SUBBAND-BASED SPEECH RECOGNITION

Hervé Bourlard[1,2] *& Stéphane Dupont*[3]
Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez
B-7000 Mons, Belgium
Email: bourlard,dupont@tcts.fpms.ac.be

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 divide up the full frequency band (represented in terms of critical bands) into several subbands, compute phone probabilities for each subband on the basis of subband acoustic features, perform dynamic programming independently for each band, and merge the subband recognizers (recombining the respective, possibly weighted, scores) at some segmental level corresponding to temporal anchor points. The results presented in this paper confirm some preliminary tests reported earlier. On both isolated word and continuous speech tasks, it is indeed shown that even using quite simple recombination strategies, this subband ASR approach can yield at least comparable performance on clean speech while providing better robustness in the case of narrowband noise.

1. INTRODUCTION

In current automatic speech recognition (ASR) systems, the acoustic processing module typically employs feature extraction techniques in which 20 to 30 ms of speech is analyzed once per centisecond, leading to a sequence of acoustic (feature) vectors that each describe local components of the speech signal. Each acoustic vector is typically a smoothed spectrum or cepstrum. Hidden Markov Model (HMM) states, which are typically associated with context independent or context-dependent phones such as triphones, are then characterized by a stationary probability density function over the space of these acoustic vectors. Words and sentences are then assumed to be piecewise stationary and represented in terms of a sequence of HMM states. In state-of-the-art ASR systems, each 10-ms speech segment is often described in terms of several (dependent or independent) parameters such as instantaneous spectral and energy features, complemented by their first and second time derivative. These parameters are then combined in a single acoustic vector, defining a large dimensional space on which the statistical parameters are estimated. To avoid undersampling of the resulting space, it is usually required to assume that the different features are independent (e.g., by assuming diagonal covariance matrices). Another solution, based on the same assumptions, is to consider the different features as independent parameter sequences that are recombined in the probability space. In both cases, it is however assumed that the streams are entirely synchronous. As discussed in [3], another way of processing the information is to consider the features in terms of different streams being treated independently up to some recombination point (e.g., at the syllable level). In this context, the different streams are not restricted to the same frame rate and the underlying HMM models associated with each stream do not have to have the same topology.

This paper mainly focuses on one particular form of this *multistream* approach, referred to as *subband-based ASR* (or "multiband" approach). The basic idea can be summarized as follows:

1. *Divide up the full frequency band into subbands:* Number, definition and possible overlap of these subbands are still open issues.
2. *Derive appropriate feature vectors for each subband:* It seems that subband PLPs are significantly better than straightforward critical band energies.
3. *Train independent recognizers for each subband region:* The work discussed in this paper has been performed in the context of hybrid HMM/ANN systems where artificial neural networks (ANN) are trained with acoustic vectors (with context) at their input to perform phonetic discrimination in each band.
4. *During recognition, combine the different subband (local) probability estimates at some segmental level:* Preliminary comparisons between state (equivalent to combining the probabilities before the decoding process), phone and syllable combination were inconclusive [5]. Consequently, all the experiments reported here have been obtained with state recombination (i.e., before the decoding process).

On top of psychoacoustic studies [4], we see several motivations for the subband approach:

1. Better robustness to noise in the case of different (and not observed in the training data) signal-to-noise ratio per band. For example, the message may be impaired (e.g., by noise, channel characteristics, reverberation...)

[1] Also affiliated with Intl. Computer Science Institute, Berkeley, CA.

[2] Now with IDIAP, Switzerland.

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

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. This was recently confirmed by several experiments [5].

2. Recent theoretical and empirical results in [2] have shown that auto-regressive spectral estimation from subbands is more robust and more efficient than full-band auto-regressive spectral estimation. Our ASR systems could thus benefit from subband all-pole modeling, which was already shown in [5].
3. As already discussed in the introduction, transitions between more stationary segments of speech do not necessarily occur at the same time across the different frequency bands, which makes the underlying HMM piecewise stationary assumption more fragile. The subband-based 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. For example, different time/frequency resolution tradeoffs may be chosen (time resolution and width of analysis window depending on the considered frequency band).
5. Some subbands may be inherently better for certain classes of speech sounds than others.

In the current paper, we extend the work previously reported by the authors [5, 6] and others [8].

2. FORMALISM

We address here the problem of recombining multiple (independent) input streams (frequency subbands) in a HMM-based ASR system. Briefly, this problem can be formulated as follows: assume K input streams X_k to be recognized, and assume that the hypothesized model for an utterance M is composed of J sub-unit models M_j ($j = 1, \ldots, J$) associated with the sub-unit level at which we want to perform the recombination of the input streams (e.g., syllables, themselves built up, as in standard HMMs from sequences of states). To process each stream independently of each other up to the defined sub-unit level, each sub-unit model M_j is composed of parallel models M_j^k (possibly with different topologies) that are forced to recombine their respective segmental scores at some temporal anchor points. The resulting statistical model is illustrated in Fig. 1. In this model we note that:

- The parallel HMMs, associated with each of the input streams, do not necessarily have the same topology.
- The recombination state (illustrated in Fig. 1 by the "$\otimes$"symbol) is not a regular HMM state since it will be responsible for recombining (according to the possible rules discussed below) probabilities (or likelihoods) accumulated over a same temporal segment for all the streams. Since this should be done for all possible segmentation points, a particular form of HMM decomposition [1], referred to as HMM recombination, has to be used [5].

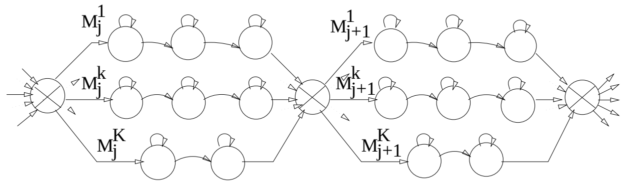

Figure 1. General form of a K-band recognizer with anchor points between speech units (to force synchrony between frequency bands).

As discussed in [3], the statistical recombination of the streams can be formulated in terms of a likelihood-based criterion or a posterior-based criterion. This is still a research issue (and not the topic of this paper). However, most of the results so far have been obtained in a likelihood framework. In the case of a likelihood-based system, we have to find the model M maximizing:

$$p(X|M) = \prod_{j=1}^{J} p(X_j|M_j)$$

Assuming that we have a different "expert" E_k for each input stream X_k (frequency band in the case of subband ASR) and that those experts are mutually exclusive (i.e., conditionally independent) and **collectively exhaustive**, we have:

$$p(X|M) = \prod_{j=1}^{J} \sum_{k=1}^{K} p(X_j^k|M_j^k) P(E_k|M_j) \qquad (1)$$

where X_j^k represents the k-th stream of the sub-sequence X_j, M_j^k represents the sub-unit model for the k-th stream, and $P(E_k|M_j)$ represents the reliability of expert E_k given the considered sub-unit.

Conceptually, the analysis above suggests that, given any hypothesized segmentation, the hypothesis score may be evaluated using multiple experts and some measure of their reliability. Generally, the experts could operate at different time scales, but the formalism requires a resynchronization of the information streams at some recombination point corresponding to the end of some relevant segment (e.g., a syllable).

In the specific case in which the streams are assumed to be statistically independent, we do not need an estimate of the expert reliability, since we can decompose the full likelihood into a product of stream likelihoods for each segment model. For this case we can simply compute:

$$log\ p(X|M) = \sum_{j=1}^{J} \sum_{k=1}^{K} log\ p(X_j^k|M_j^k) \qquad (2)$$

Since we do not have any weighting factors, although the reliability of the different input streams may be different, this approach can be generalized to a weighted log-likelihood approach. We then have:

$$\log p(X|M) = \sum_{j=1}^{J} \sum_{k=1}^{K} w_j^k \log p(X_j^k|M_j^k) \qquad (3)$$

where w_j^k represents to reliability of input stream k. In the multiband case, these weighting factors could be computed, e.g., as a function of the normalized SNR in the time (j) and frequency (k) limited segment X_j^k and/or of the normalized information available in band k for sub-unit model M_j.

More generally, we may also use a nonlinear system to recombine probabilities or log likelihoods so as to relax the assumption of the independence of the streams:

$$\log p(X|M) = \sum_{j=1}^{J} f\left(W, \{\log p(X_j^k|M_j^k),\ \forall k\}\right) \qquad (4)$$

where W is a global set of recombination parameters.

In the particular case of subband ASR [6], three different strategies have been considered for estimating the recombination weights: (1) normalized phoneme-level recognition rates in each frequency band, (2) normalized S/N ratios in each frequency band, and (3) linear or non linear multilayer perceptron.

3. PREVIOUS EXPERIMENTS

Experiments have been reported (and compared with a state-of-the-art full band approach) in [5]. It was shown on a speaker independent task (108 isolated words, telephone speech) that for "clean" (telephone) speech, the subband approach is able to achieve results that are at least as good as (and sometimes better than) the conventional fullband recognizer. When some frequency bands are contaminated by noise, the multiband recognizer yields much more graceful degradation than the broadband recognizer.

In [5], we also reported results on the BELLCORE database consisting of 13 isolated American English digits and control words. We have been comparing the performance of the multiband approach and the fullband approach in terms of acoustic features. Three sets of acoustic parameters were considered. The first one was directly composed of critical band energies (CBE). The second set used lpc-cepstral features independently computed for each subband on the basis of a subset of critical band energies (subband PLP) and possibly followed by cepstral mean subtraction (CMS) or LOG-RASTA processing [9]. One of the main conclusion was that all-pole modeling of cepstral vectors improve the performance of the subband approach.

Further tests were finally performed on the BELLCORE database contaminated by car noise. In this case, we used subband PLP features processed with J-RASTA [9], known to be efficient in broad band noise conditions. We thus used lpc-cepstral features independently computed for each band limited critical band energies previously J-RASTA processed. We obtained significantly better recognition performance using J-RASTA and the multiband approach than with the classical J-RASTA fullband approach.

In most of these experiments, we compared the recognition performance in the case of three bands, four bands and six bands. Results suggests an optimum at 4 (or perhaps 5) independent frequency bands, each band roughly encompassing one formant. These results are however still too preliminary to draw any definite conclusions regarding the optimal design of the subbands (spans and possible overlaps), which certainly needs to be further investigated.

4. NEW RESULTS

The multiband system was further tested on two continuous speech tasks: connected numbers and conversational speech over the phone.

4.1. NUMBERS'93 corpus

NUMBERS'93 is a continuous-speech database collected by the CSLU at the Oregon Graduate Institute. It consists of numbers spoken naturally over telephone lines on the public-switched network [7]. The Numbers'93 database consists of 2,167 spoken numbers strings produced by 1,132 callers. We used 1,534 utterances for training (877 for adjusting the weights of the MLPs and 657 for cross-validation purposes) and 384 utterances for testing. We used single state HMM/ANN context independent phone models. Multilayer perceptrons (MLPs) were used to generate local probabilities for HMMs. The subband-based system had four bands and used subband LOG-RASTA-PLP features. Recombination was done at the state level with a multilayer perceptron with one hidden layer. Results, reported on Figure 2, clearly show that the multiband approach yields much more graceful degradation than the classical approach in the case of band limited noise.

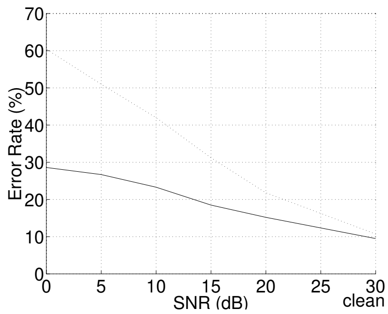

Figure 2. Error rate for speech + band limited noise in the first frequency band (first formant) and various SNR levels. Solid line is for the multiband system. Dotted line is for the classical fullband system.

4.2. SWITCHBOARD corpus

During the SWITCHBOARD workshop held this summer at the Johns Hopkins University (Baltimore) [10], the multiband system was also tested on the SWITCHBOARD conversational telephone speech database. The training data consisted of 4 hours of male speaker utterances. The test set was composed of 240 male speaker utterances. We used 4 frequency bands. The acoustic parameters for each frequency band were subband PLP-CMS. We used single state HMM/ANN context independent phone models. Each of the four subband MLPs had 500 hidden units, while the fullband MLP had 2,000 hidden units. Recombination, according to Equation 4, was done at the state level by an MLP without hidden units. As reported in Table 1, the multiband approach yielded better recognition performance than the fullband approach. Finally, we used the same recombination formalism to merge the probability estimates from the multiband system with those from the fullband system which yielded further improvement.

Error Rate	*FB*	*MB*	*FB* & *MB*
clean speech	63.6%	61.4%	59.7%

Table 1. Word error rates on continuous conversational speech recognition (SWITCHBOARD database). *FB* refers to regular fullband recognizer. *MB* refers to subband-based approach.

5. MULTI-STREAM ASR

As an extension of the subband-based ASR tested in this paper, we see several additional reasons to investigate the proposed formalism as a framework for multistream speech recognition, including:

- A principled way to merge different sources of knowledge such as acoustic and visual inputs.
- The possibility to incorporate multiple time resolutions (as part of a structure with multiple unit lengths, such as phone and syllable). For example, introducing long-term information in current ASR systems could indeed give the possibility of proper syllable modeling in ASR systems basically based on the assumption of stationary HMM states.
- As a particular application of the first two points, this multistream approach could provide us with a principled way to use concurrently different kind of acoustic information, such as instantaneous spectral features and prosodic features, which is known to be a difficult problem.

6. CONCLUSIONS

In this paper, we presented the framework of a new automatic speech recognition architecture: the multiband approach. The general idea is to divide the whole frequency range into several subband, to compute phoneme probabilities for each subband on the basis of subband acoustic features, to perform dynamic programming independently for each band, and finally to force the subband recognizers to recombine their respective score at some segmental level. Although our results are very promising, several open issues remain to be investigated:

- *Definition of frequency bands*: The frequency range as well as the possible overlap of these bands still need to be optimized. The issue of number of subband is further discussed in [8].
- *Recombination criterion*: So far, only a likelihood based recombination has been tested.
- *Weighting scheme*: 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.
- *Training scheme*: Embedded Viterbi training of the band limited recognizers.
- *Recombination level*: Clearly, our experiments (not reported here) were not conclusive with respect to the recombination level. This should be investigated further, especially on tasks with greater temporal variability (e.g., for natural continuous speech).

ACKNOWLEDGMENTS

We are indebted to Hynek Hermansky, Nelson Morgan, Steve Greenberg, Nikki Mirghafori and Sangita Tibrewala for many useful discussions. We also thank the European Community for their support in this work (SPRACH Long Term Research Project 20077).

REFERENCES

[1] A. Varga and R. Moore, "Hidden markov model decomposition of speech and noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 845–848, 1990.

[2] S. Rao and W. A. Pearlman, "Analysis of linear prediction, coding, and spectral estimation from subbands," *IEEE Trans. on Information Theory*, vol. 42, pp. 1160–1178, July 1996.

[3] Bourlard, H., Dupont, S., and Ris, C., "Multistream Speech Recognition," *FPMS-TCTS*, December 1996.

[4] 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.

[5] Bourlard, H. and Dupont, S., "ASR based on independent processing and recombination of partial frequency bands," *Proc. of Intl. Conf. on Spoken Language Processing*, Philadelphia, October 1996.

[6] H. Bourlard and S. Dupont and H. Hermansky and N. Morgan, "Towards sub-band-based speech recognition," *Proc. of European Signal Processing Conference*, Trieste, Italy, pp. 1579-1582, September 1996.

[7] Cole, R.A., Fanty, M., Lander, T., "Telephone Speech Corpus at CSLU," *Proc. of Intl. Spoken Language Processing*, Yokohama, Japan, Spetember 1994.

[8] Hermansky, H., Pavel, M., Tibrewala, S., "Towards ASR On Partially Corrupted Speech" *Proc. of Intl. Conf. on Spoken Language Processing*, Philadelphia, October 1996.

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

[10] "WS96 Workshop Page" `http://www.clsp.jhu.edu/ws96/ws96_workshop.html`, 1996.