SUB-BAND BASED RECOGNITION OF NOISY SPEECH

Sangita Tibrewala[1] *Hynek Hermansky*[1,2]

[1]Oregon Graduate Institute of Science and Technology
Portland, Oregon, USA.
[2]International Computer Science Institute,
Berkeley, California, USA.

ABSTRACT

A new approach to automatic speech recognition based on independent class-conditional probability estimates in several frequency sub-bands is presented. The approach is shown to be especially applicable to environments which cause partial corruption of the frequency spectrum of the signal. Some of the issues involved in the implementation of the approach are also addressed.

1. INTRODUCTION

When speech signal is partly degraded e.g. by a frequency selective noise, some part of the speech spectrum may still carry a valid information. A typical signal representation used in automatic speech recognition (ASR) consists of a series of feature vectors, each vector representing the entire short-term frequency spectrum at a given time instant. Even one or a few corrupted elements in the feature vector lead to severe degradation of the recognition performance.

Earlier work by Fletcher on articulatory index [1] (review in [2]) suggests that the human auditory mechanism decodes the linguistic message independently in different frequency sub-bands and the final decision is based on merging the information from these sub-bands. One interpretation of Fletcher's suggestion can be that as soon as any sub-band combination yields sufficiently confident and reliable information, the information from the remaining (possibly corrupted) sub-bands does not have to be used for subsequent decoding of the linguistic message. ASR machine could benefit if it had the human ability to de-emphasize the unreliable frequency sub-bands. Towards this end we have recently proposed and started to investigate a new ASR approach which utilizes several information sub-streams.

2. SUB-BAND ASR MODEL

In the sub-band model (Fig. 1) the frequency spectrum is divided into several sub-bands. Independent probability estimation for each class is done in each sub-band. The class conditional probability estimates from each sub-band classifier are then merged to give the final result. This scheme allows for selective de-emphasis of unreliable sub-bands.

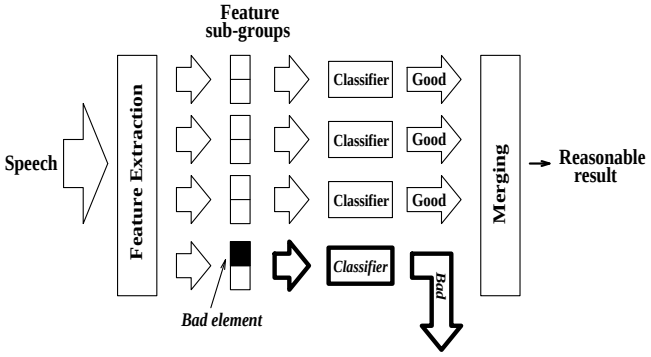

Figure 1. Sub-band Model

3. ISSUES INVOLVED IN THE SUB-BAND MODEL

There are several issues involved in designing the sub-band model: 1) the definition of the frequency sub-bands, 2) the features to be used in each sub-band, 3) the temporal unit at which information should be merged, 4) the merging technique.

Our preliminary experiments (reported in [3]) were based on a 13-word vocabulary consisting of the ten isolated digits and control words (yes,no) from the Bellcore digit database. The training set consisted of 150 speakers and 50 speakers comprised the test set. Each speaker uttered the vocabulary once. The features used were the power spectrum values obtained after the PLP critical band filter analysis followed by cube-root compression, and equal loudness equalization [4]. Each of the sub-band classifiers was a phoneme-based HMM/MLP hybrid classifier [5]. The input to the merging process were the sub-band class conditional log-likelihoods derived for all 13 classes in each sub-band.

3.1. Merging Techniques

A critical part of the sub-band model is the merging technique of the sub-band classifier outputs. We used

the sub-band class-conditional log-likelihoods (13 isolated words for word-level merging and 61 phonemes for frame-level merging) as inputs to the merging process and experimented with both linear and non-linear classifiers for the information merging. The merging using a non-linear classifier (multi-layer perceptron (MLP)) trained on the log-likelihoods from the training data systematically outperformed all linear schemes. The MLP used had a 3-layer architecture e.g. for 7-band word-level merging it had 91 inputs (13-words X 7-bands), 26 hidden units and 13 outputs.

3.2. Choice of the frequency sub-bands

Another choice to be made is the number of sub-bands and the frequency range spanned by each sub-band. Narrower sub-bands may allow greater flexibility in isolating frequency-localized degradation but the class discrimination within the sub-band decreases with decreasing amount of information in narrower sub-bands.

We experimented with 2,4 and 7 sub-bands. In the 2 sub-band system each band had approximately 7 critical bands (frequency ranges are 0-1140 Hz, 1046-4000 Hz). The 7 sub-band system had roughly two critical bands per sub-band as suggested by Allen [2] (0-360 Hz, 330-640 Hz, 580-950 Hz, 860-1360 Hz, 1265-1920 Hz, 1800-2700 Hz, 2515-4000 Hz). The 4 sub-band system used bands in 0-765 Hz, 700-1640 Hz, 1515-2700 Hz, and 2100-4000 Hz ranges. The sub-bands overlapped to some extent due to overlapping shapes of the underlying critical bands.

Results (Table 1) show that as the sub-bands become narrow, the performance in the individual sub-bands decreases, since the information content in each sub-band decreases. However, for all studied cases, the performance of the sub-band system is at least equivalent but likely somehow better than the performance of the conventional full-band system.

Baseline	4.6	
Recognizer	*Individual sub-bands*	*MLP merged system*
2-band model	11-13	2.46
4-band model	27-53	2.62
7-band model	30-60	4.31

Table 1. Error % with different number of sub-bands (merging at the word level).

3.3. Features

Throughout most of our work we have been mostly concerned about relative benefits of the multi-band approach. Therefore, rather than striving for the lowest error rates, we used the simple static cubic-root compressed short-term critical-band power spectrum energies as the sub-band features for most of the isolated digit experiments reported in this paper. However, to demonstrate a potential for further improvements of our systems we also ran a comparative experiment with cepstral coefficients of the all-pole (PLP) model of the above features. The mean-subtracted cepstral coefficient features coupled with the delta features and energy features improved the performance (Table 2). Such PLP cepstral sub-band features were subsequently used in our large-vocabulary (SWITCHBOARD database) experiments.

Recognizer	*Critical band energies*	*cepstral features*
Baseline	4.6	1.5
4-band model	2.62	2.0
7-band model	4.31	2.15

Table 2. Error % with different features (merging at the word-level)

3.4. Merging level

In some of our initial experiments which used word-level merging we have observed noticeable differences in optimal sub-band paths. Thus, we have speculated that one of the possible advantages of the sub-band model could be a relaxation of the temporal synchrony between different sub-bands which can be achieved by merging at higher-than-frame (i.e. phoneme, syllable, word, phrase,..) levels. After experimenting with both the word level and the frame level merging, we have so far not seen any benefits from relaxing the time-synchrony between sub-bands. As a matter of fact the best results were obtained by a simple merging at the frame level. However, the differences in performance are insignificant and further experiments are required to make any conclusions on this issue. (Table 3)

Recognizer	*Word-level*	*frame-level*
4-band model	2.0	1.7
7-band model	2.15	1.4

Table 3. Error % using merging at different levels

4. EXPERIMENTS ON SWITCHBOARD DATABASE

When adopting the frame-level merging, an extension of the multi-band approach to the large-vocabulary continuous-speech phoneme-based ASR is straightforward since only the initial frame-level classification needs to be modified.

The training data in our experiments using the conversational speech (SWITCHBOARD) database consisted of 4 hours of male speech. Test data were 240 male-speaker utterances. The merging network was trained on an independent set of 2 hours male speech[1]. We experimented with 4-band model and 7 band model. The sub-bands of the 7-band model were as defined earlier while the 4-band model had

[1] thanks Fred Jelinek for the suggestion of using independent set

sub-bands defined such that the frequencies from 0-200 Hz were not included (telephone quality speech). The bands were 180-670 Hz, 550-1200 Hz, 985-1980 Hz and 1670:4000 Hz. All the classifiers were trained using the HMM/MLP hybrid training software from ICSI[2] and the decoding was done using the lattice-decoder of STRUT[3].

Baseline	60.9	
Recognizer	*7-band model*	*4-band model%*
Individual sub-band classifiers	68-73	67-69
MLP merging	59.0	59.5

Table 4. Error % on the switchboard database.

The MLP-based merging again yielded the best performance with about 2% absolute improvement in the error rates. Also, consistent with our earlier experiments on isolated digits, there is no significant difference in performance between the 4-band model and the 7-band model. While we also experimented with various linear merging schemes, they again were not as effective as the non-linear merging.

This improvement in performance should not be attributed only to the multi-band paradigm. When we trained an MLP on independent data to re-classify the outputs of the baseline system, the performance of the baseline system also improved to 59% error which was similar to what we achieved by MLP merging of the sub-bands. However, we can conclude that the sub-band model can be adopted for large-vocabulary continuous-speech ASR where it yields performance similar to the conventional full-band baseline system.

5. EXPERIMENTS ON SPEECH CORRUPTED WITH ADDITIVE NOISE

Results reported in previous sections indicate that for well-matched training and test conditions, there is no loss of performance from the multi-band approach. However, as already indicated by our earlier work [3] and as further amplified below, the multi-band paradigm offers a significant advantage when there is a mismatch between training and test conditions.

5.1. Initial experiments with sinusoidal noise

In earlier reported work [3] we had experimented with various subsets of the sub-bands of the 7-band model. Independent nonlinear merging networks were trained on the clean training data using all possible combinations of subsets of the 7 sub-bands. The system thus consisted of 127 different merging networks. When a sinusoidal noise at 900 Hz was added to the original test set at different SNRs ranging from 30dB to 0dB the results showed that there exists at least one sub-band combination (among the available 127) which is capable of yielding a very good result even in the presence of a significant degradation by selective noise.

Several techniques for integrating the decisions of the 127 trained networks were investigated, among them SNR thresholding (i.e. leaving out the sub-bands which yielded SNR estimate [6] below a certain threshold) yielded results close to that obtained by cheating when picking the best merging network. Besides SNR thresholding we have also investigated several other strategies based on the outputs from the 127 merging classifiers. All decision techniques yielded results which were better than results of the conventional full-band recognizer.

5.2. Experiments with Real Noise

To test the performance of the sub-band model with real noise we used some of the noise samples (factory2, destroyer-engine, pink, white, volvo, babble and high frequency radio channel noise) from the NOISEX-92 database. Each noise was added to test speech data after being scaled so that the performance of the conventional full-band ASR noticeably degraded (from the baseline error of about 5% to about 25%).

Three decision strategies adopted from our previous work with sinusoidal noise were investigated, namely: 1) Using all 7 sub-bands in a single merging classifier trained on the clean data (i.e. the strategy optimal for the uncorrupted speech), 2) SNR thresholding (i.e. leaving out the sub-bands in which the on-line estimated SNR was worse than 5 dB), 3) a majority vote among the 127 combining networks which represented all possible combinations of sub-bands. In addition, we have also experimented with adaptation for selection of the best sub-band combination for a specific noise situation. For this, 10 words from a single speaker were used to pick the best combination for the particular noisy condition. In cases where more than one sub-band combination was selected by the adaptation process, the vote among all selected combinations was used to make the decision.

The Fig. 2 shows averaged error rates for 5 noise cases viz. destroyer-engine, factory, pink, babble and volvo noise samples. Each of these noise conditions was characterized by frequency-selective degradation, such that at least some of the sub-band classifiers exhibited very little degradation. The 7-band MLP system yields about half the error compared to the conventional full-band baseline system. The results obtained using the adaptation are similar to results obtained by the SNR thresholding and the majority voting.

We repeated the experiment with baseline performance of approximately 15% and conclude that for moderate noise, all decision-making techniques yield

[2] ICSI is The International Computer Science Institute at The University of California at Berkeley

[3] STRUT is Speech Training and Recognition Unified Tool developed at the Faculté Polytechnique de Mons

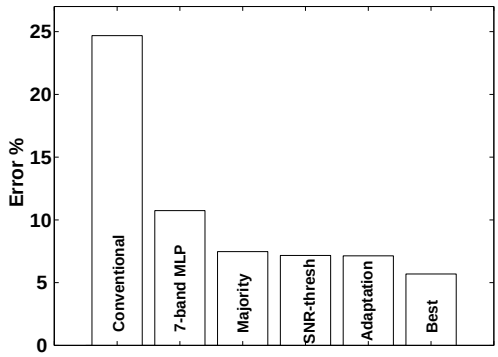

Figure 2. Average error rates for speech corrupted with factory, destroyer-engine, volvo, babble and pink noise. ("Best" refers to the best sub-band combination picked-up by "cheating" when looking at performance on the test data.)

only slight improvement over the 7-band MLP performance. However, when we corrupted the test data with stronger noise (so that the performance of the conventional full-band recognizer degraded to about 80% error) the 7-band MLP was not very efficient while the other techniques still yielded significant improvements over the full-band approach. The result obtained by using the 7-band MLP is important since it supports the notion of some inherent noise robustness of the multi-band approach (already observed by Bourlard et al. [7] for the linear merging case) even when we employ the MLP merging which we previously found to be more efficient on the clean data.

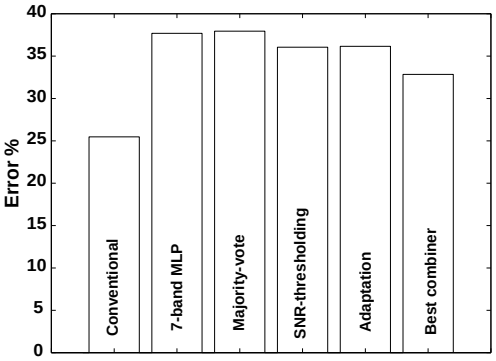

Figure 3. Average error rates for speech corrupted with white, high-frequency radio channel and car noise.

However the multi-band approach was ineffective (Fig. 3) for white noise, high frequency channel noise and one type of car-noise (recorded in our laboratory). A common characteristic of these noise conditions was that all the sub-bands were significantly corrupted by noise and hence exhibited significant degradation in performance. For these noise cases the baseline system performed better than the multi-band ASR.

6. CONCLUSION

For uncorrupted speech (same training and testing environment), the sub-band approach yields results which are at least similar but likely better than results from the conventional ASR system. More importantly, it improves performance under degradation by frequency localized noise. For moderate noise levels, the multi-band system itself, without use of any prior or posterior decision-making, significantly reduces error rates. For stronger noises, decision-making techniques using either prior information about the signal (such as use of adaptation speech or estimation of sub-band SNR), or posterior decision techniques such as majority vote, appear to be efficient. With merging done on the frame level, the approach generalizes to conversational speech.

7. ACKNOWLEDGEMENTS

We acknowledge the collaboration of Misha Pavel, Nikki Mirghafori, Nelson Morgan, Steve Greenberg, Christophe Ris, Stephane Dupont, Herve Bourlard, Mark Ordowski, and Jordan Cohen on this project. The work was supported by DoD (MDA-904-94 C-6169 and support during 1996 JHU Summer Workshop), and by NSF/ARPA (IRI-9314959).

REFERENCES

[1] Fletcher, H., SPEECH AND HEARING IN COMMUNICATION, New York: Krieger, 1953.

[2] Allen, J.B., "How do humans process and recognize speech?," IEEE TRANS. ON SPEECH AND AUDIO PROCESSING, vol. 2, no. 4, pp.567-577, 1994.

[3] Hermansky, H., Tibrewala, S. and Pavel, M., "Towards ASR on partially corrupted speech," PROC. ICSLP96, October 1996.

[4] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech," JOURNAL ACOUST. SOC. AM., vol. 87, no. 4, pp. 1738-1752, 1990.

[5] Bourlard, H. and Morgan, N., CONNECTIONIST SPEECH RECOGNITION — A HYBRID APPROACH, Kluwer Academic Publishers, 1994.

[6] Hirsch, H.G.: "Estimation of noise spectrum and its applications to SNR estimation and speech enhancement," TECHNICAL REPORT TR-93-012, International Computers Science Institute, Berkeley, CA, 1993.

[7] Bourlard, H. and Dupont, S., "A new ASR approach based on independent processing and recombination of partial frequency bands," PROC. ICSLP96, October 1996