Towards ASR On Partially Corrupted Speech

Hynek Hermansky, Sangita Tibrewala and Misha Pavel

Oregon Graduate Institute of Science & Technology,
P.O.Box 91000, Portland, Oregon

ABSTRACT

A new highly parallel approach to automatic recognition of speech, inspired by early Fletcher's research on Articulation Index, and based on independent probability estimates in several sub-bands of the available speech spectrum, is presented. The approach is especially suitable for situations when part of the spectrum of speech is corrupted. In such cases, it can yield an order-of-magnitude improvement in the error rate over a conventional full-band recognizer.

1. Introduction

The first step in most of the current automatic speech recognizers (ASR) is to convert the incoming speech signal into series of short-term vectors. Each vector represents a short segment of the signal. Each element of the vector describes some part of the information carried by the signal; for example each element of the short-term spectral vector represents energy of the speech signal in a given frequency range.

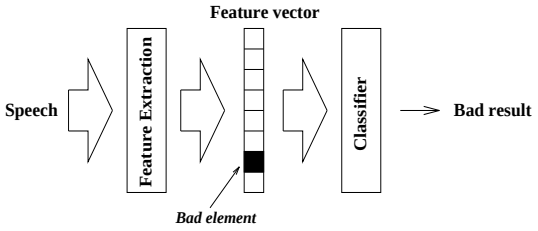

Figure 1: Conventional ASR.

A typical ASR uses the whole feature vector for its subsequent classification into one of the linguistic classes (phoneme, syllable, word, utterance,..). Time-alignment (e.g. Viterbi search, dynamic time warping) of the incoming stream of feature vectors is typically also involved in the classification.

2. The Problem

Consider the case when some of the elements of the short-term vector contain corrupted or even misleading information, while the remaining ones are still uncorrupted (Fig. 1). This can occur e.g. when the speech signal gets corrupted by selective noise. In the current main-stream ASR the entire feature vector is used as one entity and even a single corrupted spectral element can severely degrade the performance of the recognizer. This can be rather annoying to the user who has relatively little difficulty in understanding such partially corrupted speech [6].

3. The Human Way

Early Fletcher's work [5] on Articulatory Index (see [1] for a review) suggests that human auditory perception works differently than the current ASR. Specifically, Fletcher suggests that 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. According to Fletcher [5], the probabilities of erroneous recognition in the sub-bands $P\{E_i\}$ multiply to yield the overall error rate

$$P\{Error\} = \prod_i P\{E_i\}.$$

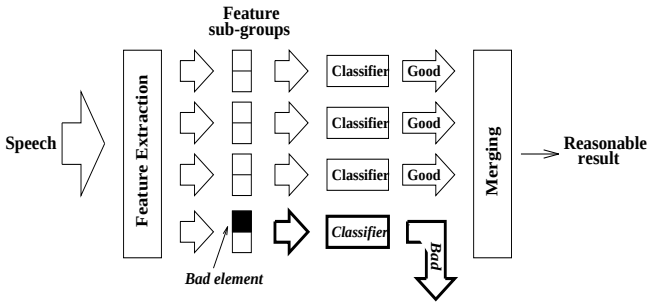

Figure 2: Sub-band Model

One interpretation of Fletcher's work is that as soon as any sub-band combination yields sufficient information, the information from the remaining (possibly corrupted) sub-bands does not have to be used for subsequent decoding of the message.

This notion receives some support from the recent work of Lippmann [8] who showed that when the important frequency band (800-4000 Hz) is left out of the decoding process, neither of the remaining low or high frequency bands alone has sufficient information to

decode the linguistic message. In contrast, if both the high and low remaining bands are heard, then the message can be decoded with a reasonable accuracy. Similarly, Riener et al [9] showed that although the word intelligibility in sentences is about 23% in individual high and low 1/3 rd octave frequency bands, when the two bands are presented simultaneously, the intelligibility increases to 77%.

In many situations, an ASR recognition performance could be improved if the ASR would have the human ability to de-emphasize the unreliable frequency sub-bands in the decoding stage; provided the remaining clean sub-bands could supply sufficiently reliable information. Furthermore, this approach could be beneficial if some sub-bands were inherently better for certain classes of speech sounds than others. Possibly, even different recognition strategies could be applied in different sub-bands.

This approach requires the use of information about the reliability of different sub-bands. For humans the decision about reliability of a particular frequency band could rely on semantics and syntax of the linguistic message. In fact, it may not be unreasonable to expect ASR to be eventually able to do the same. Moreover, classifiers used in ASR should be eventually capable of delivering confidence levels for their decisions which could then be used in deciding on reliability of the information on which they act. Finally, signal processing techniques which can estimate noise level at particular segments of speech directly from the signal (see e.g. [10]) can supply the additional a priori knowledge about the reliability of the information in a given speech segment and sub-band.

This paper describes an approach (Fig. 2) to ASR based on the described sub-band paradigm. We show its feasibility and capability of functioning on partially corrupted speech with relatively minor loss of performance.

4. Sub-band ASR

In this study we subdivide the available speech spectrum into a number of frequency sub-bands and extract spectral features from each of the sub-bands. Recognition is done independently in each of the sub-bands; each recognizer yielding conditional probability estimates for all the classes to be recognized. These estimates are then merged to give the final result. The problem of how to merge the decisions from the sub-bands is nontrivial. The merging process should provide for de-weighting or entirely eliminating any of the unreliable sub-band recognizers from the final decision process (Fig. 2).

There are several issues involved in designing such a sub-band model: 1) the number of 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.

5. Experiments

Our experiments are based on a 13-word vocabulary consisting of the isolated digits (zero, oh, one, two, three, four, five, six, seven, eight, nine) and control words (yes,no) from the Bellcore database. We used 200 speakers from the database, with 150 speakers for training and 50 speakers for testing. The baseline system is the conventional recognizer of Fig. 1.

The issue of how to subdivide the available frequency band is open. The more and narrower the sub-bands, the better is the chance to alleviate frequency-localized degradation. On the other hand exceedingly narrow bands could yield a poor discrimination between the linguistic classes. We have experimented with about two sub-bands (roughly 7 critical bands per sub-band) and with 7 sub-bands (about 2 critical bands per sub-band as suggested by Allen [1]).

The features that we use are the power spectrum values obtained after the PLP critical band filter analysis followed by cube-root compression, and equal loudness equalization [7]. The sub-bands in the seven-band experiments have certain overlap due to overlapping shapes of the underlying critical bands. Each of the sub-band recognizers is a phoneme-based HMM/MLP hybrid recognizer [2].

In addition to the frequency subdivision, we have to decide on the temporal unit (frame, HMM state, phoneme, syllable, word, sentence,...) at which the merging should be done. In our experiments, we have merged classifier outputs at the word level (11 words in our digit vocabulary are monosyllabic and 2 words consist of two syllables).

We have experimented with both linear and non-linear merging techniques using sub-band logarithmic likelihoods derived for all the words in the vocabulary. For the linear merging technique, we have tried a weighted sum of the log likelihoods and have examined two types of weighting: 1) equal weighting, and 2) weighting by the accuracy of each vocabulary word in the individual sub-bands (weights were derived from the performance of the individual sub-bands on the cross validation data set). For the non-linear merging technique, we have used an MLP trained on the training data. The MLP merging technique gave the best results in our experiments.

Table 1 and Table 2 show the results on the original (uncorrupted) speech. The reported results are the average results on 4 jack-knifed portions of the database.

Recognizer	*Error %*
Baseline	3.85
sub-band 1 (111 : 1330Hz)	14.00
sub-band 2 (1220 : 4000Hz)	10.65
Linear Merging	
equal weights	4.2
accuracy weights	4.3
MLP Merging (26:26:13)	2.73

Table 1: Uncorrupted Speech: 2 sub-band System (frequency ranges and the MLP architecture as indicated in brackets).

These results demonstrate that partial information from the individual sub-bands can be successfully merged. Linear weighting schemes appear to be feasible for the 2-band approach but we were not able to use them successfully for merging the outputs from the narrow-band (7-band) classifiers. The best results were always obtained by using MLPs for merging.

Recognizer	*Error %*
Baseline	3.85
sub-band 1 (111- 485Hz)	58.93
sub-band 2 (426- 759Hz)	30.93
sub-band 3 (686-1110Hz)	34.05
sub-band 4 (1020-1580Hz)	33.1
sub-band 5 (1450-2200Hz)	29.7
sub-band 6 (1030-3050Hz)	40.95
sub-band 7 (2820-4000Hz)	49.8
Linear Merging	
equal weights	14.9
accuracy weights	12.58
MLP Merging (91:26:13)	3.77

Table 2: Uncorrupted Speech: 7 sub-band System (frequency ranges and the MLP architecture as indicated in brackets).

In comparison to the conventional full-band ASR approach, the improvement of the performance for the 2-band system using the MLP merging technique is significant relative to the standard error of a binomial distribution that was approximately 0.5%. The improvement for the 7-band system appears to be less significant. (The issue of the optimal sub-division of the available spectrum into sub-bands and the choice of appropriate features in each sub-band is to some extent studied in the companion paper [11]).

6. Towards ASR on partially corrupted speech

In the previous section we have shown that ASR in sub-bands is feasible and may be beneficial. The companion paper [11] suggests that linear weighting schemes are capable of superior performance in noise. However, in our experiments the non-linear MLP merging technique systematically outperformed the linear schemes and we are primarily interested in extending the MLP merging techniques for applications in the case of the environmental mismatch.

We are considering various nonlinear merging schemes which would allow for a selective weighting of partially corrupted bands. As the first step, however, we decided to investigate merging using a hard threshold approach i.e. switching-off the less reliable bands.

We conducted a series of experiments with various combinations of the sub-bands in the 7-band model. Two techniques were investigated. The first one was based on the MLP merging technique using subsets of sub-band recognizers (Fig. 2). The second technique was similar to the technique of Green et al. [6] and used a single recognizer trained on a limited number of sub-bands. Each experiment resulted in 127 different configurations (different MLPs acting on all possible combinations of classifier outputs of 7 sub-band recognizers in the first experiment and 127 different recognizers trained on different combinations of the 7 sub-bands features in the second one) to explore all possible combinations of the 7 sub-bands.

In both cases, there was a gradual decrease of performance with decreasing number of available sub-bands, supporting [6]. For a small number of dropped bands (3 or less), the performance of both schemes was comparable. For more than 3 bands dropped, the performance of the the MLP merging scheme was significantly better.

Fig. 3 shows the maximum, minimum, and median error rates obtained by the MLP merging scheme by dropping different number of sub-bands for all possible combinations of the sub-bands.

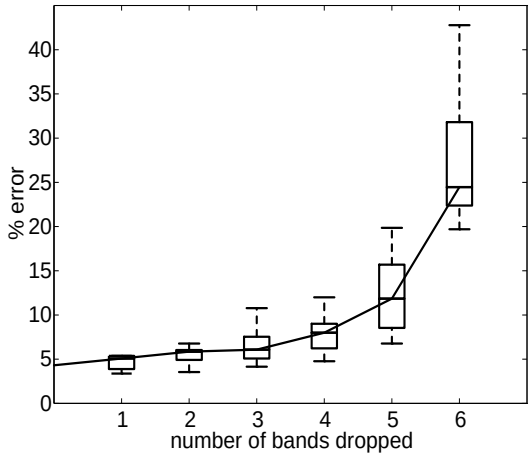

Figure 3: Range of error rates obtained with different combinations of sub-bands using the MLP merging technique

Typically, we observed that leaving out adjacent bands yielded maximum error rates while the minimum error rates were obtained when the remaining bands interleaved with the ones that were left out. Leaving out the first (111-485 Hz) band always resulted in improved performance, indicating the lack of reliable linguistic information in telephone speech at such low frequencies.

7. Experiments with frequency-localized additive noise

Sinusoidal noise at 900 Hz was added to the original test set at different SNRs. The outputs of all seven sub-band recognizers were merged using all 127 MLPs trained on the clean training data using all possible combinations of the 7 sub-bands from the previous experiment. The results of this experiment are shown in Fig 4.

7.1. How well we could do (if only..)?

..we could find the right sub-band combination The figure shows that the conventional sub-band recognizer seriously deteriorates with the increasing noise level. The sub-band recognizer which uses all seven sub-bands deteriorates somehow slower but still rather fast. When allowing for leaving out some of the sub-bands, the error rate for the noisy speech could improve dramatically (more than an order of magnitude !! for the 0 dB SNR).

7.2. Which sub-bands?

Adaptation: The results show that there exists at least one sub-band combination (among the available 127 ones) which is capable of yielding a very good result even in the presence of a significant

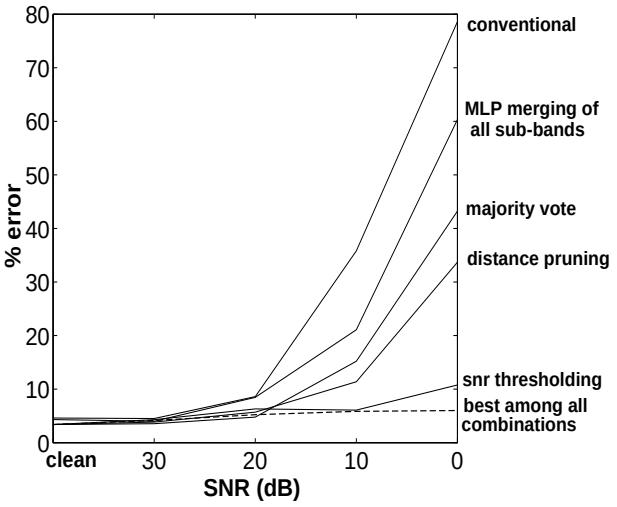

Figure 4: Performance of different merging techniques on speech corrupted with additive sinusoidal noise at 900Hz.

degradation by selective noise. This can be obviously used with advantage if the application allows for any adaptation to a given environment.

Prior knowledge about the signal: Another possibility we explored was to obtain some prior knowledge from the SNR estimates in the individual sub-bands [10] and to leave out the sub-bands with SNRs below a certain threshold (10dB in our case). When this can be done reliably (as was the case with our experimental sinusoidal additive noise) it could yield results close to the optimal (also shown in the Fig 4).

Combining decisions from the individual combinations: We currently do not have any definite way to do so. Majority vote among all available sub-band combinations yields about half the error rates of the conventional full-band approach. (This result is also shown in Fig 4).

If we had a reliable way to determine confidence for decisions from any given sub-band combination, then we would be able to prune sub-band combinations which had low confidence levels from the voting committee. We have tried several confidence level estimates. The best results were so far obtained by distance pruning based on the maximum difference between the 2 best outputs of the MLP used in merging. The top 15% of the MLPs having the maximum distance were used in the majority vote. (This result is also shown in the Fig 4).

Prior knowledge about the message: If a powerful language model would be available, it might also be used in pruning out the erroneous decisions from the individual sub-band combinations. We have not pursued this possibility yet.

8. Conclusion

Sub-band based recognition appears to be a powerful way to deal with partially corrupted speech. The system we have described in this paper can be readily used with adaptation and in situations where some estimate of the environmental noise is available. Even when these are not available, in situations where speech is corrupted by selective noise, the technique still significantly outperforms the conventional single-band recognizer. As our capability of deriving and using indications of reliability of decisions from individual sub-band recognizers improves, we expect further advances from using the proposed scheme.

9. Acknowledgements

This work was inspired by discussions with Jont Allen during the 1993 DoD Workshop on Frontiers in Speech Processing and developed in joint effort with Herve Bourlard and Morgan at ICSI Berkeley. We thank Johan Schalkwyk for discussions which led to MLP merging of subsets of the sub-band recognizers.

Finally, we thank the organizations that have provided support for our work over the years: DoD under their MDA 904-94-C-6196 the National Science Foundation and ARPA, through IRA-9314959, and the member organizations of CSLU.

10. REFERENCES

1. 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.
2. Bourlard, H. and Morgan, N., CONNECTIONIST SPEECH RECOGNITION — A HYBRID APPROACH, Kluwer Academic Publishers, 1994.
3. Cooke, M. P., P.D. Green, and M.D. Crawford, "Handling missing data in speech recognition," PROC. INTL. CONF. ON SPOKEN LANGUAGE PROCESSING 94 pp. 1555-1558, (Yokohama, Japan), 1994.
4. Duchnowski, P., A NEW STRUCTURE FOR AUTOMATIC SPEECH RECOGNITION, PhD Thesis, MIT, September 1993.
5. Fletcher, H., SPEECH AND HEARING IN COMMUNICATION, New York: Krieger, 1953.
6. Green, P.D., M.P. Cooke, and M.D. Crawford. "Auditory scene analysis and hidden markov model recognition of speech in noise," PROC. IEEE INTL. CONF. ON ACOUSTICS, SPEECH, & SIGNAL PROCESSING (Detroit, MI), pp. 401-404, 1995.
7. Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech," JOURNAL ACOUST. SOC. AM., vol. 87, no. 4, pp. 1738-1752, 1990.
8. R.P. Lippmann: Accurate consonant perception without mid-frequency speech energy, IEEE TRANS. ON SPEECH AND AUDIO PROCESSING, vol. 4, no. 1, pp.66-69, 1996.
9. K. R. Riener, R. M. Warren, and J.A. Bashford, Jr.: "Novel findings concerning intelligibility of bandpass speech," JOURNAL ACOUST. SOC. AM., 91 (4), S2339, 1992.
10. H.G. Hirsch: "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.
11. H. Bourlard and S. Dupont: "A new ASR approach based on independent processing and re-combination of partial frequency bands," PROC. ICSLP96, October 1996