MULTI-BAND SPEECH RECOGNITION IN NOISY ENVIRONMENTS

Shigeki Okawa[*], *Enrico Bocchieri and Alexandros Potamianos*

AT&T Labs - Research
180 Park Avenue, Florham Park, NJ 07932-0971, USA
{okawa, enrico, potam}@research.att.com

ABSTRACT

This paper presents a new approach for multi-band based automatic speech recognition (ASR). Recent work by Bourlard and Hermansky suggests that multi-band ASR gives more accurate recognition, especially in noisy acoustic environments, by combining the likelihoods of different frequency bands. Here we evaluate this likelihood recombination (LC) approach to multi-band ASR, and propose an alternative method, namely feature recombination (FC). In the FC system, after different acoustic analyzers are applied to each sub-band individually, a vector is composed by combining the sub-band features. The speech classifier then calculates the likelihood from the single vector. Thus, band-limited noise affects only few of the feature components, as in multi-band LC system, but, at the same time, all feature components are jointly modeled, as in conventional ASR. The experimental results show that the FC system can yield better performance than both the conventional ASR and the LC strategy for noisy speech.

1. INTRODUCTION

Robustness is a very important issue in the field of automatic speech recognition (ASR) research, especially to provide high recognition accuracy in practical applications [1]. There are numerous studies concerning the problem of robustness to additive noise conditions, that provide us with reasonable guidelines for noisy speech recognition [2, 3]. However, many techniques are based on the assumptions of ideal or artificial noise conditions such as white additive noise. As a result their use in practical applications (e.g. *colored* or *band-limited* noise) is limited.

Traditionally, speech recognition is performed by extracting a set of acoustic feature vectors, which are calculated from the whole frequency band of input speech. Even if only a part of the frequency band is corrupted by noise, all the feature vector components are affected. Recently, there have been a few studies which model sub-band features independently [4, 5]. The acoustic likelihoods are computed independently for each sub-band, and then combined before classification. Their preliminary experimental results showed robustness under noisy/mismatched conditions.

We believe that multi-band ASR should be investigated for the following reasons:

- There is a psychoacoustic evidence, as analyzed in a recent paper by Allen [6]. In the paper, he mentioned *The Independent-Channel Model* introduced by Fletcher *et al.* According to Fletcher, human beings processes narrow frequency sub-bands independently of each other in auditory perception. At some point of the processing, the outputs from each sub-bands are recombined into a global decision.
- Statistical models of sub-band features may be more accurate than full-band models, because of the higher dimensionality of the full-band feature space (curse of dimensionality).
- Ambient noise may be *colored* and severely corrupt only few frequency bands. Sub-band recombination strategies can be designed to reduce the corrupted sub-band contribution to the classification decision.

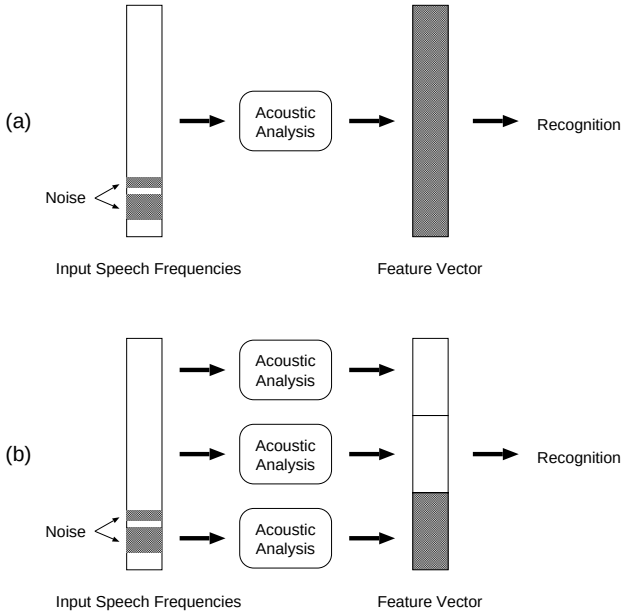

Figure 1: Schematic diagrams of (a) full-band ASR (conventional), and (b) multi-band ASR. The input speech is partially corrupted by band-limited noise, which spreads over all features in (1), but only the corresponding band in the case of (b).

Figure 1 explains the main motivation and basic concepts of multi-band ASR. The input speech is here corrupted by low frequency noise. All the feature vector components obtained by conventional acoustic analysis are affected by the noise. In the multi-band approach, however, only the feature vector corresponding to the corrupted frequency band is corrupted by the noise, the information in the other bands is not affected.

[*] Also affiliated with Waseda University, Tokyo, Japan. Since April 1998, he is with Chiba Institute of Technology, Narashino, Japan

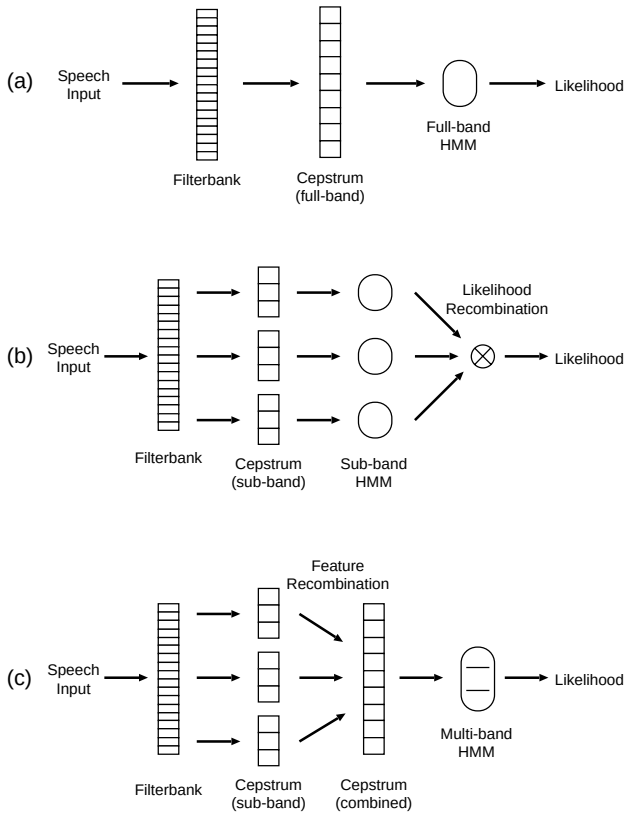

Figure 2: Diagrams of (a) full-band ASR (conventional), (b) multi-band ASR (likelihood recombination), and (c) multi-band ASR (feature recombination).

2. MULTI-BAND ASR

The basic strategy of multi-band ASR is to recognize speech by using multiple frequency bands whose acoustic features are extracted individually. The original idea of this approach was proposed by Bourlard and Hermansky *et al.* [4, 5]. They basically applied different classifiers for each band, then recombine the likelihoods at some recombination level such as HMM state, phone, word, with or without weighting functions.

In our work, we introduce another scheme to recombine the multiple inputs, which composes a single feature vector by joining the sub-band feature vectors together. Therefore, instead of sub-band *likelihood* recombination we use sub-band *feature* recombination. The advantage of this approach is: (1) it is possible to model the correlation between each sub-band feature vectors, (2) acoustic modeling becomes simpler, (3) we can avoid considering complicated weighting strategies.

Figure 2 illustrates basic concepts of the conventional ASR, multi-band ASR with likelihood recombination and multi-band ASR with feature recombination.

2.1. Acoustic Analysis for Multi-Band

As frontend of the recognizer, we use filterbank analysis, then mel-cepstrum analysis based on the DCT *(Discrete Cosine Transform).* In the full-band system, the DCT is applied to the whole filterbank to obtain a series of the mel-cepstrum feature vector. In the multi-band system, the filterbank output is split into several disjoint bands, then the DCT is applied to each of the sub-bands individually (see Figure 2-(b), (c)). In the case of the feature recombination, a single mel-cepstrum vector is created by combining all of the sub-band mel-cepstrum vectors.

2.2. Likelihood Recombination (LC)

In the likelihood recombination approach, each sub-band is modeled independently. During the recognition process, different speech classifiers are applied also independently to each sub-band, and each classifier provides a set of recognition hypotheses and recognition scores. Then all classifier outputs are combined to obtain global recognition scores and a global decision.

According to Bourlard [4], recombination at the HMM state level gives almost the same accuracy as recombination at higher levels like phone, syllable or word level. State level recombination is obviously much simpler to implement. Therefore, in this study we adopt the HMM state as the recombination level.

Let o_i and s_j be an observation vector at frame (time) i and an HMM state j. After calculating 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:

$$p(o_i|s_j) = \prod_b p(o_i^b|s_j). \quad (1)$$

However, it seems very improbable that all sub-band features have the 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.

A solution [7] is to weight the contribution from each sub-band using *probability exponents* as follows:

$$p(o_i|s_j) = \prod_b p(o_i^b|s_j)^{w_b}, \quad (2)$$

where w_b is the weighting factor corresponding to the sub-band b. In this paper, we investigate weights computed from the sub-band signal-to-noise ratio (SNR) and from the inverse conditional entropy of each band.

2.3. Feature Recombination (FC)

The main difference between the traditional (full-band) approach and the multi-band feature recombination approach is at the acoustic analysis level. After the cepstral feature for each sub-band is extracted individually, they are combined into one single vector, which is the input to the classifier. Intuitively, feature recombination gives both the advantages of the conventional ASR and of the multi-band ASR with likelihood recombination, namely:

- Band-limited noise affects only few of the feature components, as with likelihood recombination.
- All feature components can be jointly represented by statistical models, without any independence assumption, as in conventional ASR.

Obviously feature recombination can be performed only at the state level.

3. EXPERIMENTS

We use ARPA's ATIS *(Air Travel Information Service)* continuous speech recognition task to test the multi-band approach. The speech data is recorded with a close-talk microphone in laboratory environments. The training dataset consists 19,507 sentences by 528 speakers. We run ASR experiments on the official Dec.94 test set of 981 sentences.

Our recognizer is based on AT&T's ATIS Speech Recognizer [8]. In the full-band (referred as conventional) experiments, we use context independent phone HMM's with 3 states, 16 mixture Gaussian distribution, and a word bigram language model. The frontend is based on the mel-cepstrum analysis of the input speech sampled at 16kHz. The digitized waveform is analyzed with a 20ms window, that is shifted by a 10ms interval. Through the FFT computation, we obtain 31 mel-frequency energy components, that are processed by the cosine transform to provide vectors of 12 mel frequency cepstrum coefficients (MFCC) at a 100Hz frame rate. For every input sentences, we subtract from all the MFCC vectors the average (per sentence) MFCC vector *(Cepstrum Mean Subtraction).*

Every frame feature vector is made of 39 components, consisting of the 12 MFCC's and of the frame energy in dB with their 1st and 2nd derivatives. In the multi-band experiments, we used a number of mel-cepstrum features per band proportional to the number of filters in each band.

In the full-band based (conventional) system, there are 31 filterbanks as an input. For the multi-band approach, we use 2, 3, 4 and 6 sub-bands, defined by equal partitions of the mel-frequency scale:

- 2 bands: (0-1850) (1691-8000) Hz
- 3 bands: (0-1155) (1050-2996) (2723-8000) Hz
- 4 bands: (0-950) (850-1860) (1691-3625) (3295-8000) Hz
- 6 bands: (0-650) (550-1155) (1050-1860) (1691-2996) (2723-4824) (4386-8000) Hz

In our ASR experiments, we add several types of noise onto clean speech data to test the recognizer under *mismatched* conditions. We add the noise to the test speech waveform, but not to the training data. The HMM's are always trained under *ideal* (no noise) conditions.

3.1. Likelihood Recombination (LC) with Weights

In this section, we evaluate multi-band ASR by likelihood recombination, in which all classifier outputs are recombined at each HMM state level. We add "lp-white" noise at 10dB SNR to test data. "Lp-white" noise is an *ideal* type of noise, which is white noise added only to the first frequency band, by applying an FIR filter. Three sub-bands are used for all experiments in this section.

The acoustic likelihood of the three bands are recombined according to Equation (2) using weights w_b that are: (i) constant, (ii) equal to sub-band SNR computed at the frame level, and (iii) the inverse of the conditional entropy of each sub-band. The sub-band SNR (ii) is computed using the background noise level estimated from minimum energy frame in the sentence. The conditional entropy (iii) is computed from the *a posteriori* probabilities of all HMM states. Sub-band weights are equal to the inverse of the conditional entropy. All weights are normalized to sum up to the number of sub-bands.

System	Word error %
Full-band (conventional)	19.4
No weighting ([1:1:1])	14.5
Sub-band SNR weighting	14.3
Entropy weighting	14.0
Constant weighting [0:1:1]	16.4
Constant weighting [0.5:1:1]	14.1
Constant weighting [0.75:1:1]	13.9

Table 1: Word error rate for various weighting strategies: likelihood recombination, 3-bands, added "lp-white" noise at 10dB level.

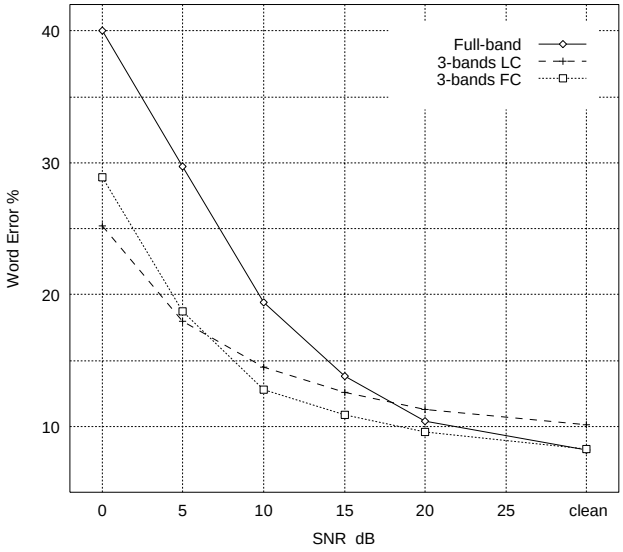

Figure 3: Word error rate for "lp-white" noise with various SNR's on full-band, 3-band LC and 3-band FC system.

Table 1 shows the recognition accuracy (word error rate) for the various weighting strategies. In the table, "constant weighting" refers to the use of a constant value for each band, as shown. The "no weighting" system has constant weights equal to [1:1:1]. Since the "lp-white" noise includes white noise only in the first sub-band, to reduce the weight of the first band seems reasonable.

For "lp-white" noise, the recognition accuracy significantly improves (from 19.4% to 14.5% word error) by using the three band system with "no weighting." Further modest improvement is observed when we apply (ii) sub-band SNR weighting (relative error reduction by 1.4%) and (iii) entropy weighting (3.5%).

Using the sub-band SNR as weights is a reasonable assumption. However, we still have some difficulty to estimate the SNR precisely, especially when the additive noise is nonstationary. The entropy weighting is also reasonable and intuitive from the point of view of information theory. Weighting the contribution of each sub-band is a promising approach but further work is required into investigating an effective weighting scheme.

Noise	Full-band	LC		FC	
		2bands	3bands	2bands	3bands
babble	21.5	20.6	22.1	19.3	18.9
buccaneer1	37.1	35.7	50.5	34.3	42.6
buccaneer2	37.9	43.3	59.9	41.3	57.4
destroyerengine	30.6	29.3	29.0	25.2	25.9
destroyerops	20.0	21.9	26.4	19.5	19.7
f16	31.7	30.0	35.6	28.2	30.7
factory1	29.5	28.4	32.9	26.6	28.0
factory2	15.5	15.7	17.1	13.6	13.8
hfchannel	36.9	39.0	43.8	34.0	33.9
leopard	10.8	11.9	12.6	11.0	10.9
m109	14.5	15.1	16.1	13.3	13.3
machinegun	13.4	11.6	12.2	11.0	10.9
pink	35.5	35.2	43.2	34.0	41.5
volvo	9.0	10.0	11.1	8.8	9.0
white	40.8	52.3	66.5	50.5	65.5

Table 2: Word error rate for various types of noises with full-band, two and three band LC and FC. The noise is added to clean speech data at 10dB level.

3.2. Robustness to Various Types of Noise

In this section, we investigate the recognition performance of the likelihood recombination (LC) and feature recombination (FC) multi-band approaches for various types of noise. In addition to "lp-white" noise, other types of noise (babble, buccaneer1, destroyer-engine, *etc.*) from the NOISEX-92 database are added to the test data.

Figure 3 shows the change of the recognition accuracy for "lp-white" noise with various SNR's, for LC and FC. FC is more accurate except for very low SNR's (0-5dB). Note that the LC system assumes complete independency between each sub-band likelihood. On the other hand, the FC system models the correlation between each sub-band.

Table 2 summarizes the recognition results for 15 kinds of additive noises using two and three band LC and FC as well as the full-band system. In each case, the noise is digitally added to the clean speech data at 10dB SNR level.

The FC system gives better performance than the LC system, for all noise conditions in Table 2. Both two and three band system implementations perform better than the baseline full-band system for most types of noise. The two band FC system gives the best overall results. The performance improvement over the baseline full-band system depends on the type of noise and goes up to 18% error reduction for "destroyer-engine" type of noise. The best results for the multi-band system are obtained for the ideal "lp-white" noise case (see Figure 3), where a 25% error reduction over the full-band system is achieved.

For several noise types such as "babble," "destroyer-engine," "factory2," "hfchannel," and "machinegun," the multi-band system gives better accuracy. These noise types have similar characteristics, with signal energy concentrated on portions of the frequency spectrum. On the other hand, the multi-band approach is less accurate (up to 25% error increase for "white" noise case) than the conventional ASR under conditions like "buccaneer2," "leopard," "pink" and "white" noise, in which the noise energy is spread all over the frequency spectrum. This result agrees with [9].

4. CONCLUSION

In this paper we studied the multi-band speech recognition method. In particular we examined two different approaches.

1) Likelihood recombination, in which the sub-band likelihoods are considered independent.
2) Feature recombination, in which acoustic analysis is applied to each band individually, and the resulting sub-band feature vectors are modeled jointly.

We performed several ASR experiments after adding different kinds of noise signals to the input speech. In general, we found that multi-band ASR is more robust than conventional ASR when the corrupting noise is concentrated on a portion of the spectrum. We have also shown that the proposed feature recombination is more effective than the likelihood recombination, at least in our HMM framework.

Acknowledgments

We acknowledge David Roe and Rick Rose for many useful discussions. The first author thanks the Japan Society for the Promotion of Science (JSPS) for their support.

5. REFERENCES

[1] J.-C. Junqua and J.-P. Haton. *Robustness in Automatic Speech Recognition — Fundamentals and Applications*. Kluwer Academic Publishers, Boston, 1996.

[2] D. Van Compernolle. Increased noise immunity in large vocabulary speech recognition with the aid of spectral subtraction. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1143–1146, 1987.

[3] A. Varga and R. Moore. Hidden Markov model decomposition of speech and noise. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 845–848, 1990.

[4] 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.

[5] H. Hermansky, S. Tibrewala, and M. Pavel. Towards ASR on partially corrupted speech. In *Proc. Int. Conf. on Spoken Language Processing*, pages 1579–1582, Philadelphia, October 1996.

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

[7] 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.

[8] E. Bocchieri, G. Riccardi, and J. Anantharaman. The 1994 AT&T ATIS CHRONUS recognizer. In *ARPA Spoken Language Systems Technology Workshop*, pages 265–268, Austin, January 1995.

[9] 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.