TOWARDS SUBBAND-BASED SPEECH RECOGNITION

Hervé Bourlard[†,‡], *Stéphane Dupont*[†], *Hynek Hermansky*[*,‡], *and Nelson Morgan*[‡]

[†] Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez, B-7000 Mons, Belgium
Email: bourlard,dupont@tcts.fpms.ac.be

[*] Oregon Graduate Institute, Portland, OR, USA
[‡] Intl. Computer Science Institute, Berkeley, CA, USA

ABSTRACT

In the framework of hidden Markov models (HMM) or hybrid HMM/Artificial Neural Network (ANN) systems, we present a new approach towards speech recognition. The general idea is to split the whole frequency band (represented in terms of critical bands) into a few subbands on which different recognizers are independently applied and then recombined at a certain speech unit level to yield global scores and a global recognition decision. The preliminary results presented in this paper show that such an approach, even using quite simple recombination strategies, can yield at least comparable performance on clean speech while providing significantly better robustness in the case of speech corrupted by narrowband noise.

1 INTRODUCTION

Current ASR systems treat any incoming signal as one entity. Even when only a single frequency component is corrupted (e.g. by a selective additive noise), the whole feature vector is corrupted, and typically the performance of the recognizer is severely impaired.

The work of Fletcher and his colleagues [4] (see the insightful review of his work in [1]) suggests that human decoding of the linguistic message is based on decisions within narrow frequency subbands that are processed quite independently of each other. Recombination of decisions from these subbands is done at some intermediate level and in such a way that the global error rate is equal to the product of error rates in the subbands.

Whether or not this is an accurate statement for disparate bands in continuous speech (the relevant Fletcher experiments were done with nonsense syllables using highpass or lowpass filters only), we see some engineering reasons for considering some form of this subband approach:

1. The message may be impaired (e.g., by noise) only in some specific frequency bands. When recognition is based on several independent decisions from different frequency subbands, the decoding of linguistic message need not be severely impaired, as long as the remaining clean subbands supply sufficiently reliable information.

2. Some subbands may be inherently better for certain classes of speech sounds than others.

3. Transitions between more stationary segments of speech do not necessarily occur at the same time across the different frequency bands, which makes the piecewise stationary assumption more fragile. The subband approach may have the potential of relaxing the synchrony constraint inherent in current HMM systems.

4. Different recognition strategies might ultimately be applied in different subbands.

The approach may yield a recognizer that is basically independent of the phase between the different frequency subbands (up to the average duration of the speech units used for recombination).

Preliminary work in this direction has recently been reported, e.g., in [3]. Although the recombination scheme in [3] was quite simple, and no optimization of the frequency bands was performed, this work yielded results that were quite similar to the results of conventional full-band recognizers used for comparison. However, the resulting system was not tested for conditions of narrowband noise degradation (for which this kind of approach should prove to be most interesting).

The work described here has some similarity to this earlier work, though in the new effort there has been an attempt to (1) better formalize the problem from a statistical pattern recognition viewpoint, (2) determine the optimal way of recombining frequency subband recognizers, and (3) test the systems both under the "clean" condition and for the case of narrowband noise degradation.

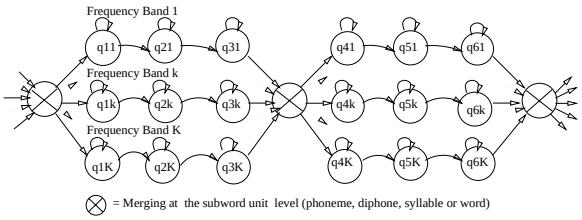

Figure 1: *General form of the HMM with intra-word anchor points. There are K frequency bands. For each frequency band, a HMM model or an ANN/HMM model is used. The merging is done at any level with an appropriate probabilities combination formalism.*

2 APPROACH

It is perhaps obvious that a core issue in the design of any subband-based system is the choice of the number and position of the constituent subbands. Once these are determined, the approach presented here will fundamentally consist of the combination of the output of multiple recognizers, one for each band, at some level of representation. Fundamentally, each of these recognizers consists of a probability estimator and a time-warp engine.

Of course, there is less information in a subband than in the whole band; the partial decisions may thus be less reliable. To avoid too much flexibility in choosing the time-warping path it is necessary to re-introduce some constraints at a higher level. This is done by forcing synchrony (in terms of the underlying segmentation) of the different independent frequency band recognizers at some level, as shown Figure 1. In other words, the scores of the different subband recognizers are recombined at a certain speech unit level (i.e., over a certain time segment) to yield a global score and a global decision. Up to now we have done this at the state, phoneme or word levels, although we are interested in looking at other units for this purpose (e.g., syllable). We note here that while this is quite easy at the HMM state level (and at the word level, in the case of isolated word recognition), it is no longer straightforward at any intermediate subword unit level (simply using the standard one-pass dynamic programming approach). Rather, the system can either use an approach based on the classic 2-level dynamic time warping algorithm, or else an adaptation of HMM decomposition [8]. This will be discussed further in an upcoming paper.

Although Fletcher's recombination criterion [1, 4] suggests an attractive optimum (since zero error in any band yields zero error overall), we are not aware of any statistical formalism for achieving this. Thus, we decided to perform statistical recombination of the frequency subbands according to

$$P(X|M) = f(\{w_k\}, \{P(X_k|M_k)\}) \qquad (1)$$

where M represents the word/sentence model, M_k the same word or sentence model for the k-th frequency band, X the full band acoustic vector sequence, X_k the sequence of (critical band) acoustic parameters for the k-th band, and w_k are the recombination parameters. In the work reported here we tested two different recombination functions $f(\cdot)$:

$$f(\cdot) = \sum_{k=1}^{K} w_k \log P(X_k|M_k) \qquad (2)$$

or

$$f(\cdot) = MLP_{w_k}(\log P(X_k|M_k)) \qquad (3)$$

where MLP_{w_k} represented a multilayer perceptron (MLP) parametrized in terms of w_k's and with $\log P(X_k|M_k)$, $\forall k$, at its input[1]. $P(X_k|M_k)$ thus represents the likelihood of a partial (frequency limited) sequence X_k given a HMM M_k. This can be computed with a standard HMM or a hybrid HMM/ANN. The latter solution has been considered in this paper.

3 RECOMBINATION STRATEGIES

In our current work, the recombination has been tested at both the HMM state level as well as at the word level.

Two different strategies have been considered for recombination at the state level:

1. Normalized phoneme-level recognition rates in each frequency band.

   Normalized phoneme-level recognition rates inside each frequency band were used as weighting factors. These weighting factors represent the relative amount of information (between 0 and 1) present in each frequency band for each speech unit class. They are normalized to sum to 1.

   These weights are computed on the clean training data set only and are not adapted to the test data. As later reported in Table 1, it is quite striking that this strategy alone already yields good robustness to narrowband noise.

2. Normalized S/N ratios in each frequency band.

   As usually done for spectral subtraction [7], the S/N ratio in each frequency subband was estimated on the basis of the subband energy histogram (see Figure 2 for an example). However, unlike the case of spectral subtraction, these histograms are used to compute the relative reliability of each frequency subband. As shown in the example of Figure 2, the distance between the two peaks of the histograms (for positive log S/N ratios, the lowest energy peak corresponding to silence+noise frames and the highest energy peak corresponding the speech+noise frames) is a function of the S/N ratio.

[1] Ultimately, we intend to use discriminant forms of (1)-(3).

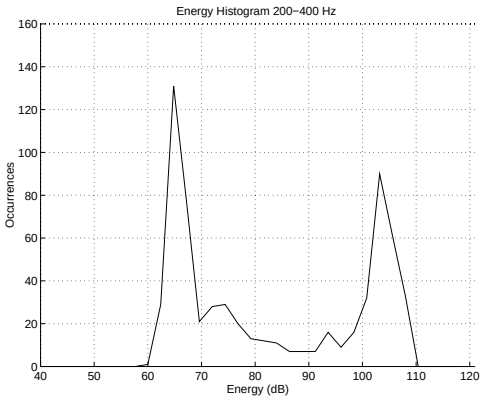

Figure 2: *Energy histogram for 6.5 sec of clean speech for the [200-400Hz] frequency subband.*

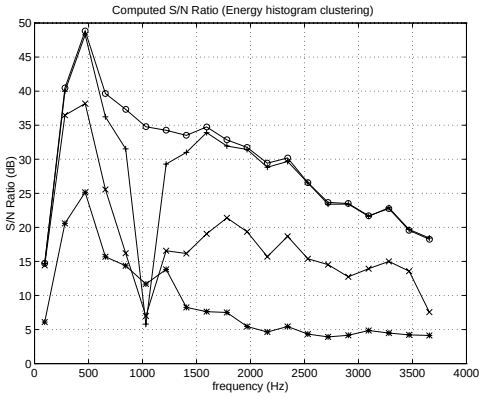

Figure 3: *Resulting freq. subband dependent SNR (20 equally spaced freq. bands) for: (o)=clean speech, (+)=added sine wave at 1000 Hz, (x)= Gaussian noise filtered around 1000 Hz, SNR=20dB, (*)= Gaussian white noise, SNR=10 dB.*

The S/N ratio in each subband is often estimated by fitting two Gaussians on the histogram. In the present work, we instead implemented a simple one-dimensional form of online clustering algorithm to compute and perform online adaptation of the two peak energies (E_1 for the lowest energy and E_2 for the highest energy). The SNR in each subband is then computed as the ratio $\frac{(E_2-E_1)}{E_1}$ in dB. Examples of resulting SNRs for 20 frequency subbands with different noise conditions are presented in Figure 3. These S/N ratios, normalized to sum up to 1, were used as w_k's in equation (2).

The recombination at the word level has been performed via an MLP, according to equation (2).

4 EXPERIMENTS

In the following experiment, we used 3-state phone models. 18 critical bands were used for the full band system and three subbands (spanning [0-1058], [941-2212], and [1994-4000] Hz) for the three subband HMM/ANN recognizers. Note that the overlap is only due to the critical band filter characteristics. Each band roughly encompasses one formant. The database consisted of 108 German isolated command words, telephone speech, with 15 speakers in the test set.

	FB	No-W	Acc-W	SNR-W	MLP
clean	3.6%	3.7%	3.7%	3.2%	2.7%
noisy	25.5%	9.2%	6.7%	6.3%	—

Table 1: *Isolated word recognition (108 German words, telephone speech) and noise was additive white noise in the 1st frequency band, 10dB SNR. Critical band energies were used as features. "FB" refers to regular full band recognizer; "No-W" refers to subband recombination at state level without any weighting; "Acc-W"= state recombination with weights proportional to phonetic subband accuracy; "SNR-W"= state recombination with weights proportional to automatically estimated subband SNR. The column "MLP" refers to subband recombination at word level using an MLP. All percentages refer to a test set with 15 speakers uttering each word once.*

The features used for each recognizer were critical band energies complemented by their first temporal derivatives, and 9 frames of contextual information were used at the input of the ANN. State level and word level recombinations were tested. In the case of word level merging, an MLP with 108 (words) $\times$ 3 (bands) input units and 108 output units was trained on normalized log likelihoods from the clean training data.

Results from the experiments are reported in Table 1. Recognition performance of the different recombination strategies are compared with the full band approach, in case of clean speech and noisy speech (additive white noise in the 1st subband, 10dB SNR). For clean speech we have been able to achieve results that were at least as good as the conventional full-band recognizer (though for this size test set the differences are not statistically significant at $p < .05$).

When one of the frequency bands is contaminated by selective noise, the multi-band recognizer yields much more graceful degradation than the broad-band recognizer. The best results have been achieved using weightings derived from S/N estimates. However, we have observed that even without any knowledge about the S/N ratio in subbands [using equal weighting ("No-W") or subband accuracy weighting ("Acc-W")] the subband recognizer still yields much better results than the conventional full-band recognizer.

5 CONCLUSIONS

In this paper, we presented the basis of our subband-based speech recognition system and preliminary experimental results. We believe that these results are quite striking and also particularly promising. These results have also been achieved with very little tuning and at an early stage of development for the method. Among other factors, we still have to consider:

- Number of frequency bands: So far, we have used 3 frequency bands roughly centered on the typical range for the first three formants. However, the number of these subbands, their width, as well as their possible overlap, still need to be optimized. The issue of number of subband is further discussed in [5].
- Recombination level: So far we tested the recombination at the HMM state level and at the word level. However, there could be another intermediate level that is more appropriate to this kind of approach. Candidate units could be the phone, the diphone, the demisyllable, or the syllable.
- Recombination criterion: So far we have mainly tested a likelihood based recombination.
- Weighting scheme: Other techniques able to estimate online the reliability of each frequency subband relatively to the others and taking larger time information into account should be investigated.
- Features: In this paper we have reported results using critical band energies. As will be reported in [2], we are now experimenting with cepstral processing of each frequency subband and this seems to further improve performance. In other work, we have recently observed that the subband approach also yielded better robustness to narrowband noise when compared to standard speech recognition approaches with noise cancellation capabilities. However, these techniques can also be applied to the subband approach. We are planning to combine subband approaches, which appear to have improved robustness to narrowband noise, with J-RASTA approaches, which appear to have improved robustness for steady-state wideband noise [6].

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., Dupont, S., Morgan, N., Mirghafori, N., and Hermansky, H., "A new ASR approach based on independent processing and recombination of partial frequency bands," to be published in *Proc. of Intl. Conf. on Spoken Language Processing*, Philadelphia, October 1996.

[3] Duchnowski, P., "A new structure for automatic speech recognition," *MIT PhD Thesis*, September 1993.

[4] Fletcher, H., *Speech and Hearing in Communication*, New York: Krieger, 1953.

[5] Hermansky, H., Pavel, M., Tibrewala, S., Morgan, N., Mirghafori, N., and Bourlard, H., "Towards ASR Using Partially Corrupted Speech" *Proc. of Intl. Conf. on Spoken Language Processing*, Philadelphia, October 1996.

[6] Hermansky, H. and Morgan, N., "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4 pp. 578-589, 1994.

[7] Hirsch, H. G., "Estimation of noise spectrum and its application to SNR-estimation and speech enhancement," *ICSI Technical Report TR-93-012*, Intl. Comp. Science Institute, Berkeley, CA, 1993.

[8] Varga A.P. and Moore R.K., "Hidden Markov Model decomposition of speech and noise," *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 845-48, 1990.