ASYNCHRONY IN MULTI-BAND SPEECH RECOGNITION

Christophe Cerisara, Dominique Fohr and Jean-Paul Haton
LORIA- Université Henri-Poincaré, Nancy 1
BP 239
54506 Vandoeuvre-les-Nancy, France
cerisara, fohr, jph@loria.fr

ABSTRACT

In this paper, an algorithm for continuous speech recognition systems based on the Multi-Band principle is proposed. This algorithm allows the bands to be asynchronous and has a practical complexity that is very close to the complexity of the classical Viterbi algorithm. The question of whether the bands should be constrained to be synchronous or not is discussed. We show that it is advantageous to let the bands asynchronous, as the increase of complexity, compared to the Viterbi algorithm, is low with our algorithm. Moreover, the accuracy must be at least as good as when the bands are synchronous, and, more importantly, different models than phones, can be used in the bands.

1. INTRODUCTION

Asynchrony between the bands is one of the hypothesis which is at the basis of the Multi-Band model. However, some issues concerning its application are still open. That is the reason why most of present Multi-Band systems use synchrony constraint between the bands at the state level [5]. On the one hand, synchrony allows the complexity of a Multi-Band system to be reduced: this reduction of complexity comes from the fact that the alignment between the states of the Hidden Markov Models (HMMs) and the frames of the speech signal is computed one single time for the whole sentence. Moreover, this alignment can be efficiently computed using the Viterbi algorithm. On the other hand, synchronism is a particular case of asynchrony, and results obtained with an asynchronous system must be at least as good as when synchrony constraints are applied. Another advantage of using an asynchronous system is the possibility to use in each band acoustic models that are not phone models. This advantage is likely to be more important than the potential increase of phonetic accuracy. We have proposed [2] an algorithm to build such models which consider the actual acoustic information which is present in each band, and we think that this approach may lead to a more interesting development of the Multi-Band principle than the synchronous one.

We describe in this paper an adaptation of classical dynamic programming algorithms in order to allow asynchrony between the bands in the phone models. Our goal is to show that asynchrony can be used with only a minor increase of the theoretical and practical complexity. In section 2, we present the problem of designing an asynchronous continuous Multi-Band speech recognition system and we discuss some solutions proposed in the literature. In section 3, we present the asynchronous recognition algorithm used in our Multi-Band system, and we present some results that have been obtained with the TIMIT database.

2. ASYNCHRONY IN MULTI-BAND SYSTEMS

21. Position of the Problem

If asynchrony is allowed between the bands, it is no longer possible to use the Viterbi algorithm to compute the best possible alignment of HMM states. This is due to the fact that the Multi-Band system with asynchrony is a segment-based and not a frame-based recognition system. That means that one frame and one state can no longer be aligned. The alignment can only be carried out between a whole segment of the speech signal and a phone model.

The two solutions that have been proposed in the Multi-Band literature consist either to synchronize the band frame by frame, or to use another algorithm than the Viterbi algorithm to align the speech signal and the phone models. The former solution is the simplest, as a score can be computed after each frame of the speech signal, and the Viterbi algorithm can then be used. However, as discussed above, this solution presents some problems, like the fact that no other unit than phones can be modeled in the bands. Two algorithms have thus been proposed: an adaptation of the HMM-decomposition algorithm [3] and the two-level dynamic programming algorithm [8]. A complete description of the former algorithm is made in [7]. In the next two sections, we present the two-level DP algorithm, as well as another programming algorithm that can also be used for the same task, the Level Building DP algorithm [6]. In section 3, we propose a new algorithm that consists of an adaptation of these two DP algorithms, together with some results that have been obtained with this new algorithm.

22. The Two-level DP Algorithm

The two-level DP algorithm [8] is originally an enhancement of the rough search of the best path through all possible alignments. Its main advantage lies in the fact that the computation of the whole search is divided into two stages: the first one consists of computing the best alignment between each model and each allowed segment in the sentence, whereas the second one consists of searching for the best overall alignment possible. In this algorithm, the best alignment between each segment and

each model considered in the current path is not computed any more as it is done in the basic search algorithm, since it has already been computed during the first stage: the computational cost is thus considerably reduced.

23. The Level Building Algorithm

231. Definition

The level building DP algorithm [6] is characterized by the fact that the best path[1] is computed level by level, a level corresponding to the number of phonemes in a path. Thus, the algorithm begins by computing the scores of all the paths of one phoneme-length. The best paths for all possible ending frames are saved. Then, the alignments between each model and each segment beginning at the end of all the best paths of the preceding level are considered. These one-phoneme length alignments are added to the best paths of the previous levels, creating new paths of two-phoneme length. The best alignments for each ending frame are saved and the process is iterated for all the following levels.

232. Comparison with the Two-level DP

The cost of the level building DP algorithm is lower than the cost of the two-level DP algorithm, as alignments are not compared over the whole sentence for the level building DP, but only for a single model at a time. Actually, the alignments are compared only for the current level, i.e. for one phoneme, whereas in the two-level DP they are compared for the entire sentence.

On the other hand, the two-level DP algorithm is synchronous whereas the level building DP algorithm is asynchronous. This is due to the fact that the level building DP often backtracks in the sequence of frames during the decoding process, as the ending frames of one level can be posterior to the beginning frames of the following level.

3. A CONTINUOUS ALGORITHM FOR MULTI-BAND RECOGNITION

31. Basic Principle

The algorithm we have implemented in our Multi-Band system is intermediate between the two-level DP and the level building DP. It uses the same recurrence principle classically used in dynamic programming algorithms. Its recurrence step is the following:

-Assuming that the best path ending at frame t is known, all the segments $[t, t + d]$, with d varying between 0 and the maximum length D of a phoneme, are aligned with all the phone models. The scores of these new paths are computed and saved if they are greater than the scores of the previously computed paths ending at the same frame. Then, the next frame is considered, and so on until the end of the sentence is reached.

Our algorithm is synchronous, as the two-level DP. However, similarly to the level building DP, it considers the alignments only for one model and adds these alignments to the best paths previously computed. The complexity of our algorithm is also lower than the complexity of the two-level DP or of the level building DP. This point is addressed in section 3.2.

32. Adaptation to Continuous Multi-Band ASR

Several adaptations of the above-described general principle have been made before incorporating it into our Multi-Band system. They are presented below.

321. Theoretical Adaptations

Originally, the two-level DP and the level building DP have been designed in order to use either a finite-state network grammar or no grammar at all [7]. We wanted to use a statistical grammar (actually a phone bigram) in our system. We have thus demonstrated that, in that case, the optimal alignment can still be obtained [1]. However, it is necessary to save more than a single best path ending at each frame. Actually, the best path ending with each possible model must be saved. Thus, *Nmod* paths must be kept for each frame, where *Nmod* is the total number of phone models use.

This constraint considerably increases the complexity of the algorithm. We have thus used a beam-search procedure in order to reduce the complexity. This method has been implemented by saving only *Nbest* paths for each frame, with $Nbest < Nmod$. Some preliminary experiments have shown that results are generally good enough when $Nbest = 1$. Nevertheless, the algorithm does not any more theoretically guarantee to return the optimal path.

322. Practical Adaptations

Each time a model is aligned with a segment of the speech signal, a maximum length for this segment is used. We have compared a phone-dependent maximum length and a constant one. No noticeable differences in the accuracy as well as in the complexity was found. Using such a maximum length makes the algorithm sub-optimal, but it is common to use this heuristic in ASR, e.g., in the level building DP algorithm. Moreover, to deal with long silences in a sentence, we have used a post-processing module that merges consecutive silences in the final path.

The practical complexity of the system has been divided by a large factor just by saving in a table associated with a sliding window the Gaussian probabilities computed in each state of the HMMs. With this heuristics, the system becomes nearly as fast as the reference Viterbi algorithm.

4. EXPERIMENTS

41. Experimental Setup

All the tests presented here have been carried out on the TIMIT database. The tests correspond to the continuous mode, and the accuracy is computed on the sequence of phones. A total number of 48 context-independent phones are modeled, corresponding to the manual segmentation of the corpus, and the accuracy is computed with only 39 phones, as it is usually done on this

[1] A path is an alignment between a sequence of models and the sequence of frames in the signal.

database. Our Multi-Band system is made up of five bands: four sub-bands with frequency limits [0...538 Hz], [461...1000 Hz], [923...2823 Hz] and [2374...7983 Hz], and the full-band. Each of the 48 phones are modeled with the use of a second-order HMM [4]. The linear recombination of the likelihoods returned by the HMMs is done according to the following equation:

$$S(M) = \sum_{b=1}^{B} \alpha_{b,M} \; P(X / M, b)$$

where *S(M)* is the score associated with the model *M*, $\alpha_{b,M}$ are the coefficients of the recombination, and *P(X/M,b)* is the likelihood returned by the HMM modeling *M* in band *b*.

The coefficients $\alpha_{FB,M}$ associated with the full-band for all the models *M* have been set to the same value, ranging from 0 to 1. The coefficients of the other sub-bands are then computed with the equation:

$$\alpha_{b,M} = \frac{1-\alpha_{b,M}}{B-1}, \forall b, M$$

where *B* is the number of bands, including the full-band.

42. Experimental Results

421. Phone Recognition in Clean Speech

Results on the core test part of TIMIT are presented in figure 1, for clean speech.

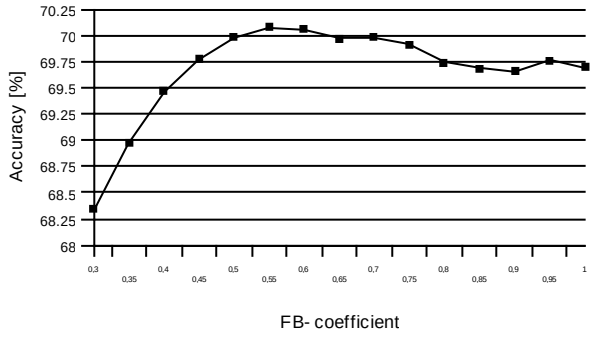

Figure 1. Phone accuracy in clean speech on TIMIT.

Several remarks can be made on this figure:

- In clean speech, it is necessary to use the full-band and the sub-bands together [1], since, when the full-band coefficient FB-coefficient) is close to zero, accuracy is low.
- The optimal accuracy (i.e., 70.1% recognition rate) is obtained for a full-band coefficient close to 0.55. That means that half of the final recognition decision is taken by the full-band, and the sub-bands take the other half. This fact shows that the sub-bands are not only a "part" of the full-band, but also contain different information derived from the acoustic models. These models are different in the sub-bands, since they are built using a different amount of information than in the full-band.
- This figure is quite similar to what has been obtained for isolated phone recognition [1]. This demonstrates experimentally that our continuous algorithm achieves its goal, which is to make Multi-Band ASR possible for continuous speech when asynchrony between the bands is allowed.

422. Is the Optimum Synchronous?

With the previous experiments, we cannot answer the question of whether using asynchrony between the bands gives a better accuracy than when the bands are synchronous. Actually, we should have tested our system when the bands are constrained to be synchronous, and compared these results with those presented here. We have not had time to do these experiments. However, in order to know if the solution given by our continuous algorithm is close to synchrony or not, we have observed the alignment proposed by our algorithm and compared the alignments between the bands.

The comparison has been done for each pair of bands: we have first discarded from the test all the first and the last frames of a phone model, since these frames are necessarily aligned with the same state in all the bands (i.e. with the first and the last state of the chosen model). Then, for the two bands considered, the states with which the remaining frames are aligned are compared. For each frame, if the two states are different, then these frames are considered asynchronous, else, they are synchronous.

The proportion of asynchronous frames for each pair of bands for one sentence of the test corpus is reported in Table 1.

Bands	1	2	3	4	5
1		28 %	30 %	44 %	33 %
2			32 %	37 %	27 %
3				36 %	29 %
4					31 %
5					

Table 1. Proportion of asynchronous frames for each pair of bands.

Table 1 shows that approximately one third of the frames are asynchronous between the bands. Since the HMMs have only three states, 1/3 of the frames represents a large proportion. This leads us to conclude that the alignments that have been proposed by our algorithm are clearly asynchronous between the bands.

423. Comparison with Some State of the Art Systems

Table 2 shows that the error rate obtained with our Multi-band system, with context-independent phone models, on the core test of TIMIT compares favorably with those already published.

Authors	*Method*	*Core test*
Lamel, Gauvain, 1993 [10]	Continuous density HMM	30.9%
Goldenthal, Glass 1994 [11]	Trajectory model	30.5%
Robinson, 1994 [12]	Recurrent ANN	26.1%
Cerisara, Fohr, Haton, 1999 [2]	Multi-Band HMM2	29.9%

Table 2. State-of-the-art phone error rate on TIMIT.

Our system gives the best accuracy on TIMIT, compared to other systems that are also based on context-independent phone HMMs.

The best results on TIMIT have been obtained with systems based on recurrent neural networks. It should thus be interesting to implement recurrent neural networks instead of HMMs into the Multi-Band paradigm. It is worth noticing that the results presented here have been obtained with clean speech. The usefulness of the Multi-band model will be more obvious with noisy speech, as shown by the partial results that we have already obtained [2].

5. CONCLUSION

51. Summary of the Work

We have shown in this paper that continuous Multi-Band speech recognition can be achieved at a low computational cost. We have obtained with our algorithm a practical complexity that is about only three times the cost of a full-band Viterbi. With such a complexity, the answer to the question of whether or not the bands must be synchronous might be not, as synchronizing the bands is only a particular case of asynchronous bands. The accuracy in the latter case is so at least as good as in the former case. And beyond the simple question of accuracy, it seems more promising to let the bands asynchronous, as fewer acoustic models than phones can be used in the bands [2].

52. Future Work

All the results presented here have been achieved with a linear Multi-Band system. However, a recombination using a Multi-Layer perceptron is generally preferred, at least in clean speech [1]. We have not presented results with such a system, since the MLP delivers scores that are similar to *a posteriori* probabilities, and not likelihoods, as it is the case with linear recombination. As the system is also a segment-based system, it cannot easily compute the segmentation by itself [9]. The solution that is usually used is to implement an independent module to compute the segmentation. It is possible to do that at low cost in our system by using for example the segmentation delivered by the linear system, but the problem is then to train the MLP on all the segments computed by this module. Work is in progress in this direction.

6. REFERENCES

[1] C. Cerisara, *Contribution of Multi-Band principle to automatic speech recognition (*In French). Ph.D. thesis, INPL, Nancy, France, September 1999.

[2] C. Cerisara, J.-P. Haton and D. Fohr, "Towards a global optimization scheme for Multi-Band Speech recognition", Eurospeech'99, Budapest, 1999.

[3] M.J.F. Gales, *Model-based techniques for noise robust speech recognition*. Ph.D. thesis, Cambridge University, September 1995.

[4] J.-F. Mari, J.-P. Haton and A. Kriouile, "Automatic word recognition based on second-order hidden Markov Models", IEEE Trans. ASSP-5(1), pp. 22-25, Janvier 1997.

[5] N. Mirghafori and N. Morgan, "Sooner or later: exploring asynchrony in Multi-Band Speech Recognition", Eurospeech'99, Budapest, 1999.

[6] C.S. Myers and L.R. Rabiner, "Connected Digit Recognition using a Level Building DTW Algorithm", IEEE Trans. ASSP-29(3), pp. 351, 1981.

[7] L.R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[8] H. Sakoe, "Two-level DP-matching. A dynamic programming-based pattern matching algorithm for connected word recognition", IEEE Trans. ASSP-27(6), pp. 588, 1979.

[9] J. Verhasselt, J.-P. Martens, I. Illina, J.-P. Haton and Y. Gong, "The Importance of Segmentation Probability in Segment Based Speech Recognizers", ICASSP'97, Munich, 1997.

[10] L.F. Lamel and J.L. Gauvain, "High Performance Speaker-independent Phone Recognition Using CDHMM", Eurospeech'93, Madrid, 1993.

[11] W.D. Goldenthal and J.R. Glass, "Statistical Trajectory Models for Phonetic Recognition", ICSLP'94, Yokohama, 1994.

[12] A. Robinson, M. Hochberg and S. Renals, "IPA: Improved Phone Modelling with Recurrent Neural Networks", ICASSP'94, Adelaide, 1994.