IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Multi-Stream Speech Recognition

Hervé Bourlard [a] Stéphane Dupont [b]
Christophe Ris [b]

IDIAP–RR 96-07

December 1996

Dalle Molle Institute for Perceptive Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

a IDIAP
b Faculté Polytechnique de Mons, Belgium

IDIAP Research Report 96-07

Multi-Stream Speech Recognition

Hervé Bourlard Stéphane Dupont Christophe Ris

December 1996

Abstract. In this paper, we discuss a new automatic speech recognition (ASR) approach based on 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 (as they usually are in state-of-the-art systems). However, as discussed in the paper, it may be required to permit some degree of asynchrony between streams. This paper introduces the basic framework of a statistical structure that can accommodate multiple (asynchronous) observation streams (possibly exhibiting different frame rates). This approach will then be applied to the particular case of multi-band speech recognition and will be shown to yield significantly better noise robustness.

1 Introduction

In current automatic speech recognition (ASR) systems, the acoustic processing module typically employs feature extraction techniques in which 20 to 30 milliseconds 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-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 a consequence:

1. The segmentations of the different feature streams into piecewise stationary segments are constrained to occur at the same instant (i.e., HMM-state transitions occur at the same moment). It is however easy to see that this is a strong constraint that could seriously hurt HMM-based ASR systems since different streams could very well be "stationary" at different moments. This could already be the case with the widely used instantaneous and first time derivatives of spectral (cepstral) features. In the worst case, it could very well be that merging (purely at the centisecond level) two sequences that actually are piecewise stationary results in a sequence that is no longer piecewise stationary.

2. The underlying HMM topology for the different streams is the same, which implies that (1) the number of stationary segments is the same for each stream, and (2) the temporal resolution is the same. Obviously, these could also be strong limitations, especially in the case where the streams are encoding quite different informations like, e.g., spectral characteristics and micro and/or macro prosodic clues.

Finally, the way the recombination of the different streams is achieved during recognition is usually independent of their respective reliability (which could be different than the one observed on the training set, e.g., due to different noise conditions).

The core idea of the multi-stream system discussed in this paper is to divide the information content of the incoming speech signal into several sub-streams, each representing different properties of the speech signal and 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.

There are many potential advantages to this multi-stream approach, including:

1. A principled way to merge different sources of knowledge such as acoustic and visual inputs.

2. 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.

3. As a particular application of the first two points, this multi-stream 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.

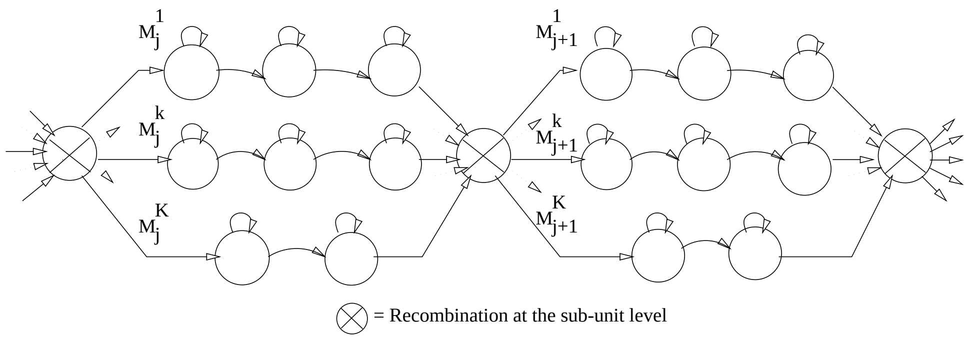

Figure 1: General form of a K-stream recognizer with anchor points between speech units (to force synchrony between the different streams). Note that the model topology is not necessarily the same for the different sub-systems.

4. Independent processing and recombination of partial frequency bands, as a very particular case of multi-stream recognition. As discussed in this paper (Section 5), there are many potential advantages to this sub-band approach, including (1) more robustness to speech impaired by selective narrowband noise, and (2) possibility to apply different time/frequency tradeoffs and different recognition strategies in the sub-bands. This sub-band technique appears to be very promising and will be discussed in more details in this paper.

In the following, we first introduce the statistical framework of the multi-stream approach. We then discuss its application to multi-band ASR and present some of the results achieved so far.

2 Multi-Stream Statistical Model

2.1 Formalism

We address here the problem of recombining multiple (independent) input streams in an 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 ($\bigotimes$ 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. Since this should be done for all possible segmentation points, a particular form of HMM decomposition [21], referred to as HMM recombination, has to be used [4].

We note here that this approach has some common motivations and interesting relationships with the stochastic segment approach discussed in [15].

Most of the work discussed here has been performed in the context of particular HMM systems using an artificial neural network (ANN) to estimate the local probabilities. Such systems are referred to as hybrid HMM/ANN system [7] — see Section 3 for further discussion. In the framework of such hybrid HMM/ANN system, the training and recognition problems can then be phrased in terms of posterior probabilities or else in terms of likelihoods. Depending on the assumptions being made, different solutions are available for recombining the sub-streams probabilities or likelihoods. In the following, we will briefly describe one of those possible solutions for the case of posterior probabilities as well as likelihood, including the approach that has been tested so far in the framework of multi-band ASR.

In this paper, we mainly discuss the **estimation** problem. Based on partial temporal information from several input streams, we will show how to estimate $p(X|M)$ or $P(M|X)$[1]. Although **training** is not discussed here, it is pretty obvious that standard training procedures like Viterbi training or Baum-Welch (or the equivalent standard hybrid HMM/ANN training procedures) still hold. The best path search (in the case of Viterbi) or re-estimation of probability parameters (in the case of Baum-Welch) is to be performed performed by one of the algorithm discussed below.

2.2 Posterior-Based System

In the case of a posterior-based system, the recognition problem can then be formulated in terms of finding the best model M maximizing $P(M|X)$, which given a few relatively standard assumptions (primarily the independence of the language and acoustic models), may be written as:

$$P(M|X) = \prod_{j=1}^{J} P(M_j|M_1, \ldots, M_{j-1}) P(M_j|X_j)$$

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:

$$P(M_j|X_j) = \sum_{k=1}^{K} P(M_j, E_k|X_j) = \sum_{k=1}^{K} P(M_j^k|X_j^k) P(E_k|X_j)$$

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 $P(E_k|X_j)$ the "reliability" of expert E_k given the acoustic X_j assigned to sub-unit M_j. This reliability will have to be estimated during training or automatically estimated during recognition. The global posterior probability is then given by:

$$P(M|X) = \prod_{j=1}^{J} P(M_j|M_1, \ldots, M_{j-1}) \sum_{k=1}^{K} P(M_j^k|X_j^k) P(E_k|X_j) \tag{1}$$

in which the first factor contains the language model information (including grammar, e.g., bi-grams, and transition probabilities between states) while the second factor represents the acoustic information in each segment, integrated over all possible input streams.[2] As shown in previous papers [6], the terms in the second factor can be computed in terms of local posterior probabilities.

[1] Note: In this paper, probabilities are denoted $P(\cdot)$ while likelihoods are denoted $p(\cdot)$.

[2] Note that for simplicity's sake, we have only explicitly shown 2 temporal levels here: for instance, the level of a complete utterance (modeled by M) and of a partial utterance (modeled by M_j). More generally the formalism can easily include multiple levels, such as utterance, word, syllable, phone, state ... in each case the multiple experts can be combined at any desired level.

2.3 Likelihood-Based System

In the case of a likelihood-based system, we have to find the model M maximizing:

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

As for posterior-based systems, and using the same definition of "experts" and the same assumptions, we can show that we have:

$$p(X|M) = \prod_{j=1}^{J} \sum_{k=1}^{K} p(X_j^k|M_j^k) P(E_k|M_j) \tag{2}$$

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:

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

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:

$$\log p(X|M) = \sum_{j=1}^{J} \sum_{k=1}^{K} w_j^k \log p(X_j^k|M_j^k) \tag{4}$$

where w_j^k represents to reliability of input stream k. In the multi-band case (see Section 5), 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:

$$\log p(X|M) = \sum_{j=1}^{J} f\left(W, \{\log p(X_j^k|M_j^k),\ \forall k\}\right) \tag{5}$$

where W is a global set of recombination parameters.

3 Hybrid HMM/ANN Recognition Systems

Most of the work discussed in the rest of this paper has been performed in the context of hybrid HMM/ANN systems. Typically, in such systems, an artificial neural net (ANN) is trained with acoustic vectors at its input to generate HMM state class posterior probabilities that are used as local probabilities for HMMs. This kind of approach has been successfully used with multilayer perceptrons (MLPs) [7] as well as recurrent neural networks (RNNs) [19]. This approach offers several potential advantages over standard HMMs, including:

- Model accuracy: ANN estimation of probabilities does not require detailed assumptions about the form of the statistical distribution to be modeled, resulting in more accurate acoustic models.
- Discrimination: ANNs can easily accommodate discriminant training. Of course, as currently done in standard HMM/ANN hybrid discrimination is only local (at the frame level). However, recent theoretical works show that global discriminant training of hybrid systems can also be performed [6].
- Context sensitivity: In the case of RNNs or if several acoustic vectors are used at the input of an MLP, local correlation of acoustic vectors can be taken into account in the probability distribution.
- Parsimonious use of parameters since all the probability distributions are represented by the same set of shared parameters.
- Flexibility: Using a neural network as the acoustic probability estimator permits the easy combination of diverse features, such as a mixture of continuous and categorical (discrete) measures.
- Complementarity: it is sometimes the case that neural networks can supply complementary information to that provided by an existing likelihood-based system [3].
- More recently, it was also observed that the availability of posterior probabilities (before division by priors) allowed a more efficient pruning for large vocabulary speech recognition systems [17].
- Finally, in the framework of the multi-stream approach, the fact that hybrid HMM/ANN systems are based on posterior probabilities makes it easier to merge multiple recognizers, each of them having different properties. Also, advanced techniques initially developed in the framework of neural networks to recombine statistical experts (mixture of experts [22]) can also be used.

Many (relatively simple) speech recognition systems based on this hybrid HMM/ANN approach, have been proved, on controlled tests, to be both effective in terms of accuracy (comparable or better than equivalent state-of-the-art systems) and efficient in terms of CPU and memory run-time requirements (see, e.g., [14, 18]). More recently, such a system (ABBOT from Cambridge University, see, e.g., [13]) has been evaluated under both the North American ARPA program and the European LRE SQALE project (20,000 word vocabulary, speaker independent continuous speech recognition). In the preliminary results of the SQALE evaluation (reported in [20]) the system was found to perform slightly better than any other leading European system and required an order of magnitude less CPU resources to complete the test. Another striking result is that the acoustic models for this system used several hundred thousand parameters (around 500,000 for ABBOT) while the corresponding models for the competing systems used millions of parameters.

4 Decoder Design

During recognition, we will have to find the best sentence model M maximizing $P(M|X)$, according to (1), or $p(X|M)$, according to (2).

In both cases (posterior-based systems and likelihood-based systems), 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).
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.

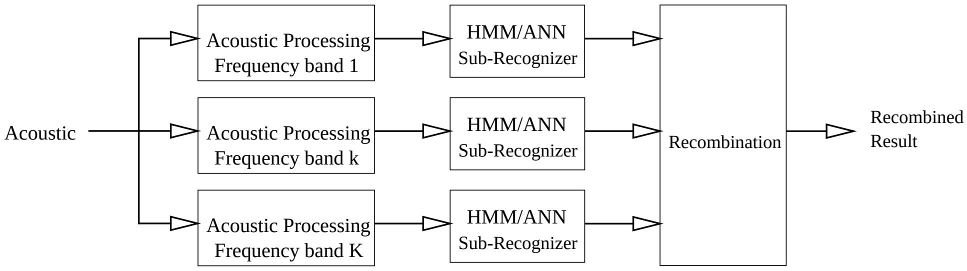

Figure 2: Sub-band-based ASR system architecture.

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 5.

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 [21]. 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. In this framework, each sub-unit HMM model (e.g., syllable model) can then be built up from parallel sub-models with topologies and processing characteristics better adapted to the properties of each of the input streams. The constraint between the parallel sub-models is implemented by forcing these models to have the same begin and end points. The sub-models, stretching out over the same temporal segment and partial likelihoods (over a same temporal segment), are then recombined to yield a global segment score (as illustrated in Fig. 1 by the "$\bigotimes$" state). This decoding process can be implemented via a particular form of dynamic programming that guarantees the optimal segmentation.

Although not discussed in this paper, it is clear that the same kind of algorithm can be used during training of such systems. For example, in the case of Viterbi training, the HMM recombination algorithm will be used to segment the training sentences for each input stream, providing the K ANNs (one for each stream) with new target outputs to re-estimate the state posterior probabilities.

5 Sub-band-Based ASR

5.1 Formalism

As a particular case of the multi-stream approach, we have been investigating an ASR approach based on independent processing and recombination of frequency bands. The general idea, illustrated in Fig. 2 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 sub-unit level to yield global scores and a global recognition decision. Acoustic processing is now performed independently for each sub-band. This lead to K input streams, each being associated with a particular frequency band. In this work, local probabilities for each frequency band are estimated with a ANN trained in a discriminant way on the associated sub-band feature stream. In this case, each HMM in Fig.1 is actually a HMM/ANN system trained independently in a regular way, but looking only at a specific frequency band. As done in usual HMM/ANN systems, each sub-band ANN will also be provided with some contextual information, typically 9 frames of acoustic vectors.

The definitions and equations of Section 2 can now be specialized to this specific multi-band formalism. Each of the K sub-recognizer is now using the information contained in one frequency band, this information being described by an input stream X_k. Equations 1 to 5 are still valid with the same assumptions. In (1), $P(M_j^k|X_j^k)$ represents the a posteriori probability of a sub-unit model M_j^k (k-th frequency band model for sub-unit M_j) given X_j^k, the time limited input stream for the k-th frequency band. $P(E_k|M_j)$ represents the "reliability" of expert E_k, working on the k-th frequency band given the acoustic X_j assigned to sub-unit M_j. In (2), $p(X_j^k|M_j^k)$ is the likelihood of the time limited and frequency limited acoustic X_j^k given the sub-unit model for the processed frequency band.

5.2 Motivations

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 [2]) 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 high-pass or low-pass 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, channel characteristics, reverberation...) 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. This was recently confirmed by several experiments [4]. Surprisingly, even when the recombination is simply performed at the HMM state level it is observed that the multi-band approach is yielding better performance and better noise robustness than a regular full-band system. It could however be argued that, in this latter case, the multi-band approach boils down to a regular full-band recognizer in which (as discussed in the introduction) several likelihoods of (assumed) independent features are estimated and multiplied together to yield local likelihoods[3]. This is however not true when using posterior based systems (such as hybrid HMM/ANN systems) where the sub-bands are presented to different nets that are independently trained in a discriminant way on each individual sub-band.

2. As for a regular full-band system, it was shown in [4] that all-pole modeling was significantly improving the performance of multi-band systems. However, as an additional advantage of the sub-band approach, it can be shown that this all-pole modeling may be more robust if performed on several sub-bands, i.e., in lower dimensional spaces, than on the full-band signal [16]. This could further explain the discussion in the previous point regarding the recombination at the state level.

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 piecewise stationary assumption more fragile. The sub-band approach may have the potential of relaxing the synchrony constraint inherent in current HMM systems.

[3] Indeed, in likelihood based systems, expected values for the full-band is the same than the concatenated expected values of sub-bands and when multiple streams of data are used it is only with the goal of improving the quality of the statistical estimates given a limited amount of training data.

4. Different recognition strategies might ultimately be applied in different sub-bands. 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 sub-bands may be inherently better for certain classes of speech sounds than others.

5.3 Estimation of the weighting factors

The approaches discussed in this paper allow the integration of partial frequency band information. In this case, each of the sub-recognizers has to make decisions based on parameters representing the information contained in a band-limited version of the speech signal, leading to one temporal stream per frequency band. Finally, recombination is done according to (4) and (5). Indeed, experimental results obtained so far showed that recombining sub-stream log-likelihoods (Equations 3, 4 or 5) is yielding slightly better performance than recombining posterior probabilities according to (2). Three different strategies [5] have been considered for estimating the recombination parameters $w_j^{k'}s$ in these equations:

1. *Phoneme-level recognition rates* – Normalized phoneme-level recognition rates inside each frequency band are then used as weighting factors in (4). 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 [12], the S/N ratio in each frequency sub-band was estimated on the basis of the sub-band energy histograms. Typically, these histograms exhibit two peaks at two different energies, the lower energy peak (E_1) corresponding to the silence (+noise) frames while the higher energy peak (E_2) corresponds to the speech (+noise) frames. For each frequency band, the distance between the two peaks is a function of the S/N ratio, which can be estimated by fitting two Gaussians on the histogram[4]. The SNR in each sub-band is then computed as the ratio $\frac{(E_2-E_1)}{E_1}$ in dB. These S/N ratios, normalized to sum up to 1, were used as w_k's in (4).

3. *Multilayer perceptron* – Since the recombination mechanism could be nonlinear, we also tested the use of a multilayer perceptron (MLP) to recombine the K partial log-likelihoods $\log p(X_j^k|M_j^k)$ according to (5). 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.

5.4 Experiments

Experiments were done both on isolated word recognition tasks and on a continuous speech recognition task. This paper only reports the most relevant results.

We have already show that each of the sub-recognizers make decisions based on acoustic parameters representing the information contained in a band-limited version of the speech signal. These acoustic parameters are computed independently for each band, on the basis of a subset of critical band energies. Critical band energies are a particular spectral representation of the signal based on psycho-acoustical knowledge of the spectral resolution and sensitivity of the human ear. In the current work, three sets of acoustic parameters were considered. The first one was directly composed of critical

[4] In the present work, we instead implemented a simple one-dimensional form of online clustering algorithm for a 2-class problem.

band energies (CBE). The second set used lpc-cepstral features independently computed for each sub-band and followed by cepstral mean subtraction (PLP-CMS). The third set was dedicated to recognition under broad band noise conditions. Since it was observed in earlier experiments that the multi-band approach alone was less efficient than other noise cancellation techniques such as spectral subtraction [12] or J-RASTA [10] in the case of wideband noise, it was decided to test the multi-band approach on J-RASTA-PLP features. We thus used lpc-cepstral features independently computed for each band limited critical band energies previously J-RASTA processed.

In a first experiment, we used 3-state HMM/ANN phone models and three sub-bands (spanning [0-1058], [941-2212], and [1994-4000] Hz) for the three sub-band-based HMM/ANN recognizer. Note that the slight 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. The features used for each recognizer were band filtered critical band energies complemented by their first temporal derivatives, and 9 frames of contextual information were used at the input of the ANNs. Recombination of the frequency bands was done at the state level with weights w_k's proportional to automatically estimated sub-band SNR. Table 1 shows that, in the case of clean telephone speech, the multi-band approach yields recognition performance that is as well as good as with the classical full-band approach. In the case of speech corrupted with additive band limited noise, the degradation is much more graceful with the multi-band system.

Error Rate	FB	MB
clean speech	3.6%	3.2%
speech+noise	25.5%	6.3%

Table 1: Error rates on isolated word recognition (108 German words, telephone speech. Noise was additive white noise in the 1st frequency band, 10 dB SNR. Training was done on clean speech. "FB" refers to regular full-band recognizer. "MB" refers to the multi-band recognizer.

In a second experiment, 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). We experimented with four bands (spanning [0-901], [797-1661], [1493-2547] and [2298-4000] Hz). 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. As opposed to the previous case, we also used PLP-CMS acoustic parameters for each frequency band. This all pole modeling of cepstral vectors improved the performance of the sub-band approach. Results are reported in Table 2. Experiments were also done with additive car noise on the test set only. As this is a kind of broad band noise, we used J-RASTA noise cancellation technique [10] which is known to yield improved robustness in this case. Table 2 shows that J-RASTA still holds in the framework of the multi-band approach. We obtained significantly better recognition performance using J-RASTA and the multi-band approach than with the classical J-RASTA full-band approach.

Error Rate	FB	MB
clean speech	1.3%	0.5%
speech+noise	12.1%	9.1%

Table 2: Error rates on isolated word recognition (13 American English words, telephone speech - *Bellcore* digits database). Noise was additive car noise, 10 dB SNR. Training was done on clean speech. "FB" refers to regular full-band recognizer. "MB" refers to the multi-band recognizer.

Previous developments and tests were done on small vocabulary isolated word databases. During the SWITCHBOARD workshop held this summer at the Johns Hopkins University (Baltimore) [1], the multi-band 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, as defined in the previous paragraph. The acoustic parameters for each frequency band were sub-band PLP-CMS. We used 1 state HMM/ANN context independent phone models. Each of the sub-band MLPs had 500 hidden units, while the full-band MLP had 2000 hidden units. Recombination was done at the state level by an MLP. Here again, the multi-band approach yields better recognition performance than the full-band approach.

Error Rate	FB	MB
clean speech	63.6%	61.4%

Table 3: Word error rates on continuous conversational speech recognition (Switchboard database)

Other results as well as a more extensive insight into the experiments can be found in a related paper [4].

6 Long-term and short-term information

As already discussed in the introduction, another potential advantage of the multi-stream approach that we envision to investigate in the near future is the possibility to combine long-term and short-term information in ASR systems. Indeed, current ASR systems only use short-term information, typically at the phoneme level. Long-term information representing temporal regions stretching over more than the typical phoneme duration is more difficult to handle in the current HMM formalism. As an attempt to include such long-term information, the RASTA approach has been quite successful in some cases [10]. This section presents some motivations for combining long-term and short-term information in the framework of the general multi-stream model.

Current ASR systems are based on phonetic units described in terms of stationary HMM states. Correlation inside these states is generally disregarded. Moreover, the current HMM state is usually assumed independent of previous acoustic vectors. It is clear that these models are far from being well suited at handling long-term dependencies. Only some kind of medium-term dependencies are captured by the topologies of the models. Recently, 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. Typically, an optimum was observed with a context covering 90 ms of speech, corresponding approximately to the mean duration of phonetic units. In addition to introducing long-term information, these approaches have another attractive feature: they tend to smooth the assumptions about the independence of acoustic frames and about the stationarity of the signal, stationarity hypothesis being extended to temporal derivative features. These approaches seem to relieve a number of problems coming from the underlying assumptions of the HMM theory. In experimental systems, they were shown to significantly improve recognition performance (see, e.g., [7, 9]).

However, although state-of-the-art systems based on these approaches work well on carefully dictated clean speech, their performance is severely compromised on conversational speech and on noisy speech. The reason could be that current feature extraction and acoustic modeling schemes do not allow make use of information from time regions covering 200 ms or more. Such long time regions could be interesting in recognizing speech corrupted by stationary noise. Indeed, it has been observed on modulation spectra (spectra of the temporal envelope of the signal) that the modulation energy of speech signals is generally maximum around 5 Hz. This corresponds to a period of 200 ms. As 200 ms is also the maximum of syllable duration distribution, we believe that syllable could be an interesting candidate to include long-term information in current recognizers.

7 Conclusions

In this paper, we presented the framework of a new automatic speech recognition architecture allowing for the processing and merging of several input streams. This general formalism has been tested in the framework of a multi-band speech recognition system. To combine several streams, we adapted the HMM decomposition algorithm which was initially used in source decomposition. Experiments were reported on isolated word recognition tasks and on a continuous speech recognition task. Although the results are very promising, several open issues remain to be investigated carefully:

- *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 [11].
- *Recombination criterion*: So far, only a likelihood based recombination has been tested.
- *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 respect to the recombination level. This should be investigated further, especially on tasks with greater temporal variability (e.g., for natural continuous speech).

Finally, we note here that the issues addresses in the multi-band approach have interesting relationships with other research topics currently investigated by other teams, such as the "missing data" paradigm [8].

Furthermore, we believe that this new approach could be a convenient way of combining short-term and long-term speech features.

Acknowledgments

We are indebted to Hynek Hermansky and Sangita Tibrewala from Oregon Graduate Institute (OGI, Portland, OR) and Nelson Morgan, Steve Greenberg, and Nikki Mirghafori from Intl. Computer Science Institute (ICSI, Berkeley, CA) for many useful discussions in many collaborative efforts. We also thank the European Community for their support in this work (SPRACH Long Term Research Project 20077).

References

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

[2] J. Allen, "How do humans process and recognize speech?," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, 1994.

[3] S. Austin, G. Zavaliagkos, J. Makhoul, and J. Schwartz, "Improving state-of-the-art continuous speech recognition systems using the n-best paradigm with neural networks," in *Proc. DARPA Speech and Natural Language Workshop, Harriman, New York* (C. Morgan Kaufmann, Los Altos, ed.), pp. 180–184, 1992.

[4] 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.

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

[6] H. Bourlard, Y. Konig, and N. Morgan, "REMAP: Recursive estimation and maximization of a posteriori probabilities in connectionist speech recognition," in *Proc. EUROSPEECH'95*, (Madrid, Spain), pp. 1663–1666, Sept. 1995.

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

[8] M. Cooke, A. Morris, and P. Green, "Recognising occluded speech," in *Proc. ESCA Workshop on Auditory Basis of Speech Perception*, (U.K.), July 1996.

[9] 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.

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

[11] H. Hermansky, M. Pavel, and S. Tibrewala, "Towards asr using partially corrupted speech," in *Proc. of Intl. Conf. on Spoken Language Processing*, (Philadelphia), pp. 458–461, Oct. 1996.

[12] H. G. Hirsch, "Estimation of noise spectrum and its application to snr-estimation and speech enhancement," Tech. Rep. TR-93-012, Intl. Comp. Science Institute, Berkeley, CA, 1993.

[13] M. Hochberg, S. Renals, A. Robinson, and G. Cook, "Recent improvements to the ABBOT large vocabulary CSR system," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, (Detroit, MI), pp. 69–72, 1995.

[14] D. Lubensky, A. Asadi, and J. Naik, "Connected digit recognition using connectionist probability estimators and mixture-gaussian densities," in *Proc. of the Intl. Conf. on Spoken Language Processing*, (Yokohama, Japan), 1994.

[15] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE ASSP Trans.*, vol. 37, pp. 1857–1869, 1989.

[16] 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.

[17] S. Renals and M. Hochberg, "Efficient search using posterior phone probability estimates," in *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., Detroit. MI*, pp. 596–599, 1995.

[18] T. Robinson, L. Almeida, J. Boite, H. Bourlard, F. Fallside, M. Hochberg, D. Kershaw, P. Kohn, Y. Konig, J. N. N. Morgan, S. Renals, M. Saerens, and C. Wooters, "A neural network based, speaker independant, large vocabulary, continuous speech recognition system: the wernicke project," in *Proc. EUROSPEECH'93*, (Berlin, Germany), pp. 1941–1944, 1993.

[19] T. Robinson and F. Fallside, "A recurrent error propagation network speech recognition system," *Computer Speech and Language*, vol. 5, pp. 259–274, 1991.

[20] J. Steeneken and D. V. Leeuwen, "Multilingual assessment of speaker independant large vocabulary speech-recognition systems: the sqale project (speech recognition quality assessment for language engineering)," in *Proc. EUROSPEECH'95*, (Madrid, Spain), 1995.

[21] 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.

[22] S. Waterhouse and A. Robinson, "Classification using hierarchical mixtures of experts," in *Proc. 1994 IEEE Workshop on Neural Networks for Signal Processing*, pp. IV–177–186, 1994.