Development of an Asynchronous Multi-band System for Continuous Speech Recognition

Yik-Cheung Tam, Brian Mak

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

Abstract

Recently, multi-band automatic speech recognition (MBASR) is proposed to combat environmental noises. In this paper, we describe the two major efforts in the development of our asynchronous MBASR system for *continuous* speech recognition. Firstly, we successfully introduce asynchrony among sub-bands under the HMM composition framework. An asynchrony limit of one state is found adequate — relaxing the limit further does not improve performance. Secondly, the sub-band log likelihoods are combined linearly at the frame level with weightings estimated by minimizing the *string* classification error (MCE) among the N-best hypotheses using simulated noisy speech. When our asynchronous MBASR system is evaluated on connected TI digits with 0db additive low-pass white noise, compared with a full-band system, (1) our synchronous sub-band system reduces the absolute string error rate (SER) and word error rate (WER) by 19.8% and 14.1% respectively; (2) the introduction of asynchrony further reduces the absolute SER (WER) by 5.2%(2.5%); (3) an estimation of sub-band weightings using N-best string MCE training gives an additional reduction of absolute SER (WER) by 19.7% (5.1%). Thus, in that test, our asynchronous MBASR system has outperformed a full-band system with a 44.7% (21.7%) reduction in absolute SER (WER). In summary, N-best MCE training can effectively emphasize the more reliable sub-band, and asynchronous recombination of sub-bands is preferred.

1. Introduction

Multi-band speech recognition has been shown to improve robustness under noisy environment recently [1, 2]. Two major issues are often encountered in designing an MBASR system: namely, asynchrony among the sub-bands, and their likelihood recombination. The asynchrony issue is motivated by the observation that transitions of sub-band are asynchronous and suspected that it may be advantageous to combine sub-band decisions in this way. Traditionally, this issue is handled by two-level dynamic programming algorithm or the level-building algorithm [3]; recently, a hybrid approach of the two methods were proposed in [4]. However, it is not easy to extend the method to continuous speech recognition in practice. More recently, new approaches such as asynchronous transition HMM [5], factorial HMMs [6], Bayesian network [7], and HMM composition [8, 9] are proposed. In our previous work [9], we showed how sub-band HMMs can successfully be combined by HMM composition. Since the output of HMM composition is just another HMM, it allows us to stay in the HMM framework and continue to enjoy the benefits provided by all HMM-based techniques which contribute to the success of ASR in the last decade. Yet, HMM composition still allows sub-band decoding to be done asynchronously at sub-word or word levels as desired.

Based on the HMM composition framework, the second issue becomes how to recombine sub-band log-likelihoods at the frame level. One approach is to recombine them linearly with sub-band weights "optimally" [1] derived by minimum classification error (MCE) training as shown in [10] and in our previous work [11]. However, both works only dealt with isolated speech recognition using word-level MCE training. In this paper, we extend our work to training sub-band weights in continuous speech recognition. *String-level* MCE is used as the optimization criterion, competitors are derived from N-best hypotheses, and simulated noisy speech are used for the estimation. In practice, we expect these weights to be estimated offline for each noise type at various signal-to-noise ratios (SNR). During recognition, an MBASR system will first estimate the type of environmental noise and its SNR so that the corresponding sub-band weights can readily be plugged into the system.

We first elaborate our approach of asynchronous HMM composition in Section 2, and derive the string-level MCE estimation formulae for estimating sub-band weights in Section 3. This is followed by recognition experiments in Section 4, discussion in Section 5, and conclusions in Section 6.

2. HMM Composition Algorithm

HMM composition algorithm originates from Parallel Model Combination (PMC) which is applied to noise compensation where clean speech models are combined with a noise model to simulate "noisy" models. Mirghafori *et al.* applied the algorithm to MBASR [8] in which sub-band HMMs are combined similarly to form a composite HMM, and asynchrony is captured by the creation of asynchronous composite states.

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. State i of the first sub-band HMM and state j of the second sub-band HMM are recombined at the "composite state", (i,j), $1 \leq i, j \leq 3$. Here, synchronous states have the form of (i, i) and asynchronous states are represented by (i, j), $i \neq j$. Notice that the sub-bands are automatically synchronized at model boundaries but asynchronous paths inside the composite model are allowed during decoding.

It is obvious that when the number of sub-bands and the degree of asynchrony increases, the number of states and tran-

[1] "Optimality" applies only to linear re-combination of sub-band likelihoods. In fact, they can be re-combined non-linearly using MLP.

sitions increases drastically. To alleviate the problem, one may impose a maximum asynchrony limit to prune away those "out-of-sync" nodes as illustrated in Figure 1.

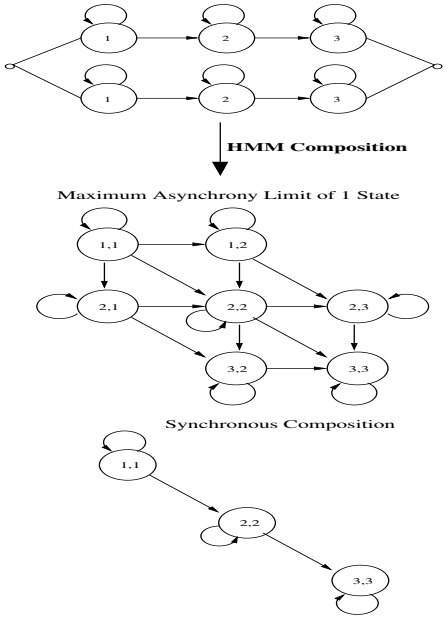

Figure 1: Synchronous and asynchronous HMM composition of two 3-state left-to-right sub-band HMMs with a maximum asynchrony limit of one state.

3. String Minimum Classification Error Estimation of Linear Sub-band Weights

For simplicity, we will only show estimation formulae [2] in which sub-band weights depend only on the sub-band and not on the model nor the state; the latter can easily be derived in a similar fashion.

Suppose we have a K-sub-band recognition system and N competing strings generated by an N-best algorithm. We define the observation sequence $X = [x_1, x_2, \ldots, x_T]$ and Λ as a set of acoustic models. The discriminant function for the class W_j is chosen to be

$$g_j(X;\Lambda) \quad = \quad \log P(X, \overline{q}_j | W_j) \tag{1}$$

where $\overline{q}_j = \{\overline{q}_j^1, \overline{q}_j^2, \ldots, \overline{q}_j^K\}$ denotes the most likely composite state sequence of W_j and $\overline{q}_j^k$ denotes the corresponding state sequence in sub-band k. Sub-band log-likelihoods are recombined at the frame level as follows:

$$\log b_{\overline{q}_{j(t)}}(x_t) = \sum_{k=1}^{K} \omega_k \log b^k_{\overline{q}^k_{j(t)}}(x_t^k) \tag{2}$$

$$\text{where} \quad 0 \le \omega_k \le 1 \text{ and } \sum_{k=1}^{K} \omega_k = 1 \tag{3}$$

[2] For a two-band system, one may find the weight efficient using linear search. But here, we are presenting a general solution for a K-band system.

assuming sub-band independence. Let $g_j(X;\Lambda)$ be g_j, and the misclassification measure $d(X)$ is then given by

$$d(X) = \log \left\{ \frac{1}{N} \sum_{n \ne j} exp(g_n) \right\} - g_j \ . \tag{4}$$

The misclassification measure is smoothed using the sigmoid function, $l(d) = 1/\{1 + \exp(-\gamma d + \theta)\}$, to obtain the total loss function over all utterances, $R_{mce} = \frac{1}{U} \sum_X l(d(X))$ where U denotes the total number of training utterances. To satisfy Eqn.(3) throughout the optimization process, parameter transformation is performed: $\overline{\omega}_k = log(\omega_k)$. Taking derivative with respect to the k-th sub-band weight, $\overline{\omega}_k$, we have

$$\frac{\partial R_{mce}}{\partial \overline{\omega}_k} = \frac{1}{U} \sum_X \frac{\partial l}{\partial d} \frac{\partial d(X)}{\partial \overline{\omega}_k}, \quad \text{where} \tag{5}$$

$$\frac{\partial l}{\partial d} \quad = \quad \gamma l(d(X))[1 - l(d(X))] \,, \tag{6}$$

$$\frac{\partial d(X)}{\partial \overline{\omega}_k} \quad = \quad \frac{\sum_{n \ne j}^{N} exp(g_n)(\Delta g_{nk} - \Delta g_{jk})}{\sum_{n \ne j}^{N} exp(g_n)} \tag{7}$$

and, $\Delta g_{nk} = exp(\overline{\omega}_k) \sum_{t=1}^{T} \log b^k_{\overline{q}^k_{n(t)}}(x_t^k)$.

The iterative gradient-descent method is employed to get a better estimate of the sub-band weights for the $(s+1)$-th iteration from their estimates from the s-th iteration as follows:

$$\overline{\omega}_k{}^{(s+1)} = \overline{\omega}_k{}^{(s)} - \epsilon_s \left(\frac{\partial R_{mce}}{\partial \overline{\omega}_k} \right) \tag{8}$$

where ϵ_s is the learning rate at the s-th iteration. Finally, $\overline{\omega}_k$ is transformed back to ω_k after the gradient-descent procedure completes.

4. Recognition Experiments and Results

Speaker-independent connected TIDIGITS was chosen for evaluating our asynchronous MBASR system. The adult training set contains 8623 digit strings while the test set contains 8700 digit strings. Only clean speech were used to train our system and noisy data for testing were created from the corresponding clean speech by adding the noise at a prescribed SNR *at the time domain* after the end-points had been detected. Noisy speech were also created from one-tenth of the whole training set (randomly picked) for the MCE training of sub-band weights in a similar manner. Corrective training was employed, and only one-best competitor was used. Accurate end-points were used for the recognition experiments.

Two sub-bands were adopted by partitioning the frequency ranges, 0 Hz — 4000 Hz, equally in the critical band scale as follows:

- Band-1: 0 – 1080 Hz
- Band-2: 1000 – 4000 Hz .

Speech data were low-passed at 4000Hz and MFCCs were extracted from a window of 20ms at a frame rate of 100Hz. The full-band acoustic vector consists of 12 MFCCs and the normalized pass-band energy + Δ + $\Delta\Delta$ while a sub-band acoustic vector consists of 6 MFCCs and the normalized pass-band energy + Δ + $\Delta\Delta$. Cepstral mean subtraction was performed as well.

All (full-band or sub-band) HMMs are strictly left-to-right, whole-word models with 6 states and 4 Gaussians mixture per state. They were all trained with *clean* speech.

Table 1: The effect of asynchrony limit (word error rate). (FB=full-band, B1=subband-1, B2=subband-2, SYNC-1:1=synchronous MB with equal weights, ASYNC-1:1= asynchronous MB with equal weights, SYNC-MCE=synchronous MB with MCE-derived weights, ASYNC-MCE=asynchronous MB with MCE-derived weights)

System	Clean TIDIGITS
FB	1.20%
SYNC-1:1	1.22%
ASYNC-1-1:1	0.98%
ASYNC-2-1:1	0.99%
ASYNC-3-1:1	0.98%
ASYNC-4-1:1	0.99%
ASYNC-5-1:1	1.00%

Table 2: Changes in the sub-band weightings of Band-1 as SNR decreases on low-pass white noise.

Weights	-10dB	0dB	10dB	20dB	clean
SYNC	0.209	0.190	0.372	0.500	0.500
ASYNC	0.217	0.203	0.389	0.499	0.500

4.1. Experiment I: Asynchrony Limit

We first tested the effect of asynchrony. Table 1 shows the recognition performance on clean speech[3] by several systems: full-band system, synchronous multi-band system, and asynchronous multi-band systems with various limits of asynchrony. To obtain these results, we found that it is crucial to train the transition probabilities of sub-band models by the Baum-Welch algorithm before HMM composition. The transition probabilities of the composite HMMs may simply be computed from their sub-band counterparts; further re-training of these transition probabilities does not make much difference [9]. From Table 1, an asynchrony limit of one state seems adequate, reducing the relative WER of an synchronous composition system by 16.7%. In addition, allowing asynchrony of more than one state does not show further improvement. After taking the time and space requirements into consideration, an asynchrony limit of one state seems to be a good choice and that is what we adopted for all the following experiments.

4.2. Experiment II: MCE Training of Sub-band Weights

4.2.1. Case 1: Additive Low-pass white noise

We first investigated how well a multi-band system can do under ideal band-limited noise when only Band-1 was corrupted by a low-pass white noise. Results are shown in Figure 2 and Figure 3.

Since only Band-1 is corrupted, Band-2 maintains good performance under all SNRs. Among all configurations, the asynchronous multi-band recognizer with MCE-derived weights performs better than the others. Compared with the full-band recognizer, the asynchronous multi-band recognizer reduces relative string (word) error rates by 55.6% (70.8%) at an SNR of 0db. Table 2 summarizes the change of the first sub-band's weight as the SNR decreases. The table shows that the first sub-band is correctly de-emphasized as the SNR decreases. The phenomenon matches with the recognition result that the first sub-band performs much worse than the second sub-band as SNR decreases.

[3]Notice that our recognition accuracies on clean connected TIDIGITS are lower than other reported results by about 1%. However, during our signal analysis, speeches are bandpassed to a bandwidth of 0–4000Hz while most reported results employ the full 10kHz bandwidth of TIDIGITS. Our system performance should better be compared with those based on Aurora since the setup is very similar except for the downsampling and filtering processes. Aurora's benchmark word accuracy on clean speech is 99.02% [12].

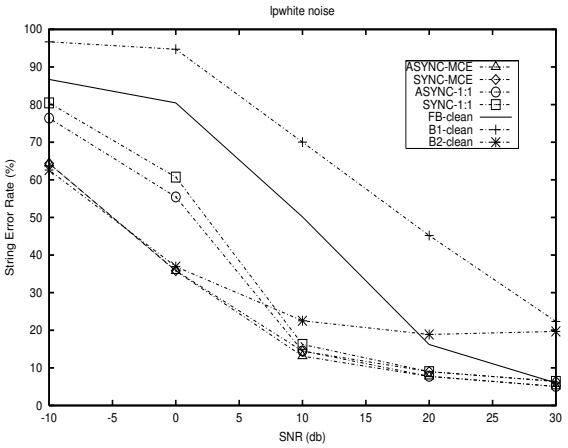

Figure 2: Recognition results with a low-pass white noise (string error rate).

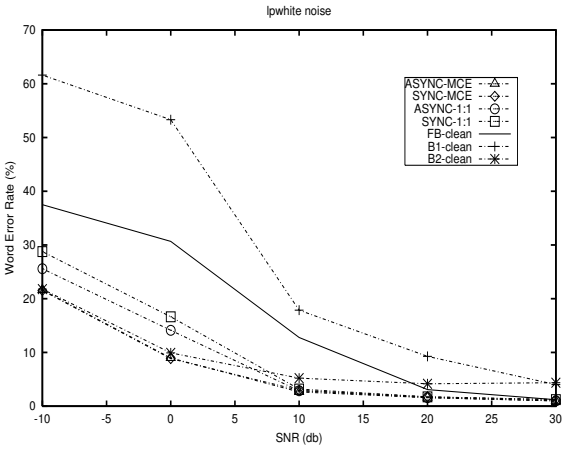

Figure 3: Recognition results with a low-pass white noise (word error rate).

4.2.2. Case 2: Additive White noise

In reality, noises often spread over a wider spectrum. So we further investigated the effectiveness of the MCE training algorithm of sub-band weights using white noise.

Recognition results are shown in Figure 4 and Figure 5. Once again, the asynchronous multi-band recognizer with MCE-derived weights outperforms the others. Compared with synchronous multi-band recognizer with equal weighting scheme, relative string (word) error rates are reduced by 12.8% (31.9%) at an SNR of 0db. Compared with the full-band recognizer, relative string (word) error rates are reduced by 6.9% (33.1%) at the same SNR.

5. Discussion

From the results of Experiments I and II, we further have the following observations:

- When the performance of the two sub-band recognizers

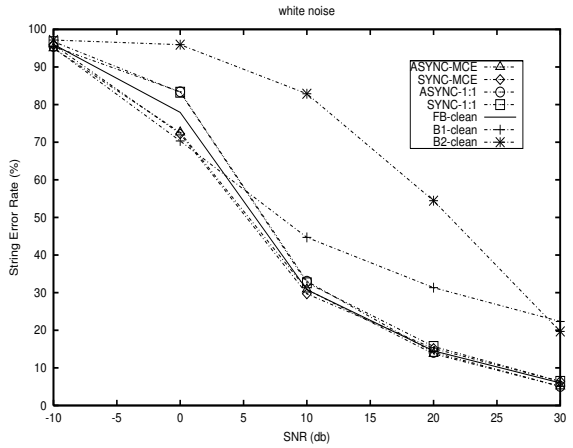

Figure 4: Recognition results with a white noise (string error rate).

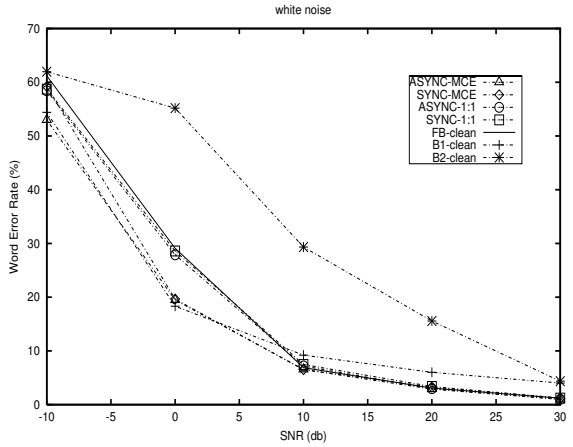

Figure 5: Recognition results with a white noise (word error rate).

diverges, MCE-derived weights effectively shift to emphasize the more reliable band; and the result is at least as good as that of the best sub-band recognizer.

- One may wonder why there is a huge difference between the two sub-band recognizers under white noise which has a uniform spectrum. It is probably due to the spectral tilt in speech and speech energy is decreasing at about 6dB per octave after about 1kHz — the dividing frequency between our two sub-bands. Consequently, the spectrally uniform white noise actually hurts the second sub-band more than the first sub-band, and thus performance of the second sub-band recognizer degrades more as well.
- With the sub-band asynchrony *and* MCE-derived sub-band weights, our multi-band system attain the best performance.

6. Conclusion

We have developed an asynchronous multi-band system using the HMM composition framework for continuous speech recognition. Linear sub-band weights are optimized by string MCE training using N-best hypotheses. Figure 6 summarizes the contribution of the asynchrony effect and the sub-band weighting in our system under low-pass white noise of 0db. The asynchrony effect and MCE-derived sub-band weights help improve performance of our multi-band system.

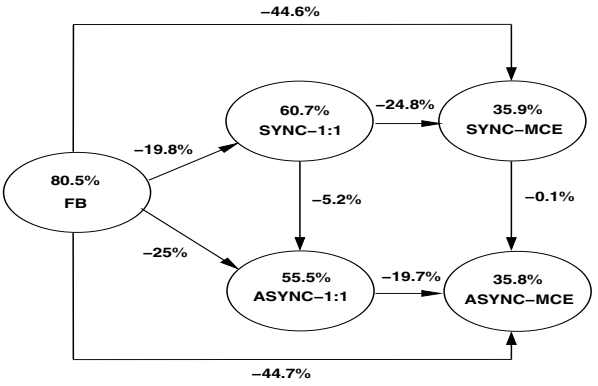

Figure 6: System improvement on low-pass white noise of 0db (absolute string error rate).

7. 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.

8. References

[1] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards ASR on Partially Corrupted Speech," in *Proc. of ICSLP*, Oct 1996.

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

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

[4] C. Cerisara, D. Fohr, and J. P. Haton, "Asynchrony in multi-band speech recognition," in *Proc. of ICASSP*, 2000.

[5] S. Matsuda, M. Nakai, H. Shimodaira, and S. Sagayama, "Asynchronous-Transition HMM," in *Proc. of ICASSP*, 2000.

[6] H. J. Nock and S. J. Young, "Loosely Coupled HMMS for ASR," in *Proc. of ICSLP*, 2000, vol. 3, pp. 143–146.

[7] D. Fohr, K. Daoudi and C. Antoine, "A New Approach for Multi-band Speech Recognition Based on Probabilistic Graphical Models," in *Proc. of ICSLP*, 2000.

[8] N. Mirghafori and N. Morgan, "Sooner or Later: Exploring Asynchrony in Multi-Band Speech Recognition," in *Proc. of Eurospeech*, 1999.

[9] B. Mak and Y. C. Tam, "Asynchrony with Re-Trained Transition Probabilities Improves Performance in Multi-Band Speech Recognition," in *Proc. of ICSLP*, 2000.

[10] C. Cerisara, J. F. Mari J. P. Haton, and D. Fohr, "A Recombination Model for Multi-band Speech Recognition," in *Proc. of ICASSP*, May 1999, vol. II, pp. 717–720.

[11] Y. C. Tam and B. Mak, "Optimization of Sub-Band Weights Using Simulated Noisy Speech in Multi-Band Speech Recognition," in *Proc. of ICSLP*, 2000.

[12] H.G. Hirsch and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," in *ISCA ITRW ASR2000*, 2000.