Submitted to Eurospeech'99, Budapest

MULTI-STREAM SPEECH RECOGNITION: READY FOR PRIME TIME?

Adam Janin, Dan Ellis, Nelson Morgan

International Computer Science Institute, 1947 Center St, Berkeley, CA 94704
Email: {janin,dpwe,morgan}@icsi.berkeley.edu

ABSTRACT

Multi-stream and multi-band methods can improve the accuracy of speech recognition systems without overly increasing the complexity. However, they cannot be applied blindly. In this paper, we review our experience applying multi-stream and multi-band methods to the Broadcast News corpus. We found that multi-stream systems using different acoustic front-ends provide a significant improvement over single stream systems. However, despite the fact that they have been successful on smaller tasks, we have not yet been able to show any gain using multi-band methods. We report various insights gained from the experience in applying these methods in a large-vocabulary task.

1. INTRODUCTION

Previously, we and others have shown that, for a number of smaller tasks, the merging of multiple probability streams derived from different acoustic representations improves speech recognition across a range of acoustic conditions [1]. In collaboration with our colleagues at Cambridge and Sheffield, we developed a multi-stream system used in the 1998 DARPA Broadcast News speech transcription evaluation. The complete system was quite complex and is described in [2]. Here, we will describe the results of some experiments at the International Computer Science Institute that show the utility of multi-stream approaches for this task.

We will begin in section 2 with a brief overview of our hybrid connectionist/hidden Markov model (HMM) system, including comments on factors that confound the analysis. Next, in section 3, we discuss multi-stream systems based on multiple acoustic front-ends. Then, in section 4, we deal with multi-band systems, in which the feature streams are derived from distinct spectral regions. Finally, some conclusions are given in section 5.

2. SYSTEM OVERVIEW

The hybrid connectionist/HMM speech recognition framework uses neural networks to estimate the posterior probabilities for each of the 50 or so phone classes, based on an acoustic feature vector observed over some temporal window. The posteriors are converted to likelihoods and used in a conventional hidden Markov model decoder to find the best matching word string hypothesis. The neural network acoustic classifiers are trained via back-propagation against "one-hot" phoneme targets derived from a forced-alignment of training set word transcriptions. Thus, the networks are trained to optimize frame-classification accuracy (choosing the "right" label for each training pattern), whereas the resulting speech recognition systems are assessed by overall word error rate. We note in passing that the relationship between performance at the frame and word levels is certainly not monotonic. We train our networks with cross-validation and early stopping to avoid over-fitting.

System	ALL	F0	non-F0
RNN/PLP	24.5%	15.8%	28.3%
MLP/MSG	27.0%	18.5%	30.7%
Combination	23.0%	15.7%	26.2%

Table 1: 1998 Broadcast News evaluation word-error rates for two neural net acoustic models and their combination. Most of the benefit from the MLP/MSG system is in the non-F0 conditions (spontaneous and/or degraded speech).

The results described in this paper arose from our work towards submitting a hybrid system for the 1998 DARPA/NIST "Broadcast News" evaluation, which was a collaboration with colleagues at Cambridge and Sheffield Universities through the European Union SPRACH project. One enabling factor of this geographically dispersed effort was the use of several largely independent acoustic models, as described in [2]. To the core Cambridge system, based on PLP features and a recurrent neural network (RNN) estimator, we added an 8000 hidden node multi-layer perceptron (MLP) neural network based on the novel modulation-filtered spectrogram (MSG) features, and combined by the simple process of averaging the log posterior probabilities for each phone class. This average log-probability method has consistently outperformed more principled methods. Since our previous experience had shown that the greatest benefits come from combinations between the most diverse features [1], we also halved the bandwidth of the acoustic signals to 4 kHz in an effort to limit the difference between telephone and fullband signals, both of which occur in the database. As illustrated in table 1, the resulting combination was indeed beneficial, with most of the benefit coming from the more acoustically-challenging parts of the corpus (i.e. sections other than the studio-quality, prepared speech referred to by the tag "F0").

For the evaluation system and in subsequent work, we have produced numerous system variants pursuing the basic idea of improving performance through the combination of multiple information streams. These are the focus of this paper.

It is worth noting some caveats in the interpretation of our results and some causes for their variation. All the remaining word error results are obtained on a 30 minute subset of the 1997 Broadcast News evaluation, containing a mix of different acoustic conditions. This set contains 5938 words, and a simple binomial confidence test requires a 1.3% absolute difference in word-error rate for significance at the 5% level. Through the course of the project we successively refined our training target labels, and we also worked with several HMM decoders and pruning strategies, trading error rate for processing time. All these factors contribute difficulties in comparing results other than within each table, for which conditions have been controlled.

3. MULTI-STREAM

Given the success of model combinations in our Broadcast News evaluation system, as illustrated in table 1, we have conducted further experiments to control for some of the confounding factors in that system. Specifically, we replaced the RNN baseline system with an MLP network trained on PLP features (removing the difference in network architectures), and we trained MSG networks based on the full 8 kHz-bandwidth audio data (removing the difference in audio bandwidth). This section reports on these combinations.

For a fair comparison, we trained a set of MLPs using identical targets and training data (a 70 hour subset of Broadcast News). We controlled the number of parameters in the networks: each net contained either 344,000 or 758,000 parameters.

One possible explanation for our observed improvements using multi-band systems is that combining *any* set of predictors (experts) tends to lead to better results. In order mitigate this effect, we trained two nets using identical data, but different random starting points in the neural network training. Each of these nets used 12th order PLP features plus energy (with mean and variance normalized in each segment), a temporal context window of 9 frames (for a total of 117 input units), and 2000 hidden units. These nets are referred to below as PLP2000a and PLP2000b.

We also trained a net using MSG features. The particular form of this feature we used consisted of two banks of roughly log-spaced spectral channels. The first bank is filtered to pass modulation frequencies in the range 0–16 Hz, and the second bank passes 2–16 Hz, approximating the time-differential of the first bank. For 8 kHz bandwidth audio, each bank has 18 channels for a total of 36 elements per feature vector, in contrast to the 13 element PLP feature vector. It was therefore necessary to reduce the number of hidden nodes in the MSG net to 907 in order to keep the number of parameters the same as the PLP2000 nets. This net is referred to as MSG907.

Equalizing the number of parameters by reducing the number of hidden nodes is perhaps not a fair comparison, since the number of weights is only an upper bound on the effective number of degrees of freedom of the model; weights may be more or less useful in different parts of the network, and what we really want to equalize is the learning ‘capacity’ of the networks. We could have included deltas and double-deltas in the PLP features in order to make the feature dimensionality roughly equivalent, but in earlier experiments we found that PLP deltas did not help significantly with this situation, and we would therefore have been increasing the PLP model parameters without significantly improving its capacity to identify useful information. On the assumption that capacity may be governed by the size of the intermediate representation constituted by the hidden layer, we trained an MSG net with 2000 hidden nodes, called MSG2000, and for comparison a PLP net with 4407 hidden units, called PLP4407. Each of these nets contain 758,000 parameters. Note that for completeness, we might wish to test duplicates of all these nets using different initial starting points. However, resource limits precluded testing every possibility.

Features	# weights	WER
MSG907	344k	34.2%
MSG2000	758k	31.4%
PLP2000b	344k	29.1%
PLP2000b - MSG907	344k+344k	28.5%
PLP2000a	344k	28.5%
PLP2000a - PLP2000b	344k+344k	28.5%
PLP2000a - MSG907	344k+344k	28.0%
PLP2000a - PLP4407	344k+758k	27.5%
PLP2000a - MSG2000	344k+758k	27.1%
PLP4407	758k	26.7%
PLP4407 - MSG2000	758k+758k	26.4%

Table 2: Word-error rate results for various nets trained to match the number of weights or the number of hidden units between the different feature streams, as well as their principal combinations.

The results of these tests are summarized in table 2. A hyphen in the table indicates that the probability streams were combined using the average log-probability method.

3.1. Multi-stream discussion

The first trend that is apparent in table 2 is that “bigger is better.” Although we have found that increasing the size of the nets improves accuracy over a very wide range of conditions [3], we do not expect to be able to enlarge the nets indefinitely, and for smaller training sets we eventually saw asymptotic performance as the nets grew larger. However, for the full 142 hour Broadcast News training set, we hit another limit before we reached excessively diminished returns — resources. Not only do the larger training runs take exorbitant amounts of time, but we are unable to increase the size of the nets beyond a certain limit because of memory constraints on our special purpose training hardware.

Multi-stream methods can ameliorate the resource limitations. Given a finite training time and multiple machines, independent systems can be trained in parallel then used in combination (parallelism can also be employed during recognition). Similarly, although hardware limits may prevent the doubling of the size of a single model, much of the benefit can be obtained by training two models (based on different features) and combining. Finally, we expect that the fact that these models are based on different information in the training data will mean that even when individual networks have been expanded to the point of fully exploiting the training sets, there will still be a gain from combining them, which could not be obtained by other means.

Table 2 provides a guide for feature selection. If the limit is the size of the network (e.g. the number of parameters in one net), it is better to combine nets using different features, rather than repeating the same feature. This is true even if the second feature does fairly poorly on its own compared to the first feature. If the limit is *sequential* training time (e.g. training time on one computer), table 2 suggests that it is better to train using only the best feature, since training time is roughly proportional to the total number of parameters, and a linear interpolation between PLP2000a and PLP4407 predicts that a single PLP net with 688k parameters would have a WER of around 27.0%, better than the PLP2000a-MSG907 combination which has the same number of parameters.

4. MULTI-BAND

Rather than deriving the probability streams from completely different acoustic representations, it is also possible to divide a single representation into disjoint regions across the spectrum. Each of the subbands can then be used as the basis for a separate probability estimator. The output of these estimators can be combined either by averaging the log posterior probabilities for each phone class, as above, or using more complex methods including multi-layer perceptrons, weighted combinations, etc.

For a number of corpora, the multi-band approach has shown significant improvement over a fullband system, especially when used in combination with a fullband probability estimator [4]. However, for Broadcast News, we have not observed any significant improvements over a wide range of experiments. In the following sections, we will briefly present these experiments, as well as some comments on the possible obstacles to success.

4.1. MSG Multi-band

Since our intention was to develop additional acoustic models that could profitably be combined with the PLP-based RNN baseline, and in view of our experience of the value of diverse feature bases, we began our study of multi-band with systems based on MSG features (described in section 2 above). MSG features for 4 kHz bandwidth data consist of two banks of 14 features each. For the subband experiments, we divided each bank into 4 subbands, and concatenated one subband from each bank to form the subband features. Previous results on other corpora indicated that the lower frequency bands contain more information, and therefore should incorporate more features. The final division consisted 10 features in band A, 8 features in band B, and 6 features each in bands C and D.

Four MLPs were trained, one for each subband. The inputs to the MLP included the features as described above over a context window of 9 frames (i.e. the MLP for Band A used $10 \times 9 = 90$ inputs). For each case, 2000 hidden units were used. Note that since the number of inputs to the MLPs is different, the constant-size hidden layer implies that the subband MLPs have different numbers of parameters. An alternative approach would have been to hold the total number of parameters constant, but our previous discussion of the difference between parameter count and true model "capacity" applies.

Numerical results from the MSG subband experiments are summarized in table 3 in section 4.3. In this section, we will discuss the qualitative results.

The first and most striking result was that when an individual subband probability estimator was used on its own, it was so equivocal in its probability estimation that the HMM decoder was unable to decode some utterances — the decoder ran out of space for hypotheses on a machine with 1GB of memory! This was in contrast to our experience on other tasks, where individual subbands would perform far below the fullband, but still at a measurable level. The simple form of combination used in our multi-stream experiments (average log-probability) produced a the system that would decode, but gave results highly inferior to those of the fullband system. Combining the subband systems with the fullband systems (either the RNN or fullband MSG) also worsened the word-error over the fullband system alone.

Another method of merging the results of separate classifiers is to feed the posteriors directly into a multi-layer perceptron trained on the "correct" posteriors. This is tractable if the number of streams to be combined is reasonably small. The output posteriors of the 4 subband estimators were fed into a MLP with $54 \times 4 = 216$ inputs, 2000 hidden units, and the normal 54 outputs. The MLP was trained on the same data as the subbands (disjoint data would have been preferable, but were not available). The resulting estimator was significantly better than the simple combination scheme. However, when combined with a fullband system, it produced only an insignificant improvement over the fullband system alone. When combined with the baseline RNN system, the gain disappeared entirely.

Several variants of the systems described above were also investigated, including cepstral transformations, Karhunen-Loeve transformations, principal axes dimensionality reduction, and entropy weighted combinations. Although some of these procedures helped the subband systems by a small amount, none of the gains carried over to the combination of the subband plus fullband systems.

4.2. PLP Multi-band

In view of these disappointing results, we decided to try to replicate the success of multi-stream on the Numbers95 corpus as reported in [4]. We used the same ideas as described in the MSG multi-band experiments above, but based the features on PLP rather than MSG. The system consisted of cepstral transformed 12th order PLP features plus energy and deltas. Bands A and B used 4th order cepstral features, while band C and D used 3rd order. This leads to subbands A and B having 8 features each, and subbands C and D having 6 features each. Four separate MLPs were trained, one for each subband. Each MLP used a context window of length 9 frames and had 2000 hidden units.

Again, the subbands by themselves failed to decode, and the simple method of combination showed very poor results. Combining with an MLP as described in the previous section showed marked improvement. In the best case, combination of the subband system with a fullband system produced a modest gain, but again, the gain disappeared when combined with the full-blown system including the RNN.

In addition to the methods described above, we also experimented with "all-wise" combinations of the PLP subband outputs. The basic idea of the all-wise combination method is to form a linear combination of all possible subsets of the 4 subbands. Picking values for the 2^4 weights is a key aspect of this algorithm. We tried equal weighting, weights based on the subset's frame-accuracy, and weights trained using an MLP. Although even the simple, equal weights method has been successful on other tasks, we were unable to show any gain on Broadcast News over combining the subbands with an MLP. However, weights picked by an oracle based on the current alignment produced excellent results (18.3% word error), indicating that, with the right dynamic assignment of the weights, the all-wise combination method could provide additional information. Clearly, this method warrants additional investigation.

4.3. Multi-band Results Summary

The results of the subband experiments are summarized in tables 3 and 4. Models separated with "-" are combined using average log-probability. Models separated with "+" are combined using

Features (see section 4.3)	Word Error
A	Failed to decode
A-B-C-D	64.9%
A-B-C-D - MSGx4	40.5%
A+B+C+D	48.6%
MSG	38.7%
(A+B+C+D) - MSGx4	38.5%

Table 3: MSG subband results. See text for explanation.

Features (see section 4.3)	Word Error
A	Failed to decode
A-B-C-D	61.7%
A+B+C+D	44.8%
MSG	37.8%
(A+B+C+D) - MSG	35.7%
RNN	33.2%
MSG - RNN	29.9%
(A+B+C+D) - MSGx2 - RNNx4	29.9%

Table 4: As table 3, but for PLP-based subbands.

an MLP. Subbands are denoted with the capital letters A, B, C, and D. The MSG fullband is denoted "MSG", while the recurrent neural network based on fullband PLP is denoted "RNN". If a feature is followed by xN, it indicates that the feature was weighted by a factor of N in the log domain. Note that only a small, representative subset of the experiments are reported; for example, the linear multiples xN reported are the best ones we tried.

Note that the numbers from table 3 and table 4 are not directly comparable — table 4 represents a later experiment using a different alignment.

4.4. Multi-band discussion

The story of multi-band is not yet finished. Although we were unable to show improvement with this particular corpus, we were able to gain some insights.

First, combination methods must be handled with care. Although the simple average log-probability method worked better than any other method we tried for the large fullband systems, it worked very poorly for the multi-band systems. Combining the subbands using an MLP gave roughly a 30% relative improvement over the average log-probability.

Second, and somewhat contrary to the results on multi-stream PLP/MSG, combining the low accuracy multi-band net with a relatively higher accuracy fullband estimator seldom improved performance. In our case, any gain with the multi-band system was erased when it was combined with our fullband system.

The fact that a single subband estimator on its own fails to decode, and that even in combination, the subband estimators produced poor results, seem to indicate that one must start with a fairly accurate estimator before this type of combination can be useful. The most obvious difference between the Numbers95 corpus (where multi-band was successful) and the current Broadcast News experiments is that the baseline systems in Numbers95 have word-error rates of 5–8%, compared to 20–30% for Broadcast News systems. It seems plausible that a certain minimum threshold of subband performance is required to obtain gains through combination.

5. CONCLUSIONS

The combination of multiple sources of information has clear theoretical attractions and many successful practical implementations. For the large-vocabulary, acoustically-complex Broadcast News task, we found that a multi-stream system combining separate acoustic models based on different underlying features was not only very beneficial when compared to using a much larger monolithic acoustic model, but also had several practical advantages in training and execution.

Dividing the original acoustic data into multiple frequency bands to constitute our different sources of information gives multi-band systems, but our experiments with these were less encouraging, with no overall improvement using this approach despite experiments on a large number of variants. Although it is not clear why our previous success with this approach in a small-vocabulary task did not translate to this new domain, it may be that restricting the available acoustic information for a task that is already very difficult leads to a classification problem that is just too hard to be recovered through subsequent combinations. We will, however, continue to investigate variants on this approach in the hope of finding a profitable exploitation.

6. ACKNOWLEDGMENTS

We are particularly indebted to Eric Fosler-Lussier, whose service was central to the ICSI Broadcast News effort, and to Brian Kingsbury, who provided MSG features, and to Gary Cook of Cambridge whose help was indispensable in getting us started with Broadcast News. This work was partially supported by the European Union through the LTR projects SPRACH (20077) and RESPITE (28149).

7. REFERENCES

[1] S. Wu, B. Kingsbury, N. Morgan and S. Greenberg, "Performance improvements through combining phone- and syllable-length information in automatic speech recognition," *Proc. ICSLP-98*, Sydney, 854-857.

[2] G. Cook, J. Christie, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, A. Robinson, and G. Williams "The SPRACH System for the Transcription of Broadcast News," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Herndon VA, 1999.

[3] D. Ellis and N. Morgan "Size matters: An empirical study of neural-network training for large-vocabulary continuous speech recognition," *Proc. ICASSP-99*, Phoenix AZ, II-1013-1016.

[4] N. Mirghafori and N. Morgan, "Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers," *Proc. ICSLP-98*, Sydney, 743-746.