A NEW ASR APPROACH BASED ON INDEPENDENT PROCESSING AND RECOMBINATION OF PARTIAL FREQUENCY BANDS

Hervé Bourlard[†,‡] *and Stéphane Dupont*[†,1]

[†] Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez, B-7000 Mons, Belgium
Email: bourlard,dupont@tcts.fpms.ac.be
[‡] 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 automatic speech recognition (ASR). The general idea is to split the whole frequency band (represented in terms of critical bands) into a few sub-bands 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 better robustness in the case of noisy speech.

1. INTRODUCTION

Current automatic speech recognition (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 (see the insightful review of his work in [1]) suggests that human decoding of the linguistic message is based on decisions within narrow frequency sub-bands that are processed quite independently of each other. Recombination of decisions from these sub-bands 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 sub-bands.

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 sub-band 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 sub-bands, the decoding of linguistic message need not be severely impaired, as long as the remaining clean sub-bands supply sufficiently reliable information.
2. Some sub-bands 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 sub-band 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 sub-bands.

Preliminary work in this direction has recently been reported, e.g., in [4]. Although the recombination scheme in [4] 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).

As initially discussed in [2], the work described here presents an attempt to (1) better formalize the problem from a statistical pattern recognition viewpoint, (2) determine the optimal way of recombining frequency sub-band recognizers, and (3) test the systems under clean and noisy conditions.

2. APPROACH

It is perhaps obvious that a core issue in the design of any sub-band-based system is the choice of the number and position of the constituent sub-bands. 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 sub-band 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

[1] Supported by a F.R.I.A. grant (Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture).

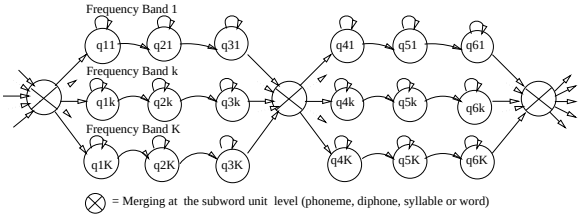

Figure 1: *General form of a K-band recognizer with anchor points between speech units (to force synchrony between frequency bands).*

different independent frequency band recognizers at some level, as shown Figure 1. In other words, the scores of the different sub-band 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, syllable or word levels, although we are interested in looking at other units for this purpose. 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 2-level dynamic time warping algorithm, or else an adaptation of HMM decomposition [8] (initially introduced to decompose the speech signal into a speech part and a noisy part). In the framework of sub-band-based speech recognition, a similar HMM decomposition formalism can be used to do multi-dimensional time warping and recombination of the frequency sub-bands. However, as opposed to standard HMM decomposition, it is not the same input signal that is fed into the different HMM models but different band limited versions of the original speech signal.

Although Fletcher's recombination criterion [1] 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 define the log-likelihood of a full-band acoustic vector X given a word (sentence) model M as

$$\log P(X|M) = \max_{J, M^j} \sum_{j=1}^{J} \log P(X_j|M^j) \tag{1}$$

where X_j represents the j-th segment of X and M^j the model associated with X_j during dynamic time wrapping. Depending on the recombination level, M^j could be a HMM state model, a word model, a phone model or any other subword unit model. For each segment, the statistical recombination of the frequency sub-bands is performed according to

$$\log P(X_j|M^j) = f(\{w_k\}, \{\log P(X_{jk}|M_k^j)\}) \tag{2}$$

where X_{jk} is the band-limited sequence of acoustic parameters associated with the k-th frequency band, M_k^j is the model associated with X_{jk}, and w_k's are the recombination parameters. $P(X_{jk}|M_k^j)$ thus represents the likelihood of a partial (frequency limited and time limited) sequence X_{jk} given model M_k^j and these can be computed with standard HMM or hybrid HMM/ANN systems.

In the work reported here, only HMM/ANN systems have been considered and two different recombination functions $f(\cdot)$ have been tested:

$$f(\cdot) = \sum_{k=1}^{K} w_k \log P(X_{jk}|M_k^j) \tag{3}$$

or

$$f(\cdot) = \log \left[MLP_{w_k}(\log P(X_{jk}|M_k^j)) \right] \tag{4}$$

where MLP_{w_k} represents a multilayer perceptron (MLP) parametrized in terms of w_k's and with $\log P(X_{jk}|M_k^j),\ \forall k$, at its input.

In this paper, all the parameters of the sub-band HMMs as well as the w_k's were estimated on the basis of the segmentation obtained from a regular full-band Viterbi alignment.

3. RECOMBINATION STRATEGIES

Three different strategies have been considered for estimating the recombination parameters $w_k{}'s$ [3]:

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

   Normalized phoneme-level recognition rates inside each frequency band are then used as weighting factors in (3). These weighting factors represent the relative amount of information (normalized to sum to 1) present in each frequency band for each speech unit class.

   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 sub-band is estimated on the basis of the sub-band energy histogram. However, unlike the case of spectral subtraction, these histograms are used to compute the relative reliability of each frequency sub-band. The estimated S/N ratios, normalized to sum up to 1, are then used as w_k's in (3).

3. *Multilayer perceptron.*

   Since it is often argued that the recombination mechanism should be nonlinear, we also tested the use of one MLP to recombine the K partial log-likelihoods $\log P(X_{jk}|M_k^j)$ according to (4). In this case, if S represents the number of speech units (used for temporal recombination, i.e., HMM states, phones, syllables or words), the MLP contains $K \times S$ input units and S output units and is trained to estimate posterior probabilities of each speech units given the log-likelihoods of all sub-bands and all speech units.

4. EXPERIMENTS

4.1. 1st Experiment: Potentiality

In a first experiment, we used 3-state HMM/ANN phone models, 18 critical bands for the full-band system, and three sub-bands (spanning [0-1058], [941-2212], and [1994-4000] Hz) for the three sub-band HMM/ANN recognizers. Note that the overlap is only due to

	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: *Error rates on 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 sub-band recombination at state level without any weighting; "Acc-W"= state recombination with weights proportional to phonetic sub-band accuracy; "SNR-W"= state recombination with weights proportional to automatically estimated sub-band SNR. The column "MLP" refers to sub-band recombination at word level using an MLP.*

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.

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 ANNs. 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.

Resulting error rates 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 sub-band, 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 weights derived from S/N estimates. However, we have observed that even without any knowledge about the S/N ratio in sub-bands [using equal weighting ("No-W") or sub-band accuracy weighting ("Acc-W")] the sub-band recognizer still yields much better results than the conventional full-band recognizer.

4.2. 2nd Experiment: Acoustic Features and Number of Bands

In this experiment, we compared the performance of the approach in terms of:

1. *Number of sub-bands.*

   As opposed to the previous case, we used only 15 critical bands. We experimented with three bands (spanning [0-948], [867-1935], and [1790-4000] Hz), four bands (spanning [0-901], [797-1661], [1493-2547] and [2298-4000] Hz) and six bands (spanning [0-495], [438-778], [707-1144], [1051-1631], [1506-2292] and [2121-4000] Hz). See Table 2.

	FB	3 bands	4 bands	6 bands
CBE	3.4%	1.6%	2.0%	2.2%
CMS	1.3%	0.9%	0.5%	—

Table 2: *Error rates on isolated word recognition (13 American English words, telephone speech). Features were either critical band energies (CBE) or lpc-cepstral features (CMS) independently computed for each sub-band and followed by cepstral mean subtraction. "FB" refers to regular full-band recognizer. For the 3, 4 and 6 sub-band-based systems, state log-likelihoods recombination was performed by an MLP.*

2. *Acoustic features.*

   Three sets of features were considered. The first one was composed of 15 critical band energies (CBE) while the second set used lpc-cepstral features (CMS) independently computed for each sub-band and followed by cepstral mean subtraction. See Table 2.

   The third set was dedicated to recognition under broad band noise. Since it was observed in earlier experiments (not reported here) that the multi-band approach alone was less efficient than other noise cancellation techniques such as spectral subtraction [7] or J-RASTA [6] in the case of wideband noise, it was decided to test the multi-band approach on J-RASTA features. We thus used lpc-cepstral features independently computed for each band limited critical band energies previously processed with the J-RASTA. In this study, a full-band recognizer and a 4-band recognizer, both trained on clean speech, were tested on speech with added car noise (10dB SNR). See Table 3.

We used 1-state HMM/ANN phone models. The database consisted of 13 isolated American English digits and control words (telephone speech — 4 $\times$ 50 speakers in a jack-knifed test). Recombination of the state log-likelihoods was performed by an MLP (trained on clean speech). For example, in the case of four bands, the MLP had an input vector of 45 (phones) $\times$ 4 (sub-bands) log-likelihoods and 45 outputs.

Table 2 compares the error rates on clean speech for CBE and CMS, and for different numbers of sub-bands. These results show that: (1) sub-band modelling yields better recognition performance when compared to a standard full-band approach and (2) all pole modelling of cepstral vectors improve the performance of the full-band system (of course, this was already known!) but also the performance of the sub-band approach. Table 2 also suggests an optimum at 4 (or perhaps 5) independent frequency bands. These results are however still too preliminary to draw any definite conclusions regarding the optimal design of the sub-bands (spans and possible overlaps), which certainly needs to be further investigated.

In the case of noisy speech, as reported in Table 3, the sub-band approach, combined with J-RASTA processing, yields better recognition performance compared to the full-band J-RASTA recognizer. In other work, not reported here, we observed that the sub-band approach also yields much better robustness to narrowband noise when compared to standard speech recognition approaches with noise can-

	FB	4 bands
$J-RASTA$	12.1%	9.1%

Table 3: *Error rates on isolated word recognition (13 American English words) of telephone speech + additive car noise (10dB SNR). Training on clean speech only. "J-RASTA" refers to lpc-cepstral features independently computed for each band limited critical band energies previously processed with the J-RASTA noise cancellation technique. Recombination was performed at the state level using an MLP.*

	State	Phone	Syllable
CMS	2.6%	2.6%	2.3%

Table 4: *Error rates on isolated word recognition (108 German words), lpc-cepstral features computed on 15 critical band energies. Recombination of the three sub-bands was performed at the state level, at the phone level and at the syllable level without any weighting ("No-W" in Table 1).*

cellation capabilities.

4.3. Recombination Level

In all the previous experiments, recombination was performed at the HMM state level (except for the MLP column of Table 1, where recombination was done at the word level).

Preliminary experiments have however been carried out to compare other recombination levels. We used the German database described in Section 4.1[2] and 3 sub-bands (spanning [0-948], [867-1935], and [1790-4000] Hz) described by CMS features (obtained from 15 CBE). Recombination at the state, phoneme and syllable levels were compared. For each of the 41 phones, a 3-state HMM/ANN model was used, and the syllable models were defined as straighforward concatenation of phone models. As shown in Table 4, the best results achieved so far have been obtained by relaxing the synchrony of the frequency bands inside syllables. However, error rates for the three merging schemes are not significantly different and we still need to investigate this further.

5. CONCLUSIONS

In this paper, we presented the basis of our sub-band-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:

- Definition of frequency bands: So far, we have used 3, 4 or 6 frequency bands. The best results were obtained with 4 bands. However, the possible overlap of these bands still need to be optimized. The issue of number of sub-band is further discussed in [5].
- 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 sub-band relatively to the others and taking larger time information into account should be investigated.
- Training scheme: Embedded Viterbi training of the band limited recognizers.
- Recombination level: Clearly the experiments reported here were not conclusive with regard to the preferred unit for recombination. We intend to consider both the levels explored here as well as other plausible units. The differences between the efficacy of different levels may become clearer as we explore the use of these techniques on tasks with greater temporal variability (e.g., for natural continuous speech).

[2] Containing 22 mono-syllabic words, 41 bi-syllabic words, 21 tri-syllabic words and 24 words containing more than three syllables.

ACKNOWLEDGMENTS

We are indebted to Hynek Hermansky, Nelson Morgan, Steve Greenberg, and Nikki Mirghafori for many useful discussions, and to Roger Moore for bringing to our attention that HMM decomposition might be applicable to multiband approaches. We also thank the European Community for their support in this work (SPRACH Long Term Research Project 20077).

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., Hermansky, H., and Morgan, N., "Copernicus and the ASR challenge – Waiting for Kepler," *Proc. of the ARPA Speech Recognition Workshop*, Arden House, New York, Feb. 18-21, 1996.
3. Bourlard, H., Dupont, S., Hermansky, H., and Morgan, N., "Towards sub-band-based speech recognition," *Proc. of European Signal Processing Conference*, Trieste, Italy, September 1996.
4. Duchnowski, P., "A new structure for automatic speech recognition," *MIT PhD Thesis*, September 1993.
5. Hermansky, H., Pavel, M., Tibrewala, S., "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.