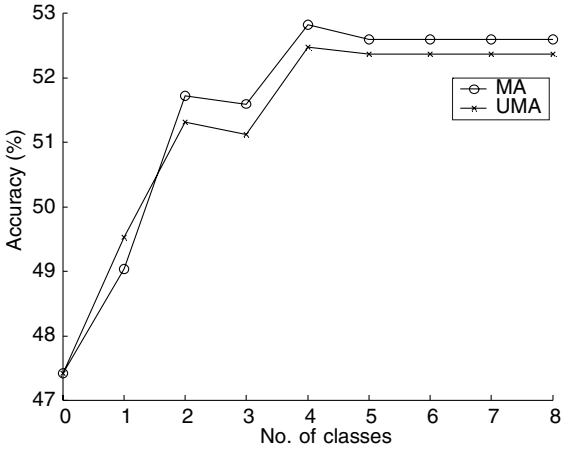

Fig. 12. Accuracy with different numbers of model classes.

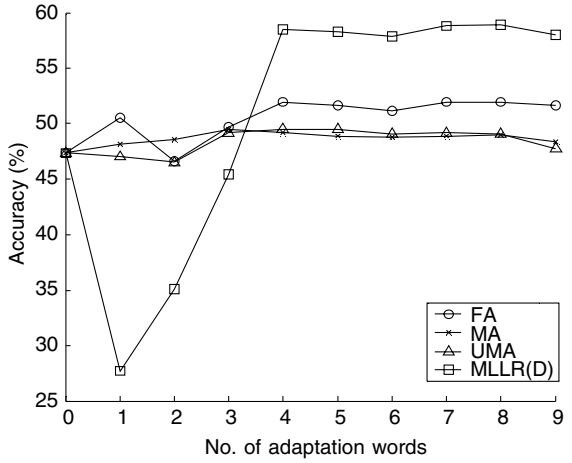

Fig. 13. Performance with different numbers of adaptation word.

Fig. 12 shows the performance with different numbers of model classes. It shows that the performance improved with the increase of model classes.

3.2.3. Comparison with MLLR

When the band scale of the noise signal is known, the weighting factors may be set properly to improve the performance, as presented in Section 2.2. However, if we are not clear about the characteristics of the noise signal, the weighting factors should be estimated by using adaptation approaches. In this case, the weighting approaches may be comparable with other maximum likelihood adaptation methods. Among them, MLLR has been proved to be an effective and robust method able to compensate mismatch between trained HMMs and observed data, which performs linear transformations on mean vectors in Gaussian components in HMMs and estimates them via maximum likelihood criteria (Legetter and Woodland, 1995). We compared their performance on the Resource Management task as follows.

In the MLLR, there is one global regression matrix. Since there are only small amounts of adaptation parameters in our methods, the methods should be able to obtain improvement on small amounts of adaptation data. Therefore, we compared our methods with the MLLR for small amounts of adaptation data. Fig. 13 shows the performance with different numbers of adaptation words. The compared methods consist of FA, MA, UMA, and MLLR with diagonal and full transformation matrices. The MLLR with full transformation matrices is not shown in the figure because its accuracy with a small amount of adaptation data is extremely low. The number of adaptation words ranged from 1 to 9. The number of sub-bands was 7. When the number of adaptation words is less than four, the MLLR with diagonal transformation matrices degraded the performance but the sub-band weighting adaptation methods (FA, MA and UMA) obtained improvement. This indicates that a very small amount of adaptation data is enough for the sub-band weighting adaptation methods but not enough for the MLLR. When the number of adaptation words was increased, the MLLR with diagonal transformation matrices outperformed the sub-band adaptation methods, because the MLLR contains more adaptation parameters. This indicates that the sub-band weighting adaptation methods can obtain improvement with a very small amount of adaptation data, only requiring a length of several words.

3.2.4. Overall results

Finally, we summarized the overall experimental results on the RM task. The number of sub-bands was set to seven. In all the adaptation

Table 5
Recognition accuracy of all methods on RM task with different SNRs

SNR(dB)	5	10	15	20	Clean	Average
FB	15.23	35.61	62.79	75.99	89.07	47.41
CSB	17.64	37.94	56.18	69.92	82.28	45.42
MLLR(D)	8.10	21.83	34.76	46.11	87.27	27.70
FA	15.79	40.39	66.98	78.98	89.13	50.53
MA	17.16	39.87	61.66	74.22	88.84	48.23
UMA	17.48	38.90	59.52	72.29	88.64	47.05

approaches, a global transformation matrix was defined and one word was used as adaptation data. Table 5 shows the accuracy of four methods: full-band (FB), concatenate sub-band (CSB), MLLR with a diagonal transformation matrix (MLLR (D)), feature space adaptation (FA), model space adaptation (MA) and unconstrained model space adaptation (UMA). The MLLR with a full transformation matrix (MLLR (F)) was not listed in the table due to its extremely low accuracy. It is shown that the average accuracy of the CSB was slightly lower than that of the FB. MLLR (D) decreased the accuracy since one word of adaptation data was insufficient for it. The FA and MA achieved higher accuracy than the FB, and the UMA obtained performance similar to the FB. The results indicate that the sub-band adaptation methods need much less adaptation data than the MLLR.

4. Conclusion

In this paper we proposed a sub-band weighting approach for robust speech recognition. Results showed that this approach achieved higher performance than both full-band approaches and conventional sub-band approaches when additive background noise signals were band-limited. In the merging step in conventional sub-band approaches, the reliability of sub-bands may be set empirically, estimated via local SNR, or determined in other ways. However, achieving reliable estimation is still an open problem. Consequently, the performance of the conventional sub-band approaches often degrades. In our approach, maximum likelihood adaptation approaches can be adopted to estimate the weighting factors because of the linear characteristics of the DCT. The results on Aurora2 and RM tasks showed that the maximum likelihood sub-band adaptation approaches consistently improved recognition performance on both tasks. We also compared our adaptation methods with the MLLR. Our methods showed an advantage in that they require only a small amount of adaptation data. In our experiment, with adaptation data including only several words, our methods can obtain improvement, but the MLLR could not.

Acknowledgements

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, ''A study of speech dialogue translation technology based on a large corpus''. The authors also acknowledge useful discussions with B. Mak.

References

Allen, B.J., 1994. How do humans process and recognize speech? IEEE Trans. Speech Audio Process. 2 (4), 567–577.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics Speech Signal Process. ASSP-27 (2), 113–120.

Bourlard, H., Dupont, S., 1996. A new ASR based on independent processing and recombination of partial frequency bands. In: Proc. ICSLP.

Bourlard, H., Dupont, S., 1997. Subband-based speech recognition. In: Proc. ICASSP. pp. 1251–1254.

Cerisara, C., Haton, J.P., Mari, J.F., Fohr, D., 1998. A recombination model for multiband speech recognition. In: Proc. ICASSP. pp. 717–720.

Cerisara, C., Fohr, D., Haton, J.P., 2000. Asynchrony in multiband speech recognition. In: Proc. ICASSP. pp. 1121–1124.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Comm. 34, 267–285.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. 39, 1–38.

Fletcher, H., 1953. Speech and Hearing in Communication. Krieger, New York.

Gauvain, J.L., Lee, C.H., 1991. Bayesian learning of Gaussian Mixture densities for hidden Markov models. In: Proc. DARPA Speech and Natural Language Workshop, Palo Alto, CA. pp. 272–277.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.

Hermansky, H., Tibrewala, S., Pavel, M., 1996. Towards ASR on partially corrupted speech. Proc. IEEE ICSLP, 462–465.

Hirsch, H.G. Pearce, D., 2000. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions, ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium". September 2000.

Kryter, K., 1960. Speech bandwidth compression through spectrum selection. J. Acoust. Soc. Amer. 32 (5), 547–556.

Legetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Language 9, 171–185.

Leonard, R.G., 1984. A database for speaker-independent digit recognition. In: Proc. ICASSP.

Lippmann, P.R., 1996. Accurate consonant perception without mid-frequency speech energy. IEEE Trans. Speech Audio Process. 4 (1), 66–69.

Macho, D., Mauuary, L., Noe B., Cheng, Y.M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., Saadoun, F., 2002. Evaluation of a noise-robust DSR front-end on Aurora databases. In: Proc. ICSLP.

Mak, B., Tam, Y.C., 2000. Asynchrony with re-trained transition probabilities improves performance in multi-band speech recognition. In: Proc. ICSLP Beijing, China, Vol. IV. October. pp. 149–152.

Miller, G., Niely, P., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Amer. 27 (2), 338–352.

Morris, A., Hagen, A., Bourlard, H., 1999. The full-combination sub-bands approach to noise robust HMM/ANN based ASR. Proc. Eurospeech, 599–602.

Okawa, S., Bocchieri, E., Potamianos, A., 1998. Multi-band speech recognition in noisy environments. In: Proc. ICASSP. pp. 641–644.

Paliwal, K.K., Chen, J.D., 2000. Robust feature extraction for speech recognition. The Seventh Western Pacific Regional Acoustics Conference 2000. pp. 61–66.

Price, P.J., Fischer, W., Bernstein, J., Pallett, D., 1988. A database for continuous speech recognition in a 1000 word domain. In: Proc. ICASSP. pp. 651–654.

Riener, K., Warren, R., J.B. Jr., 1992. Novel findings concerning intelligibility of bandpass speech. J. Acoust. Soc. Amer. 91 (4), S2339.

Sankar, A., Lee, C.H., 1996. A maximum-likelihood approach to stochastic matching for robust speech recognition. IEEE Trans. Speech Audio Process. 4 (3), 190–202.

Steve, Y. et al., 2001. The HTK Book. Cambridge University.

Tam, Y.C., Mak, B., 2001. Development of an asynchronous multi-band system for continuous speech recognition. In: Proc. Eurospeech 1, Aalborg, Denmark, September. pp. 575–578.

Tibrewala, S. Hermansky, H., 1997. Sub-band based recognition of noisy speech. In: Proc. ICASSP. pp. 1255–1258.

Tomlinson, M.J., Russell, M.J., Moore, R.K., Buckland, A.P., Fawley, M.A., 1997. Modelling asynchrony in speech using elementary single-signal decomposition. In: Proc. ICASSP.

Varga, A., Steeneken, H.J.M., Tomlinson, M.J., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition, CD-ROM available from the Speech Research Unit, DRA Malvern, UK.

Warren, R., Riener Jr., K.J.B., Brubaker, B., 1995. Spectral redundancy: intelligibility of sentences heard through narrow spectral slits. Perception Psychophys. 57 (2), 175–182.

Weber, K., Ikbal, S., Bengio, S., Bourlard, H., 2003. Robust speech recognition and feature extraction using HMM2. Comput. Speech Language 17, 195–211.