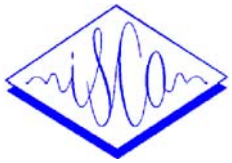

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A RECOMBINATION STRATEGY FOR MULTI-BAND SPEECH RECOGNITION BASED ON MUTUAL INFORMATION CRITERION

Shigeki Okawa[†], *Takehiro Nakajima* [‡] *and Katsuhiko Shirai* [‡]

[†] Department of Network Science, Chiba Institute of Technology, Narashino, Japan
[‡] Department of Information and Computer Science, Waseda University, Tokyo, Japan
(E-mail: okawa@net.it-chiba.ac.jp; URL: http://www.ok.net.it-chiba.ac.jp/okawa/)

ABSTRACT

This paper presents a recombination strategy for multi-band automatic speech recognition (MB-ASR). Several recent works have suggested that MB-ASR gives more accurate recognition, especially in noisy acoustic environments. The main issue in this study concerns the sub-band score recombination in MB-ASR framework. Intuitively, it seems very improbable that all sub-band features have the same amount of information for speech recognition. We therefore investigate to weight the contribution from each band at the recombination process by using a strategy derived from the information theory. The quantity of information is well determined by the mutual information between band features and target phoneme categories to be recognized. The experimental results show that the recognition accuracy improves for noisy speech by using three and six stream systems with the proposed approach.

1 INTRODUCTION

Multi-band based automatic speech recognition (MB-ASR) is a novel paradigm in speech recognition. Its basic strategy is to recognize speech by using multiple frequency bands whose acoustic features are individually extracted. Several recent works have shown that the MB-ASR could yield better performance mainly in noisy or mismatched acoustic conditions [1, 2, 3, 4].

Traditionally, speech recognition is performed by extracting a set of acoustic feature vectors, which are calculated from the whole frequency band of input speech. In this case, even if only a part of the frequency band is corrupted by noise, all the feature vector components are affected. MB-ASR solves this problem by modeling sub-bands independently. We have several reasons for believing that MB-ASR should be investigated: (i) There is a psychoacoustic evidence, that human beings are likely to process narrow frequency bands independently in auditory perception [5]. (ii) Statistical modeling of sub-band features may be more accurate than full-band models, because of the higher dimensionality of the full-band feature space. (iii) Actual existent noise may be strongly corrupt only few frequency bands.

There are two major issues in MB-ASR studies. One is about the definition of sub-band at the frontend. (How many bands? Which bandwidth?) The other is about the sub-band likelihood recombination, at which the recognizer has to obtain a global decision. In this paper, we focus on the latter issue. During the recognition process, different speech classifiers are applied individually to each sub-band, and each classifier provides a set of recognition hypotheses and scores. The problem is how to combine all classifier outputs to obtain global scores. Experientially and intuitively, it seems very improbable that all sub-band features have exactly same amount of information for speech recognition. For instance, a sub-band which has several formants may have more information than others. In another case, we should reduce the contribution from a band which has noisy elements.

In this study, we especially investigate to weight the contribution from each sub-band by using a weighting strategy derived from the information theory. In the next section, we introduce the theoretical view of the weighting strategy and its implementation onto our system. In section 3, we describe our ASR system and the experimental condition. In section 4, we report on the experimental results, and conclusions are in section 5.

2 WEIGHTED SCORE RECOMBINATION

In this section, we introduce a weighting strategy for the multi-band score recombination. Figure 1 illustrates the basic concept of MB-ASR with the recombination weighting. When the input speech frequency is split into three bands as shown in the figure, three sub-band likelihoods are computed by sub-band HMM's respectively. To recombine those three likelihoods, we attempt to multiply all outcomes with certain weighting factors derived from the posterior information of the speech input as well as the prior knowledge.

According to Bourlard's paper [1], recombination at the HMM state level gives almost the same accuracy as recombination at higher levels such as phone, syllable or word. Since state level recombination is obviously much simpler to implement, in this study we adopt the HMM state as the recombination level.

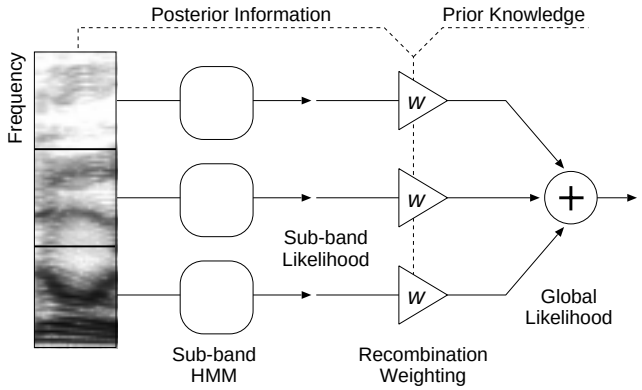

Figure 1: Schematic diagram of multi-band ASR with the recombination weighting. The sub-band likelihoods, given by a certain number of sub-band HMM's, are recombined with weighting to obtain the global score.

Sub-band Score Weighting

Here, let o_i and s_j be an observation vector at frame i and an HMM state j. After computing frame probability $p(o_i^b|s_j)$ for each band b, assuming independence of the bands, recombination of the probabilities could be realized by multiplying all outputs:

$$p(o_i|s_j) = \prod_{b=1}^{B} p(o_i^b|s_j), \tag{1}$$

where B is the number of sub-bands.

However, as mentioned in section 1, if a frequency band has relatively more information than others, the band should have larger weight value. A solution for this problem is to weight the contribution from each sub-band using *probability exponents* as follows [6]:

$$p(o_i|s_j) = \prod_{b=1}^{B} p(o_i^b|s_j)^{w_b}, \tag{2}$$

where w_b is the weighting factor corresponding to the sub-band b.

Mutual Information Weighting

There might be several reasonable strategies for selecting the weighting factor w_b in equation (2). The first author formerly attempted introducing the sub-band signal-to-noise ratio (SNR) and the inverse conditional entropy of each band [3]. In this study, we extend the entropy weighting strategy further.

The quantity of information is well determined by the conditional entropy of the frame under consideration. To treat the entropy as a relative value between each sub-band, it should be calculated using *a posteriori* probabilities. Here we compute the probability by the following approximation.

$$p(s_j|o_i^b) \cong \frac{p(o_i^b|s_j)}{\sum_j p(o_i^b|s_j)}. \tag{3}$$

The conditional entropy S (of all HMM states and phoneme categories), given the outcome of the observation o_i^b, is denoted by:

$$H(S|o_i^b) = \sum_j -p(s_j|o_i^b) \log p(s_j|o_i^b), \tag{4}$$

then the averaged entropy $H(S|O)$ for a certain length of frames is defined as:

$$H(S|O) = \sum_i \sum_j p(s_j, o_j^b) \log p(s_j|o_j^b). \tag{5}$$

This value measures the ambiguity in deciding the most probable HMM state when $\{o_i^b\}$ is observed.

Similarly, the self entropy of HMM state S can be denoted by:

$$H(S) = \sum_j -p(s_j) \log p(s_j), \tag{6}$$

from a prior knowledge (language probability *etc.*)

Using $H(S)$ and $H(S|O)$, the mutual information $I(S;O)$ is defined as:

$$I(S;O) = H(S) - H(S|O) \tag{7}$$

We attempt to employ the mutual information $I(S;O)$ as the weighting factor w_b in equation (2), allows to measure the amount of information contained in S with respect to the observation O. All the weights are normalized to sum up to the number of sub-bands. The acoustic score of all sub-bands are thus recombined at each HMM state level with the weight value.

To investigate the optimal length of frames for the weight computation, we test the weighting by (i) one frame, (ii) a segment which contains 50 ms frames, and (iii) the entire sentence (word).

3 SYSTEM OVERVIEW

In this section, we describe our speech recognition system and the experimental setup. Our recognizer is based on context independent phoneme HMM's. In this study, we choose a continuous phoneme recognition task, in which only phoneme bigram is applied as a language model. Each HMM has 4 states, 3 loops, 16 mixture Gaussian distribution.

Frontend

As the frontend of the recognizer, we use the MFCC analysis after filter-bank processing. In full-band (conventional) case, the DCT calculation is applied to the whole filter-bank to obtain a series of mel-cepstral feature vectors. In multi-band case, the filter-bank output is split into several disjoint bands, then the DCT is applied to each sub-band individually.

The digitized waveform, which is sampled at 16 kHz, is analyzed with a 21.3 ms Hamming window, shifted by a 5 ms interval. For each frame, we obtain 31 components

of filter-banks, then a certain dimension of MFCC's. The MFCC dimension is decided in proportion to the number of sub-bands, *i.e.*, 12 for full-band, 4 for three-band and 2 for six-band case, and so on. We add the 1st and 2nd derivatives of the MFCC's as well as the frame energy.

Sub-band Partitioning

In full-band based (conventional) system, there are 31 filter-banks as an input. For the multi-band case, we use three and six sub-bands defined as follows:

- 3 bands: (0-1155) (1050-2996) (2723-8000) Hz
- 6 bands: (0-650) (550-1155) (1050-1860) (1691-2996) (2723-4824) (4386-8000) Hz

The bandwidths of all bands are equal partitions of the mel-frequency scale. The band overlap is due to the band filter characteristics.

Speech Data

We use 52,400 Japanese common word data spoken by 10 male speakers from ATR Japanese speech database. The speech data is recorded with a close-talk microphone in laboratory environments. The test is performed by three different combinations of 47,160 words by 9 speakers as the training dataset; 5,240 words by other 1 speaker as the test dataset.

Noise Data

In our experiments, we add two different types of noise onto clean speech data digitally to test the recognizer under noisy conditions. We add the noise only to the test speech data, but not to the training data.

At first, as an *ideal* type of noise, we generate *LPW* noise, which is white noise added only to the first frequency band, by applying an FIR filter. The second noise, which is *Babble* noise generated by a hundred people speaking in a canteen, is from the SPIB database. This is the noise that we need to deal with in many practical situations.

4 EXPERIMENTS

Effect of Conditional Entropy

Before testing the weighting recombination, we investigate the change of the conditional entropy distribution when a part of the input speech is corrupted by noise. Using three-band system, we add the *LPW* noise, which contains noisy elements only on the first sub-band, and compute the conditional entropy in equation (4) for each sub-band. We also compute the same value for the clean speech.

Figure 2 compares the distribution for lower, middle and higher sub-bands. In the bottom figure, only the lower band (first band) contains noisy elements at 10 dB level. The distribution of the conditional entropy $H(S|O_j^b)$ by this sub-band shifts to the right, which means the sizable amount of information of the band decreases by the noise.

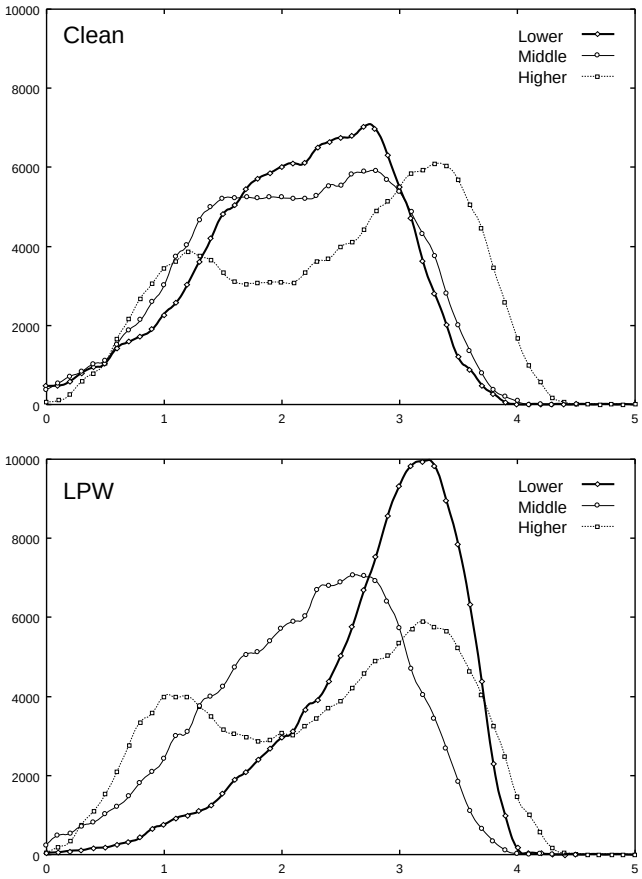

Figure 2: The distribution of the conditional entropy $H(S|O_j^b)$ of three sub-bands, for clean speech and the *LPW* noisy speech. The peak of the entropy distribution for the lower (first) band increases, while other two bands remain the same positions.

Recognition with Weighted Score Recombination

Next, we test the recognition performance of the weighting score recombination for two types of noise; *LPW* and *Babble*. The noise is added digitally at 10 dB SNR level to test data. The acoustic likelihood of each sub-band is recombined according to equation (2) using weights w_b as: (i) constant (no weighting), (ii) equal to the mutual information $I(S;O)$ for each frame, (iii) $I(S;O)$ for a 50 ms segment, and (iv) $I(S;O)$ for the entire sentence. In (iii) case, we obtain the averaged value of $I(S;O)$ from the preceding 50 ms frames. Here, all classifier outputs are recombined at each HMM state level.

Figure 3 summarizes the phoneme recognition accuracy (phoneme correct rate % - phoneme insertion rate %) for several weighting strategies. In the figure, *Full-band System* should give the baseline performance for each band number. *No Weight* refers to the use of a constant weight equal to [1:1:1] and [1:1:1:1:1:1].

For *LPW* noise, the recognition accuracy significantly improves (from 48.4 points to 60.4 points) by using the six-band system with *No Weight.* Further improvement is observed when we apply the mutual information weighting

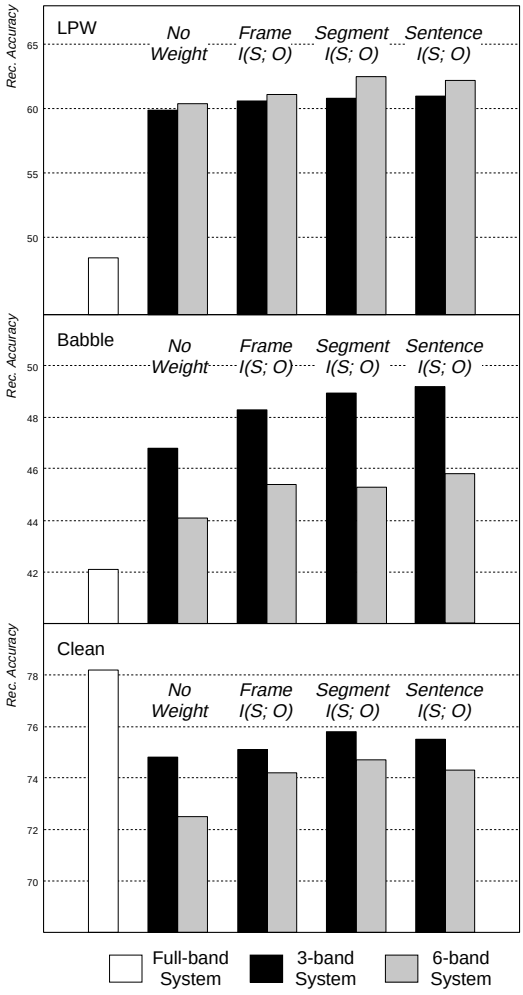

Figure 3: Summary of phoneme recognition accuracy for two types of noisy data (*LPW* and *Babble*) and clean speech with full-band, three- and six-band systems. The noise is added to clean speech data at 10 dB SNR level in all cases.

(62.5 points using $I(S;O)$ for a 50 ms segment, as the best case.)

For *Babble* noise, contrary to *LPW* noise, three-band system gives better performance than six-band system. Again, the mutual information weighting works well in both cases, especially with three-band system (49.2 points using $I(S;O)$ for the entire sentence, while *No Weight* case gives 44.1 points.)

On the other hand, for clean speech, even the weighting strategy works well in both three- and six-band system, full-band system still gives the best recognition performance. This result agrees with [3].

In general, when we apply the weight at the recombination process, the recognition accuracy increases in all conditions, compared with *No Weight* case. The performance improvement and the optimum number of bands seem to depend on the type of noise, as we observe a dramatic improvement with the proposed weighting for *Babble* noise. As the considering duration of the weight, there are few significant differences among frame, segment and the entire sentence levels.

5 CONCLUSION

In this paper, we studied a weighting strategy for the multi-band speech recognition framework. Particularly, we examined a weighting by using the mutual information criterion, which could measure the amount of information of each sub-band. We performed several ASR experiments after adding different types of noise signals to the input speech. In general, we found that multi-band ASR is more robust than conventional ASR, and the weighting strategy is effective in the most case, except clean speech.

We would like to investigate further (i) to analyze the error properties to understand the reason of the weighting effect, and (ii) to find the optimal weighting strategy.

Acknowledgments

We acknowledge Enrico Bocchieri and Alex Potamianos (AT&T Labs, USA), whom the first author collaborated with in 1996-1997, and the SIPS members of AT&T Labs for many useful discussions and suggestions in the early phase of this study.

References

[1] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *Proc. Int. Conf. on Spoken Language Processing*, pages 426–429, Philadelphia, October 1996.

[2] S. Tibrewala and H. Hermansky. Sub-band based recognition of noisy speech. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1255–1258, Munich, April 1997.

[3] S. Okawa, E. Bocchieri and A. Potamianos. Multi-band speech recognition in noisy environments. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 641-644, Seattle, May 1998.

[4] N. Mirghafori and N. Morgan. Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers. In *Proc. Int. Conf. on Spoken Language Processing*, pages 743–746, Sydney, December 1998.

[5] J. B. Allen. How do humans process and recognize speech? *IEEE Trans. on Speech and Audio Processing*, 2(4):567–577, October 1994.

[6] Y. Normandin, R. Cardin, and R. DeMori. High-performance connected digit recognition using maximum mutual information estimation. *IEEE Trans. on Speech and Audio Processing*, 2(2):299–311, April 1994.