Maximum Likelihood Sub-band Weighting for Robust Speech Recognition

Donglai Zhu†‡, Satoshi Nakamura‡, Kuldip K. Paliwal‡ and Renhua Wang†*

‡ATR Spoken Language Translation Research Labs, Japan
†University of Science and Technology of China, China
*School of Microelectronic Engineering, Griffith University, Australia
{dong.zhu, satoshi.nakamura}@atr.co.jp k.paliwal@me.gu.edu.au rhw@ustc.edu.cn

Abstract

Sub-band speech recognition approaches have been proposed for robust speech recognition, where full-band power spectra are divided into several sub-bands and then likelihoods or cepstral vectors of the sub-bands are merged depending on their reliability. In conventional sub-band approaches, correlations across the sub-bands are not modeled and the merging weights can only be set experientially or estimated during training procedures, which may not match observed data. The methods further degrade performance for clean speech. We proposed a novel sub-band approach, where frequency sub-bands are multiplied with weighting factors and merged, which considers sub-band dependence and proves to be more robust than both full-band and conventional sub-band approaches. And further the weighting factors can be obtained by using the maximum-likelihood estimation approaches in order to minimize the mismatch between the trained models and the observed features. Finally we evaluated our methods on both the Aurora2 task and the Resource Management task and showed improvement of performance on the two tasks consistently.

1. Introduction

It is well known that current ASR systems don't work as well as the human. Fletcher and his colleagues [1] 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 speech show that humans can recognize speech signals with limited spectral cues and can easily integrate acoustic cues from different frequency regions for speech perception [2,3]. To capture the phenomena, sub-band approaches were proposed and two modes of sub-band approaches have been applied: parallel sub-band (PSB) and concatenating sub-band (CSB) [4,5]. In PSB, full-band power spectrum is divided into several sub-bands and each sub-band spectrum is converted into a number of cepstral features. The features are then modeled independently, and the likelihood scores of sub-bands are merged at some segmental levels. In CSB, sub-band cepstral vectors are concatenated as a single feature vector for speech recognition. However, results showed that both PSB and CSB don't perform well in some noisy case and further degrade performance for clean speech, mainly because correlations across the sub-bands are lost in both approaches.

We proposed 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. For the DCT has the size of full-band 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 the size of the weighting factors is small, adaptation approaches are suitable for the estimation. In this paper we derived the adaptation formulae from maximum-likelihood stochastic matching theory [6].

2. Sub-band weighting

In a general feature extraction procedure for MFCC, the speech signal is converted to the spectrum via discrete Fourier transformation (DFT), the spectrum is passed through a group of Mel-frequency filter banks to get Mel-frequency FBEs, a logarithm is performed, and finally the MFCC is obtained from log FBEs via a DCT.

Figure 1 shows the feature extraction procedure with sub-band weighting. In the procedure, after the FBEs are obtained from the speech signal, they are divided into the sub-bands. In the figure there are K sub-bands. Then, log FBE bins in each sub-band are multiplied with a weighting factor depending on the reliability of the sub-band. Next they are extended to the dimension of the full filter-bank by padding zeros on other bins. After the DCT performed on the log FBE vectors of the sub-bands, the corresponding cepstral vectors of the sub-bands can be obtained. All of them have the same or lower dimension as the full-band FBE vector if truncation isn't or is performed during DCT. In the figure DXD means the size of the DCT matrix, where D is the length of full-band FBE vector. The full-band cepstrum vector can then be obtained by adding the cepstrum vectors of the sub-bands. Because of the linear characteristics of the DCT, the resulting cepstrum is equal to the general full-band MFCC if all of the weighting factors equal one. However, in CSB, the log FBE vector of

each sub-band is not extended to full-band dimension. The corresponding DCT and cepstrum vector have the dimension less than full-band. Lastly the feature vector is obtained by concatenating the cepstrum vectors of the sub-bands, which is different from the general MFCC because the sub-bands are processed independently.

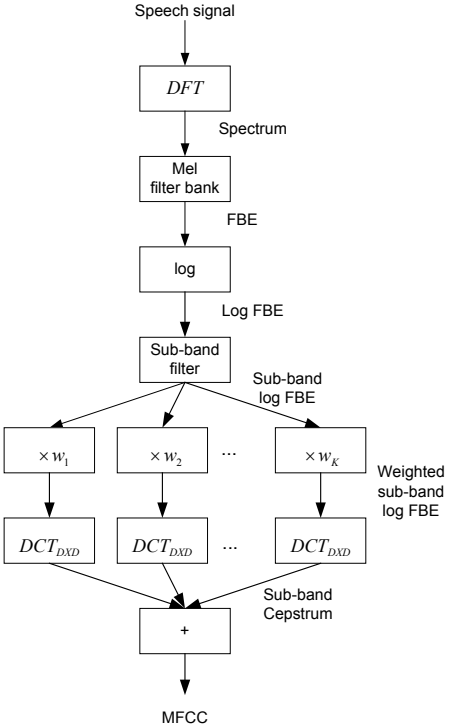


Figure 1: Feature extraction with sub-band weighting

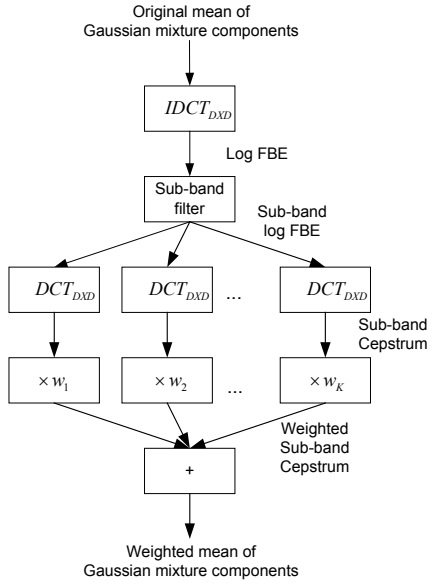


Figure 2: Sub-band weighting on mean vectors

Correspondingly, we may also perform the weighting on the model space. The procedure is shown in Figure 2. If cepstrum is adopted as the feature in the speech recognition system, the mean vectors of the Gaussian mixture components in the HMM set are values on cepstral domain. They can be converted back to log FBEs via IDCT, divided into sub-bands and weighted. In the feature vector, a zero-order cepstrum is required because of the IDCT in the procedure. In some cases weighting should be applied on both the feature space and the model space simultaneously. E.g., if the corrupted frequency component part has been known, we may multiply a small weighting factor (e.g., zero) on this frequency part in order to weaken its contribution in speech recognition, where the weighting factors must be multiplied on both feature space and model space to keep the two spaces consistent.

3. Sub-band weighting adaptation

One problem in the sub-band weighting approach is the estimation of the weighting factors. The weighting factors can be set if the reliability of the sub-bands is known. Otherwise, they need to be estimated somehow if the characteristic of the noise signal is unknown. Because the number of the weighting factors is small, they can be estimated from small amounts of adaptation data. In this section a maximum-likelihood estimation-based adaptation approach is investigated to estimate the weighting factors on model space. According to the procedure in Section 2, the model space weighting procedure may be represented by the following equations

$$\boldsymbol{\mu} = \mathbf{U}\mathbf{i} \tag{1}$$

$$\hat{\boldsymbol{\mu}} = \mathbf{U}\mathbf{w} \tag{2}$$

where, $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ are original and weighted mean vectors respectively. $\mathbf{U}$ is a cepstrum matrix whose columns are cepstrum vectors of sub-bands $\mathbf{U} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_K]$. $\mathbf{i}$ is a unit vector. $\mathbf{w}$ is the weight vector $\mathbf{w} = [w_1, w_2, ..., w_K]^T$. K is the number of sub-bands.

With these conditions and definitions, the auxiliary function in model space adaptation algorithm [6] may be written as

$$Q(\mathbf{w}, \hat{\mathbf{w}}) = \sum_{j=1}^{N}\sum_{m=1}^{M}\sum_{t=1}^{T}\zeta_t(j,m) \left\{-\frac{1}{2}\left[\mathbf{y}_t - \mathbf{U}_{jm}\hat{\mathbf{w}}\right]^T \boldsymbol{\Sigma}_{jm}^{-1}\left[\mathbf{y}_t - \mathbf{U}_{jm}\hat{\mathbf{w}}\right]\right\} \tag{3}$$

where, N, M, T are the number of states, Gaussian components and frames respectively. $\zeta_t(j,m)$ is the posteriori probability of the m th Gaussian component in the j th state given the t th observed frame. $\mathbf{y}_t$ is the observed feature vector. $\boldsymbol{\Sigma}_{jm}$ is the covariance matrix in the m th Gaussian component in the j th state.

Meanwhile, weighting factors have the following constraint

$$\mathbf{i}^T\mathbf{w} = K \tag{4}$$

In order to find the maximum of the auxiliary function in the condition of (4), we may define objective function as

$$F(\mathbf{w}, \hat{\mathbf{w}}) = Q(\mathbf{w}, \hat{\mathbf{w}}) + \lambda(\mathbf{i}^T\hat{\mathbf{w}} - K) \tag{5}$$

Differentiating (5) with respect to $\hat{\mathbf{w}}$ and solve for its zeros

$$\nabla_{\hat{\mathbf{w}}} F(\mathbf{w}, \hat{\mathbf{w}}) = 0 \tag{6}$$

Then we may get the following solution

$$\hat{\mathbf{w}} = \mathbf{X}^{-1}(\mathbf{y} + \lambda\mathbf{i}) \tag{7}$$

where

$$\mathbf{X} = \sum_j\sum_m\sum_t \zeta_t(j,m)\mathbf{U}_{jm}{}^T\boldsymbol{\Sigma}_{jm}^{-1}\mathbf{U}_{jm} \tag{8}$$

$$\mathbf{y} = \sum_j\sum_m\sum_t \zeta_t(j,m)\mathbf{U}_{jm}{}^T\boldsymbol{\Sigma}_{jm}^{-1}\mathbf{y}_t \tag{9}$$

$$\lambda = \frac{K - \mathbf{i}^T \mathbf{X}^{-1} \mathbf{y}}{\mathbf{i}^T \mathbf{X}^{-1} \mathbf{i}} \quad (10)$$

The weighting procedure on the mean vectors may be regarded as a type of transformation on them. So it is possible to remove the constraint on the weights. This case we called "unlimited sub-band weighting adaptation". The solution is slightly different from the previous one

$$\hat{\mathbf{w}} = \mathbf{X}^{-1} \mathbf{y} \quad (11)$$

where, $\mathbf{X}, \mathbf{y}$ are equal to (8) and (9) respectively.

4. Experimental results

We investigated the performance on both the Aurora2 task and the DARPA Resource Management (RM) task [7,8]. The Aurora2 corpus was created for research on distributed speech recognition under noisy environments. It was composed of connected digits from the clean TIDIGITS database. The data were pre-filtered according to the frequency characteristics of common telecommunication channels (G.712 or MIRS) and were artificially added with realistic noises at six different signal-to-noise (SNR) ratios ranging from 20 dB to -5 dB at 5 dB steps. In its baseline, there were 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. The RM database is a collection of speaker-dependent and speaker-independent continuous speech data for both training and testing speech recognition systems. The speech comes from a variety of speakers of North American English collected by Texas Instruments in quiet environments using close talking, 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 Feb89 set is used as the test set, consisting of 300 sentences from 10 new speakers.

4.1. Band-limited noise

In this experiment we investigated the performance of the weighting and adaptation approaches for the speech signals with additive band-limited noise signals on the RM task. The word pair grammar was used in the decoder. There were 49 monophone HMMs. Each HMM consisted 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, their delta and accelerator coefficients. Since the test set included only speech signals, noise signals could be added. It was assumed that the noise signals were band-limited on the frequency domain and the scale of the band was known. Hence, we could artificially set the weights to restrict the noisy parts. One setting way was to let the weights for this band equal zero and others equal one.

We added band-limited noise signal onto the test set. The amplitude of the noise signal was set as 1.0E5 (100dB). The cepstra are converted to log filter-banks via IDCT, so there are 13 bins in 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. The performance of four features (FB, CSB, CCSB and WSB) was compared on both clean test data and noisy test data. FB is the general full band MFCC. For CSB, we divided the 13 bins in 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. 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 be zero, while the others to be one. The weighting factors were multiplied on both feature space and model space to keep consistency.

Clean	FB	88.76				
	CSB	82.24				
Noisy	Noisy bins	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
	WA	74.10	85.78	85.50	87.72	83.28
	UWA	74.83	85.66	85.90	87.39	83.45

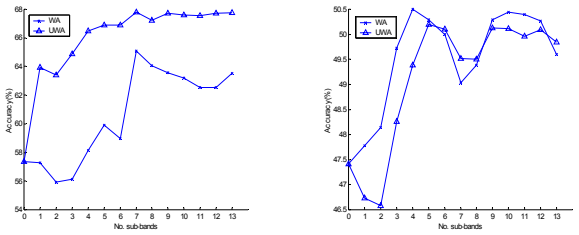
Table1: Accuracy (%) of different features for clean test data and noisy test data. FB: full band; CSB: concatenated sub-band; CCSB: concatenated clean sub-band; WSB: weighted sub-band; WA: sub-band weighting adaptation; UWA: unlimited sub-band weighting adaptation

Table 1 shows the accuracy of different features in different conditions and that of the adaptation approach as well. WA was the sub-band weighting adaptation approach where the number of sub-bands was 4 but it needn't know which sub-band was noisy. UWA was the unlimited case. For clean test data, FB achieved better performance than CSB. For band-limited noisy test data, FB degraded the performance. The noise signal added on lower bins had more degradation because it confused the formants of the speech signal. CSB improved the performance over FB when the noise signal impaired severely but degraded 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. WSB achieved much better performance than both CSB and FB in the noisy case, which was even similar to that in the clean case. WA and UWA obtained similar performance that is better than FB.

4.2. Realistic noise

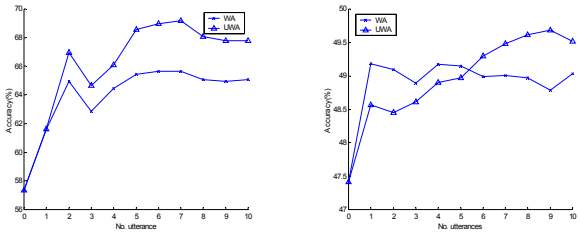
In this experiment we investigated the performance of the sub-band weighting adaptation approach for realistic noises on both the Aurora2 task and the RM task. For the Aurora2 task, test set A was taken as test data. HMMs are clean training models. Each sub-set was divided into two parts: adaptation data and test data. The adaptation procedure was performed on the previous, and the performance was

evaluated on the later. 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 cepstral coefficients including the zero-order coefficient, and their delta and acceleration coefficients. For the RM task, we added the voice babble noise signals in Noisex-92 database [9] to the Feb89 set at four SNRs from 5dB to 20dB at the step of 5dB. Hence, we had four sub-test-sets at different SNRs. The performance is defined as the average accuracy over the four sets. We investigated the performance with different number of sub-bands and adaptation data as follows.



(a) (b)

Figure 3: Accuracy with different numbers of sub-bands on (a) Aurora2; (b) RM.



(a) (b)

Figure 4: Accuracy with different numbers of adaptation utterances on (a) Aurora2; (b) RM.

4.2.1. Number of sub-bands

This experiment investigated the relationship between the accuracy and the number of sub-bands. Ten utterances in the test set were used as adaptation data and the remainder as test data. The number of sub-bands ranged from 1 to 13. Figure 3 showed the performance of WA and UWA with different numbers of sub-bands. It illustrated that the performance improved when the number of sub-bands increased in all cases. It was shown in figure 3(a) that more than 6 and 3 numbers of sub-bands were preferable for WA and UWA respectively on the Aurora2 task. Figure 3(b) showed that more than 4 and 3 were preferable on the RM task.

4.2.2. Number of adaptation data

This experiment investigates the relationship between the accuracy and the number of adaptation data. The number of adaptation utterances ranged from 1 to 10. The other utterances in the test set were used as test data. The number of sub-bands was set as seven. Figure 4 showed the performance of WA and UWA with different numbers of adaptation utterances. In all cases, the performance improved when there was only one adaptation utterance. Figure 4(a) and 4(b) showed that more than 4 adaptation utterances were preferable on the Aurora2 task and one adaptation utterance was enough for the RM task. It illustrated that the required amounts of adaptation data were small.

5. Conclusions

In this paper we proposed a sub-band weighting approach for robust speech recognition. Its advantage over the conventional sub-band approaches is that it keeps the correlations across the sub-bands. Results showed that this approach achieved higher performance than both full-band approaches and conventional sub-band approaches in the cases of band-limited additive background noise signals. We also proposed that the weighting factors could be estimated by the maximum-likelihood adaptation approaches. Results showed that the adaptation approaches improved recognition performance on both the Aurora2 task and the RM task consistently while required only small amounts of adaptation data.

6. 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”.

7. References

[1] Fletcher H., “Speech and Hearing in Communication”, Krieger, New York, 1953

[2] Lippmann P.R., “Accurate Consonant Perception Without Mid-Frequency Speech Energy”, IEEE Trans. on Speech and Audio Processing, vol. 4, No. 1, pp. 66-69, 1996

[3] Warren R., Riener K., Jr. J.B. and Brubaker B., “Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits”, Perception and Psychophysics, 57(2), pp. 175-182

[4] Hermansky H, Tibrewala S. and Pavel M., “Towards ASR On Partially Corrupted Speech”, Proc. IEEE ICSLP, pp. 462-465, 1996

[5] Okawa S., Bocchieri E. and Potamianos A., “Multi-band speech recognition in noisy environments”, Proc. ICASSP, pp. 641-644, 1998

[6] Sankar A. and Lee C.H., “A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition”, IEEE Trans. on Speech and Audio Processing, vol. 4, No. 3, pp. 190-202, 1996

[7] Hirsch H.G. and Pearce D., “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”, Sept. 2000

[8] Price P.J., Fischer W., Bernstein J., Pallett D. “A Database for Continuous Speech Recognition in a 1000 Word Domain”, Proceedings ICASSP, pp. 651-654, 1988

[9] Varga A., Steeneken H.J.M., Tomlinson M.J, and Jones D., “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, 1992