COMBINING CONNECTIONIST MULTI-BAND AND FULL-BAND PROBABILITY STREAMS FOR SPEECH RECOGNITION OF NATURAL NUMBERS

Nikki Mirghafori and Nelson Morgan

International Computer Science Institute
University of California at Berkeley
{nikki, morgan}@icsi.berkeley.edu

ABSTRACT

Multi-band automatic speech recognition is a new and exploratory area of speech recognition which has been getting much attention in the research community. It has been shown that multi-band ASR reduces word error in noisy conditions, particularly in the case of narrow band noise.

In this work we show that multi-band ASR could be used to improve the speech recognition accuracy of natural numbers for clean speech when the multi-band (MB) information stream is used in addition to the full-band (FB) one. We also observe that a similar combination method significantly reduces the error rate on reverberant speech. Finally, we analyze the error patterns of the full-band and multi-band paradigms to understand why the combination of the two streams is effective.

1. INTRODUCTION

There has been much interest generated in the speech recognition community on multi-band automatic speech recognition (ASR) [2, 11, 12, 8] since Jont Allen's cogent retelling of Harvey Fletcher's work on the articulation index [4, 1]. The main idea of this approach is to divide the signal into separate spectral bands (see Figure 1), process each independently (typically by training a multi-layer perceptron (MLP) and generating state probabilities or likelihoods for each sub-band), and then merge the information streams (for example, on a frame by frame level using another merger MLP). Some motivations for the multi-band paradigm are signal processing advantages, psycho-acoustic studies, robustness to noise, and taking advantage of parallel processing architectures.

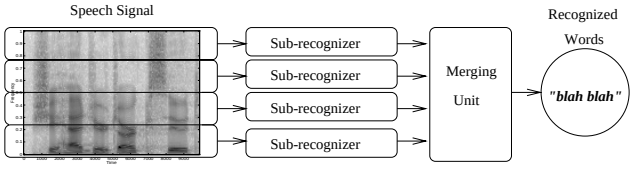

Figure 1: A simple overview of the multi-band system.

It has been shown that multi-band ASR performance is similar to that of the full-band paradigm on continuous speech, and significantly better in many noise conditions (especially for narrow band noise), when the multi-band streams are combined on a frame level [11, 2]. It is less clear if the multi-band paradigm can be used to significantly improve recognition accuracy for clean speech. Furthermore, if we can indeed confirm such improvements, we would like to understand the reason for this effect. Finally, we would like to evaluate the multi-band paradigm in a reverberant condition.

In the next section, we describe our experimental setup. In sections 3 and 4, we discuss our experiments with clean and reverberant speech, respectively. In section 5, we report on the results of our error analysis. Discussion and conclusions are in section 6.

2. DATABASE & SYSTEM DESCRIPTION

We use the Oregon Graduate Institute NUMBERS95 database, which comprises continuous digits and numbers recorded over the telephone as a part of census data collection. The database is phonetically hand-transcribed. For the purposes of this study, we use approximately two hours of the database for training and cross validation, and forty minutes as a test set.

Our baseline full-band system is an HMM/MLP based [3] system. We train the MLP phonetic probability estimator on a nine-frame window of 8th-order RASTA-PLP cepstra [5], energy, and delta-RASTA-PLP cepstral features over a 25 ms window, stepped every 10 ms. The MLP is fully connected and has 153 inputs (9 frames with 17 features per frame), 1000 hidden units, and 56 outputs (one output for each phone[1]), and is trained using back-propagation with softmax normalization at the output layer. The system is trained on hand-transcribed phone labels (without embedded realignment). We use a multiple pronunciation lexicon (derived from the hand transcriptions), a bigram language model, and a synchronous-time decoder called Y0 (described in [10]), which uses a single density per phone with repeated states for a simple durational model. The word error rate (WER) of this base-

[1]Note that some of the 56 phones do not occur in the NUMBERS database and have zero priors.

Percent Frame and Word Error for Clean Numbers							
Error	b1	b2	b3	b4	MB	FB	Mgd
Frame	40.2	37.4	42.5	49.5	22.4	23.7	-
Word	33.7	24.9	34.4	47.8	8.3	7.9	6.3

Table 1: Frame and word error, in percent, for subbands 1 through 4 (b1 through b4), multi-band (MB), full-band (FB), and the merged (Mgd) systems for clean natural numbers.

line system on the test set is 7.9%.

For our multi-band system, we divide the frequency range into four bands of [300-800 Hz][2], [700-1600 Hz], [1500-2700 Hz], and [2100-3800 Hz]. From the sub-bands, we derive [3rd, 3rd, 2nd, 2nd] order RASTA-PLP cepstral features, respectively, as well as energy and corresponding deltas. We train four MLPs on these acoustic features, that is, one on each sub-band. The input layer to each MLP has a context window of nine frames, for total input layer sizes of [72, 72, 54, 54] respectively. We choose hidden layer sizes of [497, 497, 372, 372], respectively, so that the total number of parameters in the four MLPs and the full-band system are roughly equal. There are 56 output units, one for every phone, as in the full-band MLP. The frame-by-frame information from the four sub-band streams is combined using a *merger* MLP, which takes the output of the sub-band MLPs as input, has 300 hidden units, and an output of 56 phones. Table 1 includes the frame and word errors for each subband, the full-band, and the multi-band systems. The frame error (on the cross validation set) of the four subband systems range between 37.4% and 49.5%, whereas the merged multi-band system has a reduced frame error of 22.4%, which compares favorably to that of the full-band system (23.7%). The word errors follow a similar trend. The word errors of the four subband systems range from 24.9% to 47.8%, and the merged multi-band system has a word error of 8.3% which is statistically not different from a word error of 7.9% by the full-band system.

3. EXPERIMENTS WITH CLEAN SPEECH

As we discussed in the previous section, the word error rate of the multi-band and the full-band system are similar. The question remains whether multi-band information can be used for improving the ASR performance.

We merged the probability streams by simply multiplying the likelihoods from each system, before feeding the probability stream into the decoder. The word error rate of the combined system decreased to 6.3%. In other words, errors were reduced by 20%.

We note that the combined system has roughly twice as many parameters as the other systems, so it is possible that doubling the number of parameters in the full-band system might produce a similar improvement. We trained a full-band system with twice as many parameters and its WER was 8.9%. It appears likely that the improvement in the combined system was not merely due to an increase in the number of parameters.

[2]Because we are testing on telephone quality speech, we disregard frequencies from 0 through 300 Hz.

Percent Word Error for Reverberant Numbers						
CW	b1	b2	b3	b4	MB	Mgd
9	68.1	61.2	68.7	76.2	39.9	30.3
17-11	66.2	60.5	67.9	75.9	38.2	29.5
17	65.7	59.0	67.4	75.7	42.8	31.6

Table 2: Percent frame error for bands 1 through 4, multi-band (MB), and merged (Mgd) systems for reverberant natural numbers for different sizes of feature-input context-windows (CW). The baseline FB system has a word error rate of 32.2%.

Thus, it appears that combining multi-band and full-band systems significantly[3] reduces the word error rate for our test set over either system alone, or a version of the full-band system with an extended parameter set.

In Section 5, we analyze the error patterns of the two streams to understand how they might counteract each other.

4. EXPERIMENTS WITH REVERBERANT SPEECH

We performed similar experiments on digitally-reverberated versions of the data. 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[4].

Natural reverberation usually affects low frequencies more than high frequencies, since most common room boundary materials are less absorptive at low frequencies, leading to longer reverberation times and more smearing of the spectral information at those frequencies. Our baseline system system has a feature input window of nine (four frames of context in the past and the future) for all frequency bands. We decided to increase the size of the feature input window for the low frequency subbands. More specifically, we decided to double the input window size for the lowest band, which would make it roughly equal to the length of a syllable (200 ms). We decreased the size of the neighboring higher frequency windows by two frames, therefore, the "pyramid" system has 17, 15, 13, and 11 frames of input for bands one through four. To be aware of the effects of overall window size increase, we also trained four subband systems with 17 frames of input each. The

[3]For this size test set, an absolute difference of more than 1.1% is considered statistically significant (using z-scores on binomial distributions).

[4]Although this ratio might suggest a seriously degraded signal, recent listening tests showed essentially no reduction in intelligibility with respect to tests using the clean signal [7].

WER for each subband, the multi-band, and the merged systems are reported in Table 2. The WER for the full-band system is 32.2%, which is significantly better than each of the multi-band systems (38.2% – 42.7%). However, merging the inferior multi-band stream with the full-band stream still *improves* the overall WER (29.5% – 31.6%). We also observe that although the WER of each 17-context-frame subband system was less than that of the pyramid system, this was not true when the subbands were merged together, and again, with the full-band system. In short, adding the multi-band pyramid system information to that of the full-band system reduces the WER from 32.2% to 29.5%. This is an error reduction of 8%.

For the sake of completeness, we also ran the increased window size experiments for clean speech. Neither of the conditions significantly changed the WER from the baseline setup; for instance, WER for the pyramid windows was 6.5%, in comparison with the 6.3% for the 9-frame window. Thus, it appears that using the extended windows, particularly the pyramid case, improves WER for reverberant speech without substantially hurting performance for clean speech.

5. ANALYSIS OF MULTI-BAND AND FULL-BAND ERROR PATTERNS

In addition to simply observing a reduction in WER, it is also important to at least try to understand why such reduction occurs. One explanation, inspired by the expert-merging community, is that the error rate decreases when two different experts with different characteristics (preferably orthogonal) are combined [6]. We want to understand how our full-band and multi-band recognizers are different, and how this difference affects performance [13]. If possible, we would also like to associate these differences with phonetic content: are there particular phones or features that one system is better at discriminating than the other?

	t	s	eh	sil	...
t	5722	252	31	316	...
s	258	8495	110	1159	...
eh	11	93	3118	37	...
sil	436	2733	68	40237	...
...	...	...	...	...	...

Table 3: An example of a phone-based confusion matrix.

We performed phone recognition on both the full-band and the multi-band systems. We generated confusion matrices for phone classes, both for the phone recognition results and for the frame by frame comparison of the phone decoding path. The main difference is that the latter gives more weight to long phones, since the classification for every frame is counted. A confusion matrix (CM) is simply an extended matrix of *hits* and *misses* for all classes, as in Table 3. The column headings represent the classes we intend to *transmit*, and the row headings correspond to the *received* classes. In Table 3, for example, 93 instances of /s/ are received as /eh/. We use frame level phonetic classification on the test set for generating phone CMs. We also generated detailed statistics for every phone token: whether both systems were right or wrong, and how this affected the merged system's classification. The summarized results of this analysis are in Table 4.

		Mgd $\surd$	Mgd$\times$
FB	MB $\surd$	86.8	0.2
$\surd$	MB $\times$	2.2	1.3
FB	MB $\surd$	1.9	1.2
$\times$	MB $\times$	0.3	6.1

Table 4: A summary of the analysis on the recognized phone string for the full-band (FB), multi-band (MB), and the merged (Mgd) system as compared to the correct results. $\surd$ means the phone classification of that band was correct, $\times$ means that the phone classification was incorrect.

We observe the following:

- It rarely occurs that the classification of the merged stream is incorrect when both full-band (FB) and multi-band (MB) streams have the correct phone classification (only for 0.2% of the phone tokens). Conversely, it is also unusual for the co-occurring errors of the two streams to be corrected by merging (only for 0.3% of the phone tokens).
- Nearly all of the the correctly classified phones in the merged stream were actually correct in both streams (95.1% of the correctly classified tokens). Of the remainder, which were correct in one stream only, roughly half were correct in each stream (2.5% and 2.1% for MB and FB respectively).
- Most of the phones that were incorrectly classified in the merged stream were incorrect in both steams (69.6% of the incorrectly classified tokens). Of the remainder, which were incorrect in one stream only, roughly half were incorrect in each stream (13.9% and 14.3% for MB and FB respectively).

Not shown in the table:

- Most of the MB and FB phone errors are identical (76% of the misclassified tokens).
- Examining the errors for each phone class, we see that for /sil/ and /tcl/ (t-closure) the MB system is correct significantly more often than the FB system. The reverse is true for the vowel /ao/.
- As we examine the frame-based confusion matrices, we observe that the MB system is significantly more accurate in classifying /sil/, /r/, /w/, and /tcl/ phones. The FB system, on the other hand, is significantly more accurate in classifying /ao/, /n/, /iy/, /ah/, /f/, and /s/. Research on the acoustic

cues for the perception of liquids and glides has shown that the duration of the formant transitions provides the essential cue for these speech sounds [9]. For discrimination of vowels, however, simultaneous identification of the location of the first two formants is necessary. Perhaps the divide and conquer MB strategy makes it difficult for a fine across subband information analysis necessary for accurate discrimination of vowels, whereas the transition pattern becomes more apparent, explaining better liquid and glide discrimination.

6. DISCUSSIONS AND CONCLUSIONS

In this work we have shown that multi-band ASR could be used to improve the speech recognition accuracy of natural numbers for clean speech when a multi-band information stream is used in addition to the full-band one. Specifically, this combination reduced the word error rate by 20%. We observed that a similar combination method significantly reduced the error rate on reverberant speech. We also saw that extending the input window to our neural network probability estimators, particularly for the low frequency bands, improved recognition for reverberant speech without substantially changing the performance for the clean case.

Additionally, we analyzed the error patterns of the full-band and multi-band paradigms to understand why the combination of the two streams is effective. It appears that in most cases, both systems either classify the phone either correctly or incorrectly. However, in many instances, one system is correct while the other is wrong. In 62% of these instances the correct classification prevails. Finally, about 5% of the instances when both systems are incorrect, the merged system (miraculously!) performs the classification correctly; whereas, in the 0.2% of the instances where both systems' classification is correct, the merged system guesses the wrong phone.

Besides the overall advantages, we also observed that the MB and FB system are different in their level of accuracy for various phone classes. Most notably, the MB system is inferior to the FB system in classifying some fricatives and vowels, while the MB system excels in classifying the silence and some liquids and glides.

ACKNOWLEDGMENT

We would like to thank Brian Kingsbury for sharing the effort in developing the analysis scripts, Eric Fosler-Lussier and Su-Lin Wu for helpful discussions on lexicon creation and decoding, and Dan Ellis for proof-reading the paper. We acknowledge our colleagues Hervé Bourlard and Hynek Hermansky for our long-distance multi-band collaboration. We thank Jim West and Gary Elko, from Bell Labs, and Carlos Avendano, at the University of California, Davis, for collecting the room impulse responses and making them available to us. This work was supported by a European Community Basic Research grant subcontract from Faculté Polytechnique de Mons in Belgium (Project Sprach), an IDEA grant from the U.S. Department of Defense, and the International Computer Science Institute.

7. REFERENCES

1. J. B. Allen. How do humans process and recognize speech? *IEEE Trans. on Speech and Audio Proc.*, 2(4):567–577, Oct. 1994.
2. H. Bourlard and S. Dupont. Subband-based speech recognition. In *ICASSP*, volume 2, pages 125–128, May 1997.
3. H. Bourlard and N. Morgan. *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic Press, 1994.
4. H. Fletcher. *Speech and Hearing in Communication*. Krieger, New York, 1953.
5. H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, Oct. 1994.
6. R. A. Jacobs. Methods for combining experts' probability assessments. *Neural Computation*, 7(5):867–888, Sept. 1995.
7. B. E. D. Kingsbury. *Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments*. PhD thesis, University of California, Berkeley, California, 1998. To appear.
8. N. Mirghafori and N. 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.
9. J. D. O'Connor, L. J. Gerstman, A. M. Liberman, P. C. Delattre, and F. Cooper. Acoustic cues for the perception of initial /w,y,r,l/ in english. *Word*, 13:24–43, 1957.
10. T. Robinson, L. Almeida, J. Boite, H. Bourlard, F. Fallside, M. Hochberg, D. Kershaw, P. Kohn, Y. Konig, N. Morgan, J. Neto, S. Renals, M. Saerens, and C. Wooters. A neural network based, speaker independent, large vocabulary, continuous speech recognition system: The WERNICKE project. In *EUROSPEECH*, pages 1941–1944, Berlin, Germany, Sept. 1993.
11. S. Tibrewala and H. Hermansky. Sub-band based recognition of noisy speech. In *ICASSP*, volume 2, pages 1255–1258, May 1997.
12. 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, Apr. 1997.
13. S.-L. Wu, B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. In *ICASSP*, pages 721–724, Seattle, WA, May 1998.