USING MULTIPLE TIME SCALES IN A MULTI-STREAM SPEECH RECOGNITION SYSTEM

Stéphane Dupont[†,1], *Hervé Bourlard*[‡,2]

[†] TCTS, FPMs, Mons, Belgium
[‡] IDIAP, Martigny, Switzerland
Email: dupont@tcts.fpms.ac.be
bourlard@idiap.ch

ABSTRACT

In this paper, we propose and investigate a new approach towards using multiple time scale information in automatic speech recognition (ASR) systems. In this framework, we are using a particular HMM formalism able to process different input streams and to recombine them at some temporal anchor points. While the phonological level of recombination has to be defined a priori, the optimal temporal anchor points are obtained automatically during recognition. In the current approach, those parallel cooperative HMMs will focus on different dynamic properties of the speech signal, defined on different time scales. The speech signal is then defined in terms of several information streams, each stream resulting from a particular way of analyzing the speech signal. More specifically, in the current work, models aimed at capturing the syllable level temporal structure are used in parallel with classical phoneme-based models. Tests on different continuous speech databases show significant performance improvements, motivating further research to efficiently use large time span information of the order of 200 ms into our standard 10 ms, phone-based ASR systems.

1. INTRODUCTION

The multi-stream approach discussed in this paper is a principled way for merging different sources of temporal information (possibly asynchronous and/or with different frame rate) and has many potential advantages. In this approach, it is assumed that the speech signal is described in terms of multiple input streams, each stream representing a different characteristic of the input signal. If the streams are supposed to be entirely synchronous, they may be accommodated simply. However, it is often the case that they are not and sometimes that they do not even even the same frame rate. The multi-stream approach discussed in [1] allows to deal with this. In this framework, the input streams are processed independently of each other up to certain anchor points where they have to synchronize and recombine and their partial segment-based likelihoods. While the phonological level of recombination has to be defined a priori, the optimal temporal anchor points are obtained automatically during recognition.

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 significantly better noise robustness [2, 7]. 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. This subband approach has many other potential advantages, including the possibility to better accommodate the possible asynchrony between different components of the speech spectrum [8].

Another feature that is 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. In the same framework, it is indeed possible to define subword models composed of several cooperative HMM models focusing on different dynamic properties of the speech signal. The results presented here are still very preliminary and will require much further development work. They however show quite clearly that the proposed multi-stream approach could indeed provide a promising way of modeling syllable length information spanning around 200 ms (as well as as microprosodic information) in our standard 10 ms, phone-based ASR systems.

2. MULTI-STREAM STATISTICAL MODEL

In the following, we briefly address the problem of recombining several information sources represented by different input streams (see [1] for a detailed discussion of the mathematical formalism). In this case, an observation sequence X (representing the utterance to be recognized) is assumed to be composed of K input streams X_k (possibly of different lengths and/or different frame rates). A hypothesized model M associated with X will then be built up by concatenating 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 allow the processing of each of the input streams independently of each other up to the pre-defined sub-unit boundaries (determined automatically during decoding), 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 model is il-

[1] Supported by a F.R.I.A. grant (Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture), Belgium.
[2] Also affiliated with Intl. Computer Science Institute, Berkeley, CA.

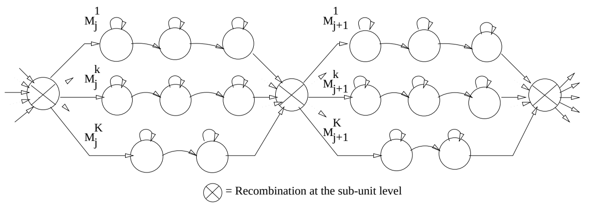

Figure 1: General form of a K-stream recognizer with anchor points between speech units (to force synchrony between the different streams). Note that the model topology is not necessarily the same for the different subsystems.

lustrated 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 ($\otimes$ 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. This should of course be done for all possible segmentation points. The problem appears to be similar to the continuous speech recognition problem where all of the concurrent word segmentations, as well as all of the phone segmentations, must be hypothesized. However, as recombination concerns sub-unit paths that must begin at the same time, and as the best state path in not the same for all of the sub-stream models, it is necessary to keep track of the dynamic programming paths for all of the sub-unit starting points. Hence, an approach such as the asynchronous two-level dynamic programming, or a synchronous formulation of it, is required. Alternatively, a particular form of HMM decomposition [9], referred to as HMM recombination, can also be used [2].

As discussed in [1], the training and recognition problems (including automatic segmentation and recombination) can be coined into different statistical formalisms based on likelihoods or posterior probabilities and using linear or nonlinear (neural network based) recombination schemes. During recognition, we will have to find the best sentence model M maximizing $p(X|M)$ (likelihood formalism). Different solutions can 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 or different topologies of the different streams, recombination at the HMM state level (where M_j's are HMM states) can also be done.

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.

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", which is an adaptation of the HMM decomposition algorithm [9, 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 (Artificial Neural Network) 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 of large acoustic contexts which will be used to catch long-term information.

In [2, 7], this multi-stream approach has been shown to be particularly robust to (unpredictable and not seen on the training data) band limited and wideband noise conditions. It was also shown in [8] to be able to better accommodate the possible asynchrony between the different frequency bands.

3. SYLLABLE-LEVEL DYNAMICS

Another potential advantage of this approach which is studied now is the possibility to combine long-term dynamic properties (using long-term features) and short-term dynamic properties (using short-term features) 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 capture and model in standard ASR systems typically looking at 10 ms frames and assuming piecewise stationarity at the level of HMM states.

Although state-of-the-art systems based on phonemes work well on carefully dictated clean speech, their performance is severely compromised on natural conversational speech and on noisy speech. The reason could be that current feature extraction and acoustic modeling schemes do not allow to make use of information from time regions covering 200 ms or more. Such long time regions could be interesting in distinguishing variable speech from 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, corresponding to a period of 200 ms. As 200 ms is also the maximum of syllable duration distribution, using long-term time regions could also allow to better catch syllable level dynamics.

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. 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. Of course, for the same amount of training data, building syllable models by simply concatenating HMM states emitting 10 ms frames will not help and will not capture long term information...

3.1. Generic syllable model

In the framework of the multi-stream approach, long-term information streams can be used to introduce long-term dependencies into current phone-based systems (or, more precisely, into HMMs based on the concatenation of states emitting 10 ms feature vectors). In this case, in parallel with the standard 10 ms based acoustic stream, we also consider another acoustic stream looking at larger temporal windows. The acoustic streams are then processed by different sub-unit HMM models better suited to the temporal properties they are supposed to capture.

In the current work, experiments were performed in view of modeling syllabic sub-unit models, which means that the different streams were forced to recombine at predefined syllable levels. Consequently, lexicon words were first transcribed in term of syllables defined as phoneme sequences containing a vowel nucleus and optional left and right consonants. Simple rules were used to obtain the lexical syllable boundaries. As illustrated in Fig 2, syllable models were then composed of two parallel models:

1. A "classical" syllable model built up from context independent HMM/ANN phone states, i.e., not really modeling the syllable structure, and using the probabilities obtained at the output of the phone ANN typically looking at 9 frames of 10 ms. As illustrated in Fig. 2, minimum phoneme duration, defined as half the mean duration of the phoneme, was also used.
2. A particular HMM model aimed at capturing the syllable level temporal structure of the speech signal. In these preliminary experiments, a simple 3-state HMM was used and was supposed to be common to all syllable models. This model was using the output probabilities of a 3-output-ANN looking at temporal segments spanning more than 200 ms.

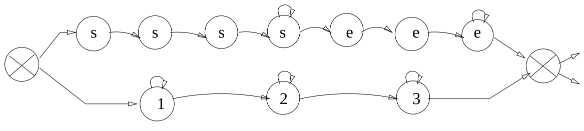

Figure 2: Syllable [se] multi-stream model.

Two continuous speech databases were used in our experiments:

- NUMBERS'93: it consists of numbers spoken naturally over telephone lines on the public-switched network [4]. We used 1,534 utterances for training and 384 utterances for testing. The training procedure of the ANNs uses a cross-validation scheme to prevent the neural network from becoming overspecialized on the training data. From the whole training set, 1400 sentences were used for adjusting the weights of the ANNs and 134 for cross-validation purposes. No grammar was used during testing
- DARPA RESOURCE MANAGEMENT: from the 3990 official sentences, 3591 were used for training and 999 for cross-validation during the training of the ANNs. The 300 february 89 sentences were used for testing. We have been using the standard wordpair grammar which has a perplexity of 60.

We note here that these two databases are quite different with respect to the speaking style. While RESOURCE MANAGMENT is read speech, NUMBERS'93 is more like spontaneous speech.

We used single state HMM/ANN context independent phone models. Feed-forward multilayer perceptrons (MLPs) were used to generate local probabilities for the different HMMs. We used LOG-RASTA-PLP parameters for the NUMBERS'93 experiments and PLPs for the RESOURCE MANAGEMENT experiments. The phoneme-based system was given 9 frames (125 ms) of contextual information and the gross syllable model was given 17 frames (225 ms). Decoding was done with the HMM decomposition/recombination algorithm. We recombined the sub-stream models log-likelihoods either linearly or with an artificial neural network. The recombination weights were optimized on the training set only. As an additional reference point, tests were also performed by constraining the search (based on phone HMMs) to match the true syllable segmentation[3] (obtained from a Viterbi alignment).

Results, reported in Table 1 and compared to a state-of-the-art phoneme-based hybrid HMM/ANN system, clearly show a significant performance improvement.

	Phone	*Linear*	*MLP*	*Cheat*
Error Rate	10.7%	10.1%	8.9%	6.8%

Table 1: 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, 15 dB SNR.

	Phone	*Linear*	*MLP*	*Cheat*
Error Rate	5.9%	5.5%	4.8%	4.2%

Table 2: Word error rates on continuous speech (Resource Management database - February 89 test set).

[3] 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.

3.2. Modeling microprosodic phenomena

As a first extension to these preliminary experiments, the HMM responsible for capturing larger time scale properties was extended to model stressed and unstressed syllables differently. Of course, the number of such models could of be increased (as in [6]) so as to cover a wider range of syllabic structures like CV, stressed CV, CVC, etc. Two "syllable" HMMs were used, one for stressed and one for unstressed syllables. These were 3-state models similar to the one introduced in the previous section for gross syllable modeling. A syllable was qualified as stressed if its vowel nucleus was lexically stressed. The approach proposed in [6] was based on the rescoring (using the syllable-based models) of the N-best hypothesis given by the phone-based recognizer. HMM scores from both types of models were combined at the end of the utterance using weights optimized on the training set. In our case, the merging of phone-based decisions and syllable-based decisions is performed so as to recombine the individual scores but also to force the models to have the same temporal begin and endpoints.

We have used the NUMBERS'93 database and the same parameterization scheme as in the previous section. Although there might be a strong mismatch between our lexically defined syllables and their acoustic realization, we can observe in Table 3 that the proposed method yields performance improvement.

	Phone	*Linear*
Error Rate	10.7%	9.6%

Table 3: Word error rates on continuous numbers (Numbers'93 database). Stressed and unstressed syllable models. To be compared with Table 1.

4. CONCLUSIONS

In this paper, we have discussed a speech recognition using multiple time scales in the framework of a multi-stream approach based on the independent processing and recombination of several feature streams. This approach was used as an attempt to define a particular HMM model able to focus on different dynamic properties of the speech signal and to model piecewise stationarity at different feature levels. Preliminary results suggest that this generic approach can indeed provide a new formalism for combining different sources of short term and long term information. In this work, gross syllable models were efficiently used to catch the syllable level dynamics somewhat constraining a phoneme-based system to align on syllables, but also allowing to make use of microprosodic features like syllabic stress. The method was shown to yield significant performance improvement.

As an alternative approach to introducing syllabic constraints, it was recently shown [10] that acoustically derived syllable onsets can improve speech recognition performance. This further motivates research towards efficient use of long time regions (covering 200 ms or more) and syllable level information.

This preliminary work will now be extended in several directions. Experiments have to be done to determine the features that are best suited to capture relevant long-term dynamic properties and to distinguish different kinds of syllabic structures (e.g. stressed and unstressed syllables). In this respect, the use of modulation spectrum features [5] could be an interesting candidate. Furthermore, as already mentioned, there might be a mismatch between lexically defined syllables and their acoustic realizations, particularly in conversational or free speech. This also highlights the need of further research towards finding appropriate lexical representations for long-term properties such as syllable level dynamics and microprosody.

Acknowledgments

We thank the European Community for their support in this work (SPRACH Long Term Research Project 20077).

5. REFERENCES

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

[2] 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.

[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. Greenberg and B. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. of ICASSP'97*, (Munich), pp. 1647–1650, 1997.

[6] M. Jones and P. Woodland, "Modelling syllable characteristics to improve a large vocabulary continuous speech recogniser," in *Proc. of the Intl. Conf. on Spoken Language Processing.*, 1994.

[7] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," in *Proc. of ICASSP'97*, (Munich), pp. 1255–1258, 1997.

[8] M. Tomlinson, M. Russell, R. Moore, A. Buckland, and M. Fawley, "Modelling asynchrony in speech using elementary single-signal decomposition," in *Proc. of ICASSP'97*, (Munich), pp. 1247–1250, 1997.

[9] A. Varga and R. Moore, "Hidden markov model decomposition of speech and noise," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pp. 845–848, 1990.

[10] S. Wu, M. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, (Munich), pp. 987–990, Apr. 1997.