Robust Speech Recognition Based on Multi-Stream Features

Stéphane Dupont[†,1], Hervé Bourlard[‡,2] and Christophe Ris[†]

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

[‡]IDIAP
1920 Martigny, Switzerland
Email: bourlard@idiap.ch

Abstract— **In this paper, we discuss a new automatic speech recognition (ASR) approach based on the independent processing and recombination of several feature streams. In this framework, it is assumed that the speech signal is represented in terms of multiple input streams, each input stream representing a different characteristic of the signal. If the streams are entirely synchronous, they may be accommodated simply. However, as discussed in the paper, it may be required to permit some degree of asynchrony between streams, which are then forced to recombine at some temporal "anchor points" associated with some (pre-defined) speech unit levels. We start by introducing the basic framework of a statistical structure that can accommodate multiple observation streams. This approach was initially applied to the case of subband-based speech recognition and was shown to yield significantly better noise robustness. After having summarized these results, the multi-stream approach will be used to combine multiple time-scale features in ASR systems (in our case, to use syllable level features in a phoneme-based HMM system).**

1. Introduction

The general motivation of the multi-stream approach discussed in this paper is to allow for the parallel processing of several feature streams, each feature stream resulting from a particular observation of the speech phenomena. These different information sources, possibly representing different properties of the speech signal are 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 necessarily have the same topology.

This multi-stream approach is a principled way to merging different sources of temporal information (possibly asynchronous and/or with different frame rate) and has many potential advantages. In the case of subband-based recognition, a particular case of multi-stream recognition, it was shown on several databases that this approach was yielding much better noise robustness [1]. As recalled in Section 3, the general idea of this subband approach is then 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.

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

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

Another feature that will be investigated in the current paper is the possibility to incorporate multiple time resolutions as part of a structure with multiple length units, such as phone and syllable. As it will be discussed in details in the full paper (and briefly presented below), it is possible to define subword models composed of several cooperative HMM models focusing on different dynamic properties of the speech signal. This could for example allows for proper syllable modeling in HMM-based ASR systems heavily depending on the assumption of piecewise stationarity (at the level of HMM states).

2. Multi-Stream Statistical Model

We address here the problem of recombining several sources of information represented by different input streams. This problem can be formulated as follows: assume an observation sequence X composed of K input streams X_k representing the utterance to be recognized, and assume that the hypothesized model M for an utterance 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). 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 Figure 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 ($\bigotimes$ in Figure 1) 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. To implement this an approach such as the asynchronous two-level dynamic programming, or a particular form of HMM decomposition [8], referred to as HMM recombination, can also be used [1].

The recognition problem for a likelihood-based system can then be formulated in terms of finding the model M

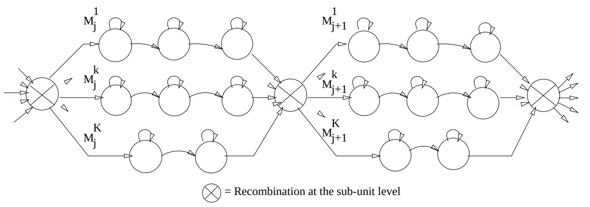

Fig. 1. General form of a K-stream recognizer with anchor points between speech units (to force synchrony between the different streams). Note that the model topology is not necessarily the same for the different sub-systems.

maximizing[3]:

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

where X_j represents the multiple stream subsequence associated with the sub-unit model M_j. Assuming that we have a different "expert" E_k for each input stream X_k (e.g., one "expert" for long-term features and one "expert" for short-term features) and that those experts are mutually exclusive (i.e., conditionally independent) and **collectively exhaustive**, we have:

$$\sum_{k=1}^{K} P(E_k) = 1$$

where $P(E_k)$ represents the probability that expert E_k is better than any other expert. We then 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 $P(E_k|M_j)$ 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:

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

[3]The a posteriori-based formulation (finding the model M maximizing $P(M|X)$) is not discussed here. For further details, see [2]

where w_j^k represents the reliability of input stream k. In the multi-band case (see Section 3), 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.

During recognition, we will have to find the best sentence model M maximizing $p(X|M)$. Different solutions will be investigated, including:

1. Recombination at the sub-unit level (where M_j's are sub-unit models composed of parallel sub-models, one for each input stream, as illustrated on Figure 1).
2. Although it does not allow for asynchrony of the different streams, recombination at the HMM state level (where M_j's are HMM states) is also discussed in this paper.

Recombination at the HMM-state level can be done in many ways, including untrained linear way or trained linear or nonlinear way (e.g., by using a recombining neural network). This is pretty simple to implement and amounts to performing a standard Viterbi decoding in which local (log) probabilities are obtained from a linear or nonlinear combination of the local stream probabilities. Of course, this approach does not allow for asynchrony, yet it has been shown to be very promising for the multi-band approach discussed in Section 3.

On the other hand, recombination of the input streams at the sub-unit level requires a significant adaptation of the recognizer. We are presently using an algorithm referred to as "HMM recombination". It is an adaptation of the HMM decomposition algorithm [8]. The HMM-decomposition algorithm is a time-synchronous Viterbi search that allows the decomposition of a single stream (speech signal) into two independent components (typically speech and noise). In the same spirit, a similar algorithm can be used to combine multiple input streams (e.g., short-term features and long-term features) into a single HMM model. The constraint between the parallel sub-models is implemented by forcing these models to have the same begin and end points. The resulting decoding process can be implemented via a particular form of dynamic programming that guarantees the optimal segmentation.

All the work presented in this paper has been carried on in the framework of hybrid HMM/ANN systems [3]. On top of the advantages already known, this approach is particularly attractive to the multi-stream experiments reported here since (1) it allows to estimate local and global posterior probabilities (directly reflecting confidence levels) and (2) allows to compute these probabilities on the basis on large acoustic contexts (which will be used in Section 4 when using multiple time scales).

3. Subband-based speech recognition

As a particular case of the multi-stream approach, a new speech recognition system based on independent processing and recombination of partial frequency bands was recently developed and tested on several clean and noisy databases. The general idea is to split the whole frequency band (represented in terms of critical bands) into a few subbands, to perform acoustic processing independently for each subband, to use these subband features to compute subband phonetic probabilities and then to recombine these sources of information at a certain speech unit level to yield global scores and a global recognition decision.

There are many potential advantages to using 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 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.

Experiments have been reported (and compared with a state-of-the-art full band HMM/ANN approach) in [1]. The most important results and conclusions are briefly summarized here.

It was first 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. Furthermore, when some frequency bands were contaminated by noise, the multiband recognizer was yielding much more graceful degradation than the broadband recognizer.

Results were also reported in[1] on a telephone database (referred to as "Bellcore Digits") consisting of 13 isolated American English digits and control words. More specifically, the performance of the multiband and the fullband approaches were also compared in terms of acoustic features. Three sets of acoustic parameters were considered: critical band energies, lpc-cepstral features independently computed for each subband on the basis of a subset of critical band energies (subband PLP [6]) possibly followed by cepstral mean subtraction or log-RASTA processing [7]. One of the main conclusion was that all-pole modeling of cepstral vectors greatly improves the performance of the subband approach. Further tests were finally performed on the same Bellcore database on which 10dB car noise was added. In this case, subband J-RASTA PLP features [7] (known to be more robust to additive broad bad noise) were used and it was shown again that the subband approach was outperforming the regular full band system.

More recently, experiments were performed on the Numbers'93 database, a continuous speech telephone database collected by the CSLU at the Oregon Graduate Institute [4]. It consists of numbers spoken naturally over telephone lines on the public-switched network. 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.

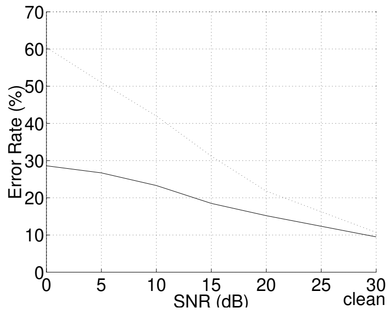

Fig. 2. Error rate for speech + band limited noise in the first frequency band (first formant) and various SNR levels. Solid line is for the multiband system, dotted line is for the fullband system.

4. Combining multiple time-scale features

In the previous section, several results were recalled to show that the multi-stream approach seems to be particularly robust to (unpredictable) band limited and wideband noise conditions. Another potential advantage of this approach which is discussed now concerns the possibility to combine short-term temporal and long-term temporal information. Indeed, current ASR systems mainly use short-term information, typically at the phoneme level, while the longer term information is supposed to be captured via the HMM topology. However, it is often acknowledged that it may be necessary to incorporate larger lexical units than the phoneme to capture all the speech variability and to model long-term dynamics. A plausible candidate is the syllable. Unfortunately, long term temporal dependencies (dynamics), e.g., between syllables, are not explicitly captured by the centisecond-based feature extraction or by the model topology. Consequently, properly handling longer temporal regions (stretching over more than the typical phoneme or HMM-state duration) is still an open issue.

Several studies have attempted to use acoustic context. This was done either by conditioning the posterior probabilities on several acoustic frames, or by using temporal derivative features (see, e.g., [3], [5]). Typically, an optimum was observed with a context covering 90 ms of speech, corresponding approximately to the mean duration of phonetic units. However, these approaches do not allow for representing higher level temporal processes (such as syllable dynamics for instance) since the underlying HMM model is still phoneme based and implicitly assumes piecewise stationarity (at the HMM state level). In fact, what we should actually (attempt to) do is to process short-term and long-term information with two concurrent HMMs assuming (via different topologies and different features) piecewise stationarity at different temporal scales. In the following we thus tested the multi-stream approach to combine short-term dependencies and features associated with them (e.g., at the level of 90 ms) with long-term dependencies and their corresponding features (e.g., at the level of 200 ms).

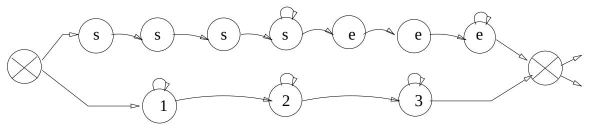

Fig. 3. Syllable [se] multi-stream model.

As a first attempt in this direction, and as illustrated in Figure 3, initial experiments were performed with syllable models described in terms of two parallel models:

1. A "regular" syllable model built up by concatenating context independent (HMM/ANN) phone models and supposed to capture the fine structure of the syllable. This model was processing acoustic vectors as usually used in HMM/ANN systems, typically 9 frames of acoustic context covering about 100 ms. Minimum phone duration was also used.
2. A second HMM model aimed at capturing the gross syllable temporal structure. In our initial experiments, this model was composed of fewer states (3-states in our case) processing larger temporal context of about 200 ms. It is however clear that both the topology and the features will be subject to optimization in the future.

Preliminary tests were performed on the NUMBERS'93 database already used in Section 3, with the same split between training, cross-validation and test data. Besides single state HMM/ANN context independent phone models, another HMM/ANN system was also used for the gross syllable models. Full band log-RASTA-PLP parameters were used, with 9 frames (125 ms) of contextual information for the phoneme-based model, and 17 frames (225 ms) for the gross syllable model. Decoding was done with the HMM decomposition/recombination algorithm. We recombined the sub-stream models either linearly (Eq. 3), or by using a multilayer perceptron (Eq. 4). As an additional reference point, tests were also performed by constraining the search (based on phone HMMs) to match the true (hand labeled) syllable segmentation[4].

Tests were done in the case of clean speech as well as in the case of speech corrupted by additive stationary white noise. Results, reported in Table I and compared to a state-of-the-art phone-based hybrid HMM/ANN system, clearly show a significant performance improvement.

5. Conclusions

In this paper, we discussed a new speech recognition approach based on the independent processing and recombination of multi-stream features. This approach was

[4]This was achieved by using time dependent syllable transition penalties, where the penalties are very high for the time slots where a syllable transition is not allowed.

Error Rate	*Phone*	*Linear*	*MLP*	*Cheat*
clean speech	10.7%	10.1%	8.9%	6.8%
speech+noise	17.2%	16.2%	16.2%	13.5%

TABLE I

WORD ERROR RATES ON CONTINUOUS NUMBERS (NUMBERS'93 DATABASE). *Phone* REFERS TO REGULAR PHONE BASED RECOGNIZER. *Linear* REFERS TO MULTI-STREAM SYSTEM WITH LINEAR RECOMBINATION OF THE TWO STREAMS. *MLP* REFERS TO A RECOMBINATION WITH AN MLP. *Cheat* REFERS TO CONSTRAINING THE DP SEARCH WITH SYLLABLE BOUNDARIES. NOISE WAS ADDITIVE GAUSSIAN WHITE NOISE, 15dB SNR.

tested in the framework of subband based speech recognition as well as on a new model combining multiple time scale features. In both cases, preliminary results suggest that, while opening many new research opportunities, this generic approach (1) does not degrade performance on clean speech, (2) is more robust to unpredictable (non-stationary) noise and (3) could provide a new formalism for combining different sources of short term and long term information. This preliminary work will now be extended in several directions, including:

- *Long-term features*: Further experiments have to be done to determine the features that are best suited to capture long-term dynamic properties.
- *Recombination criterion*: So far, only a likelihood based recombination has been tested.
- Further work with subband ASR.
- Combining subband and multiple temporal scale recognition.

Acknowledgments

We thank the European Community for their support in this work (SPRACH Long Term Research Project 20077). We also thank our colleagues Nelson Morgan, Steve Greenberg, and Nikki Mirghafori from Intl. Computer Science Institute (Berkeley, CA), and Hynek Hermansky and Sangita Tibrewala from Oregon Graduate Institute (Portland, OR) for helpful discussions.

References

[1] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. of Intl. Conf. on Spoken Language Processing*, (Philadelphia), pp. 422–425, Oct. 1996.

[2] H. Bourlard, S. Dupont, and C. Ris, "Multi-stream speech recognition," Tech. Rep. IDIAP-RR 96-07, IDIAP, Martigny, Switzerland, 1996.

[3] H. Bourlard and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic Publishers, ISBN 0-7923-9396-1, 1994.

[4] R. Cole, M. Fanty, and T. Lander, "Telephone speech corpus at cslu," in *Proc. of Intl. Spoken Language Processing*, (Yokohama, Japan), September 1994.

[5] S. Furui, "Speaker independant isolated word recognizer using dynamic features of speech spectrum," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust, Soc. Am.*, 87 (4), pp. 1738-1752, April 1990.

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

[8] 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.