SOONER OR LATER: EXPLORING ASYNCHRONY IN MULTI-BAND SPEECH RECOGNITION

Nikki Mirghafori[†‡§] *Nelson Morgan*[†‡]

[†] International Computer Science Institute, 1947 Center St, Berkeley, CA 94704
[‡] University of California at Berkeley, EECS Department, Berkeley, CA 94720
Tel: (510) 643-9153, FAX: (510) 643-7684, Email: morgan@icsi.berkeley.edu
[§] Nuance Communications, 1380 Willow Road, Menlo Park, CA 94025
Tel: (650) 847-7866, FAX: (650) 847-7979, Email: nikki@nuance.com

ABSTRACT

Multi-band speech recognition is an exploratory paradigm in which each frequency region is treated as a distinct source of information and the streams are combined after each is processed independently. A number of researchers have hypothesized that it is advantageous to combine the sub-frequency information in an asynchronous manner. This paper examines this hypothesis, using two different approaches in relaxing synchrony constraints: HMM decomposition/recombination [19] and two-level dynamic programming (DP) [16].

Drawing on this work and those of others [2, 18], we conclude that relaxing the synchrony constraints indiscriminately for all phone-to-phone transitions does not consistently and significantly reduce the word error rate. The optimal permissible asynchrony must depend on both the phone-class transitions and the training-data statistics.

1. INTRODUCTION

Multi-band approaches have generated a great deal of interest in the automatic speech recognition (ASR) community [9, 2, 8]. In this paradigm, predetermined frequency sub-regions of the speech signal are treated as distinct sources of information that are processed independently and then combined to perform recognition (Figure 1). Motivations for the multi-band paradigm include results from psycho-acoustic studies, robustness to noise, and potential for parallel processing.

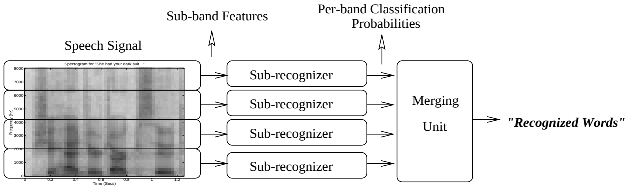

Figure 1: A simplified overview of multi-band.

In our earlier work [10], we had verified that phonetic transitions in sub-bands do not necessarily occur synchronously. Moreover, we observed that phonetic transitions patterns in sub-bands were affected by conditions such as speaking rate and room reverberation. In this paper, we try to answer the following question: can we improve recognition accuracy by relaxing synchrony constraints when combining information from different bands?

Before proceeding, let us explain what we mean by "relaxing the synchrony constraints." Such a relaxation permits a phone-state transition to occur later – or earlier – in one sub-band than another when there are sufficient acoustic cues to warrant such a decision. For example, if maximum evidence for phone transition $\alpha \rightarrow \beta$ is available in frame t in sub-band i, and the same phone transition, $\alpha \rightarrow \beta$, is best supported by the acoustic evidence in frame $t + \delta$ in sub-band j, evidence from frame t from sub-band i is combined with evidence from frame $t + \delta$ from sub-band j.

The first algorithm we considered was HMM-recombination, which is better known as either Parallel Model Combination or HMM-decomposition [19, 6]. Later, two-level dynamic programming [16] was implemented for this task. Sections 3 and 4 discuss these two algorithms and their experimental results. Section 5 includes discussion and conclusion. In the next section, we will describe the database and the system.

2. DATABASE & SYSTEM DESCRIPTION

We used the Oregon Graduate Institute Numbers95 database [4], which comprises continuous digits and numbers (total of 32 words, such as ”one”, ”sixteen”, ”forty”) recorded over the telephone as a part of census data collection. The database is phonetically hand-transcribed. For the purposes of this study, we used what is known as the ”core subset”: approximately two hours of the database for training and cross validation, and forty minutes (with non-overlapping speakers) as a test set.

We used ICSI's HMM/MLP based [3] system. For our multi-band system, we divided the frequency range into four bands of [216-778 Hz], [707-1631 Hz], [1506-2709 Hz], and [2121-3769 Hz][1], which roughly correspond to the formant regions. From the sub-bands, we derived [3rd, 3rd, 2nd, 2nd] order RASTA-PLP [7] features, respectively, as well as energy and delta RASTA-PLP and delta energy for every 25 ms window, stepped every 10 ms. We trained four MLP phonetic probability estimators on a nine-frame window of these features. The multi-layer perceptrons (MLPs) were fully connected and had [72, 72, 54, 54] inputs and [497, 497, 372, 372][2] hidden units respectively. They each also had 56 outputs (one output for each

[1] Because we used telephone quality speech, frequencies below approximately 300Hz and above 3800 were disregarded.

[2] The number of hidden units were chosen proportional to the number of input units for each net and also to make the total number of parameters equal to that of a baseline full-band system of 1000 hidden units.

phone[3]), and were trained using backpropagation with softmax normalization at the output layer. The system was trained on hand-transcribed phone labels (without embedded realignment). A multiple pronunciation lexicon (derived from the hand transcriptions), a bigram language model, and a Viterbi decode, Y0, were used for decoding.

3. HMM-RECOMBINATION

3.1. Algorithm Description

HMM-decomposition has traditionally been used for recognition in noisy conditions [19, 6]. Its main idea is to separate speech and noise into two streams, assuming that each is produced by a separate model. Similarly, HMM-recombination can combine several independent streams into a single model. An intuitive way to think of the algorithm for multi-band purposes is as follows. Consider a two-band system: if each sub-band stream is decoded independently, the acoustic data in each sub-band may best match different words. We can force both streams to consider the same word model, with identical start and finish frames, yet allow freedom for each band to transition from state to state within a word as warranted by the sub-band acoustic information. In the simple example of two uni-dimensional models, two separate streams of data, one for each sub-band, may be decoded independently using each model. Combining the two models and clamping the enter and exit states (represented as black circles) creates a two-dimensional model (as in Figure 2). The two-dimensional model could be expanded and more clearly expressed, as shown in Figure 3, where each new state is a product of two old states. In other words, assuming independence between the two bands, we have: $p(X|S_1, S_a) = p(X_1|S_1)p(X_2|S_a)$, where X_1, X_2 and X are acoustic information for band 1, band 2, and the full-band, respectively, and S_i is a state in the one-dimensional HMM. Or more generally, the likelihood may be estimated as:

$$p(X|M) = \prod_{k=1}^{K} p(X_k|M_k), \tag{1}$$

where X is the acoustic information, and M is the model.

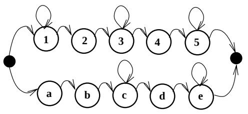

Figure 2: An unexpanded multi-dimensional HMM.

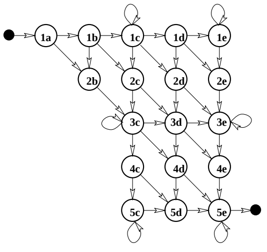

Figure 3: An expanded multi-dimensional HMM.

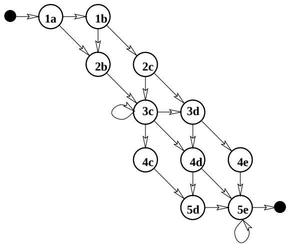

Figure 4: An expanded multi-dimensional HMM model, with maximum asynchrony limit of one state.

As the reader might suspect, the size of these multi-dimensional models can get prohibitively large as the number of sub-bands increase, and as the word models become more complex. This problem can be alleviated, to some extent, by enforcing a maximum number of states of asynchrony. For example, if a maximum asynchrony of three states is allowed, state **1e** is pruned. If asynchrony is further limited to a maximum of two states, states **1d** and **2e** are pruned. Finally, with a limit of one state, **1c, 2d, 3e** and **5c** are also pruned (Figure 4). The practical issue of the explosion of number of states is discussed in Section 3.2.

3.2. Experimental Results

The implementation of HMM-recombination was performed in the following way:

1. Multi-dimensional word models with a given maximum asynchrony constraint were created.
2. The scaled likelihoods for each new state were calculated by multiplying scaled likelihoods of the old states.
3. Viterbi decoding was run on the new multi-dimensional model, given the newly generated data likelihoods.

As suggested in Section 3.1, an explosion in the size of the models proved to be a problem, especially for a four-band system with its accompanying four-dimensional model, consistent with a similar observation by Dupont at IDIAP [5]. We therefore decided to use a two-band system where the size of the model was more manageable.

The frequency ranges for the two-band system were [216-1631 Hz] and [1506-3769 Hz]; each band of the two-band system comprised two contiguous bands of the four-band system. The 6th- and 3rd-order RASTA-PLP features, energy, delta RASTA-PLP, and delta energy for the lower and higher sub-bands were derived and MLPs with 820 and 470 hidden units were trained for the sub-bands. For a baseline, two-band comparison, a simple merging of the two bands was performed by simply adding the log likelihoods and the resulting stream was decoded using the Y0 decoder. The word error rate was 9.0%. The word error rates for HMM-recombination are listed in Table 1. Experiments with reverberant speech were also conducted, since we had earlier observed evidence for higher levels of asynchrony in reverberation [10]. It was surmised that any gains to be made for asynchronous decoding using this method would be more pronounced with reverberant speech. The reverberant data set was generated by convolving the clean set with an impulse response measured in a room having a reverberation time of 0.5 s and a direct-to-reverberant energy ratio of 0 dB. The results are reported in Table 2.

[3] Some of the 56 phones did not occur in the Numbers95 database and had zero priors.

HMM-Recombination on Clean Numbers95	
Max. Asynchrony	word error rate
None	9.0%
1	9.1%
2	9.5%
3	9.8%

Table 1: Word error rates for HMM-recombination asynchronous merging algorithm on clean Numbers95 as the maximum states of asynchrony is increased for a two-band system.

HMM-Recombination on Reverberant Numbers95	
Max. Asynchrony	word error rate
None	35.4%
1	37.1%
2	37.0%

Table 2: Word error rates for HMM-recombination asynchronous merging algorithm on reverberant Numbers95 test data as the maximum states of asynchrony is increased for a two-band system.

For both the original ”clean” speech and the reverberant speech, increasing the maximum asynchrony constraint did not improve the word recognition performance, and in fact, was slightly degraded. It may be that the synchrony requirement imposes useful constraints that negate any benefits of utilizing asynchrony. Furthermore, relaxing the synchrony requirement may increase confusion because, even though we used a 2-band system, there was still a large increase in the number of states. Also, the increased number of parameters in the expanded model may not be estimated accurately enough.

An alternative algorithm for asynchronous merging is two-level dynamic programming, which, as a bonus, does not have prohibitive space requirements, thus allowing us to experiment with a four-band system as well.

4. TWO-LEVEL DYNAMIC PROGRAMMING

4.1. Algorithm Description

The main idea of the two-level dynamic programming [16, 15] is to perform decoding in two stages: in the first stage (level 1), each individual word individual word (or some other unit) model is matched against an arbitrary portion of the test string. The second stage (level 2) of the computation pieces together the individual reference pattern scores to minimize the overall accumulated distance over the entire test string. In the case reported here, the two-pass approach permitted an efficient implementation of the desired asynchronous merging.

4.2. Experimental Results

In our implementation of the two-level dynamic programming, synchrony is enforced on the word level. In the first stage of the algorithm, for every word and every sub-band, a distance matrix is calculated. Every entry (i, j) in the distance matrix signifies how likely it is for the word to have been uttered, starting at frame i and ending at frame j. Synchrony at the start and end of the word unit is enforced by adding all the distance matrices for the sub-bands for a given word. Next, word lattices are created for each utterance. Finally, a search on the lattices[4] [13, 12, 14] is performed to generate the string with the least distance.

Two-level Dynamic Programming on Numbers95		
Condition	Simple Merge WER	2-level DP WER
Clean, 4 bands	11.5%	11.1 %
Clean, 2 bands	9.0%	9.3%
Reverb, 4 bands	39.1%	45.9 %
Reverb, 2 bands	35.4%	37.7 %

Table 3: Word error rates (WER) for two-level dynamic programming for clean and reverberant speech on the Numbers95 test set.

Without any pruning, the produced lattice can be unmanageably large, since there would be an arc for every word for every start and finish time. On-line garbage model pruning [1] was used. This is a simple on-line pruning method that has proved as effective as some of the more sophisticated off-line approaches that require training. Dynamically, the top n scores are averaged; the scores above this threshold are kept, and the rest are pruned. Experimentally, it was determined that $n = 10$ produced lattices of reasonable size in an acceptable amount of time. The word error rates are reported in Table 3.

Similar to the HMM-recombination algorithm, increasing the asynchrony limit using two-level dynamic programming did not improve the recognition results. Asynchronous merging does not seem to have a significant effect on the word error rate. The word error rates are very similar to when the streams are simply log linearly merged. It is not too surprising since, in both methods, the probability streams are multiplied. If alternative asynchronous paths are not exploited, the two methods of stream merging are essentially equivalent, which would explain the results.

For reverberant speech, the phone-transitions in the sub-bands are more temporally spread, as observed in the analysis in [10], so asynchronous combination of the sub-bands should affect the results more dramatically. We see in Table 3 that the results, however, may depend on the size of the sub-band. If the sub-bands are large enough to include sufficient acoustic cues, the probability stream would be relatively accurate, and the freedom to merge asynchronously hurts only slightly – analogous to the clean speech case. However, if the sub-bands are too narrow, in the presence of reverberation, acoustic cues would be gravely degraded, and in combination with relaxed constraints, the word error rate may increase significantly.

5. CONCLUSIONS

It was observed that relaxing the synchrony constraints when merging the multi-band streams did not improve word recognition accuracy. The results were consistent both for word-level

[4] A lattice, for the purposes of this work, is defined as a directed acyclic graph where each edge corresponds to a word with a distance score and each node corresponds to a point in time.

(in the case of two-level dynamic programming) and multiple state-level (for HMM-recombination) relaxation of synchrony constraints.

Tibrewala and Hermansky [17] had also observed differences in optimal sub-band paths and conjectured that relaxing the temporal synchrony requirement among sub-bands would improve the word error rate. The word error rate increased slightly (though, not statistically significantly) when relaxing the synchrony requirement over a word (using Viterbi decoding on each stream), compared to enforcing synchrony at every state for an isolated digit recognition task both for a four-band and a seven-band system. In a similar experiment for isolated German word recognition, Bourlard and Dupont [2] observed no improvement when the synchrony was relaxed from the frame level to the phone level, and only a small (not statistically significant) improvement when the merging was performed on the syllable level for a three-band system using HMM-recombination. Finally, Tomlinson et al. [18] have reported a slight improvement ($p < 0.1$) when synchronization is performed on a three-state, instead of a per-state, level for a two-band system, and no improvement for a three-band system using HMM-recombination.

It may be that by disregarding the synchrony information, important information is being lost. As observed in [10], some broad phone category transitions in some sub-bands occur systematically earlier or later than the full-band average. If at all, synchrony requirement should perhaps be relaxed only for particular phone transitions, and not indiscriminately for all phone transitions, as it has been done in previous work. Furthermore, the amount of the permitted asynchrony may have to depend on the phone-class transitions and the training-data statistics. Along these lines, Morris and Pardo [11] have also observed that the patterns of onsets or offsets of the phone transitions across frequency bands tend to be quite stable for each transition, suggesting that warping the sub-bands to align the transitions might remove the potentially useful information that this characteristic transition pattern could provide.

In summary, various reasons justify the observed results. The transition constraint may be aiding the Viterbi search by reducing the number of potential paths and transition options. On the other hand, it may be that there is a “gain” from relaxing the synchrony assumptions, but only for a limited number of phone-to-phone transitions, and that by allowing synchrony relaxation for all transitions, this “gain” is lost. Also, we have dramatically increased the size of the parameter space and are using a simple technique (multiplying scaled likelihoods) for estimating these parameters.

Based on the significantly higher computational costs and the currently available evidence, we are forced to conclude that relaxing the synchrony constraint, in this form, is not advantageous, and a more detailed set of constraints for multi-band asynchronous merging is warranted, as opposed to the approaches that we and others have so far explored.

Acknowledgments

We would like to especially thank Eric Fosler-Lussier, Su-Lin Wu, Brian Kingsbury, and Steve Greenberg at ICSI, and Greg De Haan at Nuance Communications. We would further like to acknowledge our colleagues Hervé Bourlard, Stéphane Dupont, Hynek Hermansky, and Sangita Sharma for multi-band collaboration over the past few years. We thank Jim West and Gary Elko, from Bell Labs, and Carlos Avendano, at the University of California, Davis, for the room impulse responses. This work was supported by the National Science Foundation grant 9712579, the European Community Basic Research grant (Project SPRACH), and the International Computer Science Institute.

6. REFERENCES

[1] Hervé Bourlard, Bart Dh̀oore, and Jean-Marc Boite. Optimizing recognition and rejection performance in wordspotting systems. In *ICASSP*, volume 1, pages 373–376, Adelaide, South Australia, April 1994.

[2] Hervé Bourlard and Stéphane Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *ICSLP*, pages 426 – 429, Philadelphia, PA, USA, October 1996.

[3] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic Press, 1994.

[4] Numbers corpus, release 1.0, 1995.

[5] Stéphane Dupont. Personal communication, 1996.

[6] M. J. F. Gales and S. J. Young. An improved approach to the hidden Markov model decomposition of speech and noise. In *ICASSP*, pages 233–236, San Francisco, CA, March 1992.

[7] Hynek Hermansky and Nelson Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[8] Hynek Hermansky, Sangita Tibrewala, and Misha Pavel. Towards ASR on partially corrupted speech. In *ICSLP*, pages 462 – 465, Philadelphia, PA, USA, October 1996.

[9] Naghmeh Nikki Mirghafori. *A Multi-Band Approach to Automatic Speech Recognition*. PhD thesis, University of California at Berkeley, Berkeley, CA, December 1998.

[10] Nikki Mirghafori and Nelson Morgan. Transmissions and transitions: A study of two common assumptions in multi-band ASR. In *ICASSP*, volume 2, pages 713–716, Seattle, WA, May 1998.

[11] A. C. Morris and J. M. Pardo. Phoneme transition detection and broad classification using a simple model based on the function of onset detector cells found in the cochlear nucleus. In *EUROSPEECH*, pages 115–118, Madrid, Spain, 1995.

[12] H. Ney, R. Haeb-Umbach, B.-H. Tran, and M. Oerder. Improvements in beam search for 10,000-word continuous speech recognition. In *ICASSP*, volume 1, pages 9–12, San Francisco, California, March 1992. IEEE.

[13] Hermann Ney and Xavier Aubert. A word graph algorithm for large vocabulary, continuous speech recognition. In *ICSLP*, pages 1355–1358, Yokohama, Japan, September 1994.

[14] Martin Oerder and Hermann Ney. Word graphs: An efficient interface between continuous-speech recognition and language understanding. In *ICASSP*, volume 2, pages 119–122, Minneapolis, Minnesota, April 1993. IEEE.

[15] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[16] Hiroaki Sakoe. Two-level DP-matching– a dynamic programming-based pattern matching algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-27(6):588–595, December 1979.

[17] Sangita Tibrewala and Hynek Hermansky. Sub-band based recognition of noisy speech. In *ICASSP*, volume 2, pages 1255–1258, May 1997.

[18] M. J. Tomlinson, M. J. Russell, R. K. Moore, A. P. Buckland, and M. A. Fawley. Modelling asynchrony in speech using elementary single-signal decomposition. In *ICASSP*, volume 2, pages 1247–1250, April 1997.

[19] A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *ICASSP*, pages 845–848, 1990.