ASYNCHRONY WITH TRAINED TRANSITION PROBABILITIES IMPROVES PERFORMANCE IN MULTI-BAND SPEECH RECOGNITION

Brian Mak and Yik-Cheung Tam

Department of Computer Science,
The Hong Kong University of Science and Technology,
Clear Water Bay, Hong Kong
{mak,cswilson}@cs.ust.hk

ABSTRACT

One of the central themes in multi-band automatic speech recognition (ASR) is to devise a strategy for recombining sub-band information. This in turn raises two questions: (1) at what phonetic unit should the recombination take place? (2) How asynchronously should the sub-bands be run? Theoretically asynchronous multi-band ASR should perform at least as well as synchronous multi-band ASR. However, in the past few years, there are conflicting results on the issue. In this paper, we study the asynchrony issue under the framework of HMM composition in which a model-based recombination strategy is used to recombine sub-band HMMs at the state level. We hypothesize that re-estimation of the transition probabilities is crucial for multi-band ASR (using HMM composition). Experiments on connected TI digits show that for both clean speech and noisy speech (with additive white noise of 10db), HMMs composed from sub-band HMMs in which transition probabilities are trained with Baum-Welch algorithm outperform those in which transition probabilities are set uniformly (e.g. 0.5 in common left-to-right HMMs) by about 20%. Recombining sub-bands with a maximum asynchrony limit of one state gives a further $\sim 15\%$ improvement over synchronous recombination on both clean speech and noisy speech. Finally relaxing asynchrony to more than one state results in worse performance.

1. INTRODUCTION

Recently, multi-band speech recognition has been proposed by Bourlard *et al.* [3] and Hermansky *et al.* [5]. This approach is motivated by the empirical findings by Harvey Fletcher of Bell Labs [1] from a thorough study of human speech recognition. In multi-band approach, the full frequency band is divided into sub-bands and a speech recognizer is built for each band. During recognition, decisions from individual sub-band recognizers are recombined to arrive at a final decision at some phonetic/linguistic level.

One major issue in multi-band ASR is how to recombine sub-band information. Some researchers observed that transitions in sub-bands occur asynchronously [7, 4] and suspected that it may be advantageous to combine sub-band decisions in this way. There are two approaches to recombining sub-bands asynchronously:

(1) Segment-based approach includes the classical two-level dynamic programming algorithm and the level-building algorithm [9]. An efficient hybrid of these two algorithms is recently suggested by [4]. One advantage of segment-based methods is that they allow recombination at larger (phonetic) units.

(2) Frame-based or model-based approach creates a composite HMM out of the sub-band HMMs [8] which encapsulates state asynchrony in the models so that subsequent decoding can be done synchronously at the frame level.

Although theoretically, an asynchronous multi-band system should perform at least as well as a synchronous system, since if synchrony is really preferred, an asynchronous system may simply fallback to the synchronous mode. However, the opposite is empirically found in [8]. In this paper, we study the effect of transition probabilities on the asynchronous recombination of sub-band HMMs. We first notice that, in practice, transition probabilities of full-band systems are obtained either from Baum-Welch training [2] or by simply setting them uniformly to the reciprocal of the number of outgoing arcs of a state. The latter case is commonly used in hybrid ANN/HMM systems. The main justifications are: (1) the exponential duration model implicitly implied for an HMM is not correct; (2) empirically, many people did not find them helpful anyway; and (3) the dynamic range of transition probabilities are much smaller than that of observation probabilities. Thus, in our study, we will investigate both ways of setting the transition probabilities . To avoid explosion of the state space in HMM composition, only 2-band systems are studied. We also follow the procedure in [8] and vary the maximum degree of asynchrony (in terms of number of states) in each experiment. In Section 2, we first present the HMM composition algorithm followed by recognition experiments in Section 3. Discussion and conclusions are made in Section 4.

2. HMM COMPOSITION ALGORITHM

HMM composition algorithm is similar to Parallel Model Combination (PMC) [10]. PMC is originally applied to noise compensation where a clean speech model is combined with a noise model to simulate a "noisy" speech model. Mirghafori *et al.* applied the algorithm to multi-band ASR [8] in which sub-band HMMs are recombined in a similar manner to form a composite HMM so as to allow asynchronous state transitions of the various sub-bands.

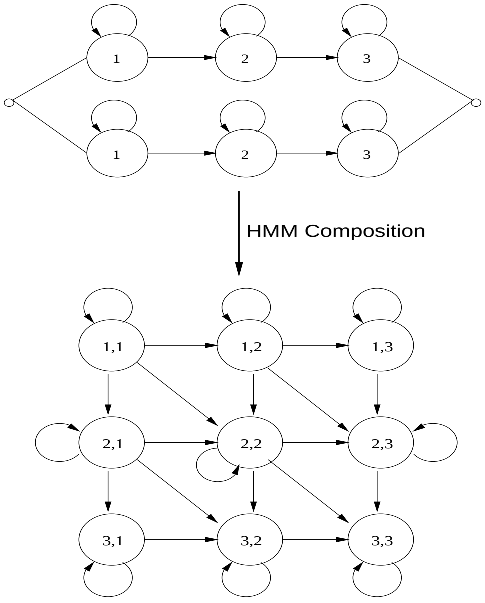

Figure 1: HMM composition of two 3-states left-to-right sub-band HMMs with a maximum asynchrony limit of two states.

For the sake of completeness, we briefly describe the algorithm below.

To illustrate the idea, HMM composition is applied to a 2-band system consisting of two 3-state left-to-right sub-band HMMs as shown in Figure 1, allowing a maximum asynchrony of two states. In this case, the composite HMM can have $3^2 = 9$ "composite states", (i, j), $1 \leq i, j \leq 3$, where state i of the first sub-band HMM recombines with state j of the second sub-band HMM. For instance, at the composite state $(3, 1)$, the third state of the first sub-band HMM recombines with the first state of the second sub-band HMM. In the composite HMM, there are two kinds of composite states: synchronous or asynchronous states. If all HMMs are strictly left-to-right with the same number of states which are numbered successively from left to right, then a synchronous state has the form (i, i) and an asynchronous state is represented by $(i, j), i \neq j$. Notice that sub-bands are automatically synchronized at the beginning and ending states but asynchronous paths inside the composite model are allowed during decoding.

In general, for a K-band system, we denote the log likelihood of a composite state $(i_1, \ldots, i_K)$ as $L_{(i_1,\ldots,i_K)}$ and the log likelihood of state i_k in sub-band k as L_{i_k}. In this paper, the log likelihood of each composite state is linearly combined with sub-band weightings as follows:

$$L_{(i_1,\ldots,i_K)} = \sum_{k=1}^{K} \omega_k L_{i_k} \ , \tag{1}$$

where ω_k denotes the weighting of sub-band k satisfying the constraint

$$\sum_{k=1}^{K} \omega_k = 1 \ . \tag{2}$$

Table 1: Total number of states N, and transitions M for various numbers of sub-bands and maximum asynchrony limits L.

#Sub-bands	L	N	M
1	0	3	5
2	0	3	5
2	1	7	19
2	2	9	25
3	0	3	5
3	1	15	65
3	2	27	125
4	0	3	5
4	1	31	211
4	2	81	625

The transition probabilities of a composite HMM can be computed as follows:

$$a_{(i_1,\ldots,i_K),(j_1,\ldots,j_K)} = \prod_{k=1}^{K} a_{i_k,j_k} \tag{3}$$

where a_{i_k,j_k} is the transition probability from state i_k to j_k in sub-band k. Then these transitional probabilities are normalized to one. i.e. $\sum_{s \in \mathcal{S}} a_{(i_1,\ldots,i_K),s}$ where $\mathcal{S}$ is the set of states that $(i_1, \ldots, i_K)$ transits to.

It is obvious that as the number of sub-bands and the degree of asynchrony increase, the number of states and transition arcs in a composite HMM increases drastically as shown in Table 1, requiring more computation during decoding. To alleviate the problem, one may impose a maximum asynchrony limit. For example, with a limit L, composite states $(i_1, \ldots, i_K)$ are not allowed when the following condition is satisfied:

$$\max\{i_1, \ldots, i_K\} - \min\{i_1, \ldots, i_K\} > L \tag{4}$$

3. RECOGNITION EXPERIMENTS AND RESULTS

We chose connected TI digits [6] as our training and evaluation corpus for two reasons: (1) No language models are used in the task so that we can be sure if there is any benefit due to asynchronous recombination of acoustic models, it will not be shadowed by the use of language models. (2) If we allow a high degree of asynchrony, the generated composite HMMs can be very complex and subsequent computation can be quite substantial. Using the simple task allows us to concentrate on the asynchrony issue and experiment in a more manageable amount of time.

The whole training set contains 4235 digit strings from male speakers while the test set contains 4311 digit strings from another set of male speakers. A two-band system with equal sub-band weights was investigated. The frequency range from 100Hz to 4400 Hz are equally partitioned in the critical band scale into two sub-bands as follows:

- Band-1: 100 – 1080 Hz
- Band-2: 1000 – 4400 Hz .

Speech data were low-passed at 4400Hz and MFCCs were extracted from a window of 20ms at a frame rate of 100Hz. Each sub-band acoustic vector consists of 6 MFCCs and the normalized energy concatenated with their delta and acceleration features. Cepstral mean subtraction was performed as well.

All sub-band HMMs are strictly left-to-right whole-word models with 6 states and 4 Gaussian mixtures per state. Only *clean* speech was used to train each sub-band HMM independently. Afterwards, composite HMMs are generated using the HMM composition algorithm subject to a maximum asynchrony limit which is varied from $0 \leq L \leq 5$ ($L = 0$ and $L = 5$ refer to a forced synchronous system and a fully-asynchronous system respectively). In any case, the sub-band state observation pdf's are not modified during HMM composition. During decoding, Viterbi search was run on a network of composite digit HMMs. No (beam) pruning is performed so that all possible paths are searched.

3.1. Full-band Results

Full-band HMMs were first trained for each digit using the clean training data. It gives a word error rate of 3.01% on the clean test set and 15.6% on the noisy test set with additive white noise of 10db.

3.2. Experiment I & II: HMM Composition Using Sub-bands with Uniform Transition Probabilities

In these two experiments, the transition probabilities of the sub-band HMMs are uniformly set (to 0.5 in this case) and only the sub-band state observation pdf's are trained. Then during HMM composition, the transition probabilities in the composite HMMs are either computed using Eqn. (3) or are re-trained using the Baum-Welch algorithm. The procedure is applied to clean speech in Experiment I and noisy speech with 10dB additive white noise in Experiment II. From the results tabulated in Table 2 and 3, we found that

- It is beneficial to re-train transition probabilities after HMM composition in multi-band recognition system. The performance with re-trained transitions are always better than that with uniform transitions.
- In most cases, and certainly in the case with re-trained transition probabilities, asynchronous recombination of sub-bands produces much better recognition.
- In all cases, limiting asynchrony to a maximum of one state gives the best results — about 20% reduction in word error rate (WER) over synchronous recombination of sub-bands.

3.3. Experiment III & IV: HMM Composition Using Sub-bands with Trained Transition Probabilities

The assumption of uniform transition probabilities results in "non-optimal" HMMs (in the maximum likelihood sense). We repeated the previous two experiments except that the sub-band transition probabilities were *trained* prior to HMM composition. Transition probabilities of the composite HMMs were then computed or re-trained. Same evaluation was performed as in Section 3.2 and the results were tabulated in Table 4 and Table 5. The following points are worth noting:

- Comparing Table 1 and 3, and Table 2 and 4, we notice that composite HMMs generated from sub-band models wherein transition probabilities are Baum-Welch trained outperform their counterparts generated from sub-band models with uniform transition probabilities, resulting in an error reduction of about 20%.
- Starting from sub-band models with Baum-Welch trained transitions, the performance gap between HMM composition with computed and re-trained transition probabilities is now small, though the latter method is still always better.
- Another main difference from Experiment I & II is that now asynchrony of more than one state does not help anymore — quite to the opposite, it hurts performance.
- Nevertheless, limiting asynchrony to maximum one state continue to gives significant gain in performance: 13.8% in clean speech and 15.2% in noisy speech with 10db white noise.

4. DISCUSSION AND CONCLUSION

HMM composition is an effective way to implement asynchronous recombination of sub-bands in multi-band ASR. Our best results show that HMM composition reduces word error error on connected TI digits from full-band's 3.01% to 2.01% (relatively 33.2%) in synchronous recombination of sub-bands, and further to 1.81% (relatively 39.9%) in the best asynchronous recombination of sub-bands for clean speech. The figures are 15.6%, 11.1% (relatively 28.8%), and 9.41% (relatively 39.7%) respectively for noisy speech with additive white noise of 10db.

Two factors seems to be important for HMM composition to be effective:

(1) The transition probabilities of the sub-band HMMs should be trained by the Baum-Welch algorithm. The common heuristics of simply setting them equal among the outgoing arcs gives inferior results.

(2) Asynchronous recombination with a maximum asynchrony limit of one state helps.

The two factors together make sense: uniformly setting the transition probabilities unnecessarily imposes the same and probably wrong duration model on the corresponding states of various sub-bands. By examining the trained transition probabilities of the two sub-band HMMs of the same digits, we find that the corresponding transition probabilities are similar but far from 0.5 — a_{ii} are more around 0.8 instead. As they are similar but not identical, synchronous recombination are preferred though there is still room for better performance by going asynchrony. Further examination of the trained transition probabilities of the composite HMMs confirms that (a) transitions from synchronous states to synchronous states are more likely; and (b) asynchronous states are less likely to make a transition

Table 2: Experiment I: % Word error rates on clean speech. The percentages in bracket are error reductions relative to $L = 0$.

L	(Uniform Sub-band a_{ij} Prior to Composition) How a_{ij} Are Set in the Composite HMMs?	
	Computed (Uniform)	Re-trained
0	2.55	2.56
1	2.27 (-11.0%)	2.03 (-20.7%)
2	2.90 (+13.7%)	2.39 (-6.64%)
3	2.87 (+12.5%)	2.37 (-7.42%)
4	2.92 (+14.5%)	2.37 (-7.42%)
5	2.95 (+15.7%)	2.37 (-7.42%)

Table 3: Experiment II: % Word error rates on noisy speech with 10db additive white noise. The percentages in bracket are error reductions relative to $L = 0$.

L	(Uniform Sub-band a_{ij} Prior to Composition) How a_{ij} Are Set in the Composite HMMs?	
	Computed (Uniform)	Re-trained
0	16.24	14.34
1	13.12 (-19.2%)	11.41 (-20.4%)
2	13.74 (-15.4%)	11.67 (-18.6%)
3	13.75 (-15.3%)	11.68 (-18.5%)
4	13.77 (-15.2%)	11.71 (-18.3%)
5	13.76 (-15.3%)	11.68 (-18.5%)

Table 4: Experiment III: % Word error rates on clean speech. The percentages in bracket are error reductions relative to $L = 0$.

L	(Trained Sub-band a_{ij} Prior to Composition) How a_{ij} Are Set in the Composite HMMs?	
	Computed	Re-trained
0	2.01	2.10
1	1.84 (-8.46%)	1.81 (-13.8%)
2	2.45 (+21.9%)	2.36 (+12.4%)
3	2.44 (+21.4%)	2.37 (+12.9%)
4	2.45 (+21.9%)	2.37 (+12.9%)
5	2.45 (+21.9%)	2.37 (+12.9%)

Table 5: Experiment IV: % Word error rates on noisy speech with 10db additive white noise. The percentages in bracket are error reductions relative to $L = 0$.

L	(Trained Sub-band a_{ij} Prior to Composition) How a_{ij} Are Set in the Composite HMMs?	
	Computed	Re-trained
0	11.19	11.10
1	9.48 (-15.3%)	9.41 (-15.2%)
2	11.34 (+1.34%)	11.20 (+0.90%)
3	11.36 (+1.52%)	11.21 (+0.99%)
4	11.40 (+1.88%)	11.20 (+0.90%)
5	11.40 (+1.88%)	11.20 (+0.90%)

that will further increase asynchrony between sub-bands. In other words, asynchronous transitions of more than one state are much less likely. This observation may explain the phenomenon that systems with maximum asynchrony limit from two to five states have similar performance.

Although our results show that asynchrony by one state improves performance over forced synchrony, further relaxing asynchrony beyond one state results in degraded performance in all reported experiments. This may be caused by inaccuracies of the state observation distributions in the composite HMMs. In other words, re-training the observation distributions of composite states may be required besides re-training the transition probabilities.

5. ACKNOWLEDGEMENTS

This work is supported by the Hong Kong RGC under the grant number CA97/98.EG02, and by the grant HKTIIT 98/99.EG01 from the Cable & Wireless HKT.

6. REFERENCES

[1] J. B. Allen. How Do Humans Process and Recognize Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, 1994.

[2] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

[3] H. Bourlard and S. Dupont. A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands. In *ICSLP*, October 1996.

[4] Dominique Fohr Christophe Cerisara and Jean-Paul Haton. Asynchrony in multi-band speech recognition. In *ICASSP*, volume 2, pages 1121–1124, 2000.

[5] H. Hermansky, S. Tibrewala, and M. Pavel. Towards ASR on Partially Corrupted Speech. In *ICSLP*, October 1996.

[6] R.G. Leonard. A Database for Speaker-Independent Digit Recognition. In *ICASSP*, 1984.

[7] 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, 1998.

[8] Nikki Mirghafori and Nelson Morgan. Sooner or later: Exploring asynchrony in multi-band speech recognition. In *Eurospeech*, 1999.

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

[10] A.P. Varga and R.K. Moore. Hidden Markov Model Decomposition of Speech and Noise. In *ICASSP*, pages 845–848, 1990.