INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 666-2900 • FAX (510) 666-2956

Stream combination before and/or after the acoustic model

Daniel P.W. Ellis

TR-00-007

April 2000

Abstract

Combining a number of diverse feature streams has proven to be a very flexible and beneficial technique in speech recognition. In the context of hybrid connectionist-HMM recognition, feature streams can be combined at several points. In this work, we compare two forms of combination: at the input to the acoustic model, by concatenating the feature streams into a single vector (feature combination or FC), and at the output of the acoustic model, by averaging the logs of the estimated posterior probabilities of each subword unit (posterior combination or PC). Based on four feature streams with varying degrees of mutual dependence, we find that the best combination strategy is a combination of feature and posterior combination, with streams that are more independent, as measured by an approximation to conditional mutual information, showing more benefit from posterior combination.

STREAM COMBINATION BEFORE AND/OR AFTER THE ACOUSTIC MODEL

Daniel P.W. Ellis

International Computer Science Institute, Berkeley, California, USA

ABSTRACT

Combining a number of diverse feature streams has proven to be a very flexible and beneficial technique in speech recognition. In the context of hybrid connectionist-HMM recognition, feature streams can be combined at several points. In this work, we compare two forms of combination: at the input to the acoustic model, by concatenating the feature streams into a single vector (feature combination or FC), and at the output of the acoustic model, by averaging the logs of the estimated posterior probabilities of each subword unit (posterior combination or PC). Based on four feature streams with varying degrees of mutual dependence, we find that the best combination strategy is a combination of feature and posterior combination, with streams that are more independent, as measured by an approximation to conditional mutual information, showing more benefit from posterior combination.

1. INTRODUCTION

As the first stage in any speech recognition system, the features are critical to the overall system performance. The ideal features reflect the 'important' information in the speech signal (e.g. the phonetic variation) in a consistent and well-distinguished fashion, while minimizing or eliminating 'irrelevant' information (such as speaker identity or background conditions). These goals are very difficult to achieve, and consequently a wide variety features has been proposed and employed, each with different strengths and weaknesses.

Stream combination is a technique which seeks to capitalize upon the practical differences between feature streams by using several at once. The basic argument is that if the recognition errors of systems using the individual streams occur at different points, there is at least a chance that a combined system will be able to correct some of these errors by reference to the other streams. An extreme example of this approach is the Rover system which combines final hypotheses of complete speech recognition systems, and which was able to show 30% relative error rate reductions over the best system in a recent NIST Broadcast News evaluation [1].

A range of other approaches was described in [2]. In this paper, we compare the two simplest of these. Concatenating the feature vectors from different extraction algorithms to create a single, higher-dimensional space for modeling is the default approach, here termed Feature Combination (FC), following the terminology of [3]. This is contrasted with combining the streams at the *outputs* of the acoustic models: In the hybrid connectionist-hidden Markov model speech recognition approach [4], the acoustic model is a neural network estimating posterior probabilities across a complete set of context-independent phones. These posteriors can be combined in several ways, but simple averaging of the log probabilities from the different estimators for each phone has consistently performed as well as or better than more complex schemes [5, 6, 7, 8]; we will call this Posterior Combination (PC). Posterior combination is close to the formally correct approach if the feature streams are conditionally independent given the phone, but its main support comes from empirical, not theoretical, considerations.

An interpretation of the success of PC is provided in [5]. When confronted with data outside its domain of expertise, each model may tend to emit relatively 'flat' (high entropy) posteriors, which will have a neutral impact on the relative probabilities of other distributions with which they are averaged. Thus, if one posterior estimator is relatively confident of the correct classification (low entropy), and the remainder are equivocal, the confident estimate will dominate. This, however, would be true of most if not all simple combination rules; the pre-eminence of log-domain averaging remains something of a mystery to us.

1.1. Feature combination versus probability combination

In previous experiments, PC has been shown to outperform FC for combining distinct feature streams. This can be explained by the following argument: Consider two feature streams with different properties and 'domains of expertise'. The training set will presumably contain a number of conditions in which each of the streams is providing useful information when the others are not. A classifier trained on the combined feature space of FC will learn about those specific cases represented in the training set, but will have difficulty generalizing to other situations where one stream is giving 'good' information but the other streams are in pathological conditions different from those represented in the training set. In PC, by contrast, the separate models learn the specific regions of 'good' information for each stream individually; at the point of combination, the fact that the different states of the feature streams may not have been observed in training is no longer relevant – the log probabilities can be averaged together regardless.

Thus the multiple, smaller feature spaces of PC achieve a factoring of the possible signal conditions by the specialties of each classifier (and can generalize to previously-unseen conjunctions of those factors), whereas the combined space of FC requires training on an enumeration of all possible factor combinations.

This advantage of PC must have limits, however, for if taken to extremes we might be tempted to subdivide the feature vector from a single stream between multiple classifiers. In fact, this is similar to the approach adopted in multi-band recognition [9], which, while advantageous in certain situations, is often inferior to full-band recognition on well-matched tests. In particular, as the information contained in each subband is reduced, leading to classifiers which are rarely able to make unequivocal decisions, the subsequent PC stage has more and more difficulty recovering the latent information distributed among the classifier outputs – even when they are combined by a second, trained classifier rather than by simple averaging [7]. We can imagine that if particular feature di-

mensions are significantly co-dependent given the particular phone class, it will be desirable to build a classifier that can model their joint distributions via FC. It is only when the feature streams are relatively independent that PC is the more appropriate choice.

This paper describes a set of experiments designed to investigate and verify these intuitions. In particular, we wanted to see if we could come up with a practical way to predict the relative merits of FC and PC for a pair of feature streams by looking at some measure of the statistical dependence between the streams. The next section describes the experimental setup we used, in terms of the different feature streams and the speech recognition task. Section 3 presents the results of these experiments, which we then discuss in the conclusion.

2. EXPERIMENTAL SETUP

2.1. Feature streams

To explore the differences between FC and PC, we experimented with four feature streams, organized as two relatively independent pairs of more closely related streams. The first pair was standard 12th-order PLP cepstral coefficients (first stream, "plp12", 13 elements per feature vector) and their deltas (second stream, "dplp12", 13 elements). The second pair consisted of the novel modulation-filtered spectrogram features (MSG) recently developed in our group [10], which also split into two banks, covering roughly the 0-8 Hz modulation frequencies (stream 3, "msg3a") and 8-16 Hz (stream 4, "msg3b"). plp12 and dplp12 are most often modeled by a single classifier i.e. combined with FC, as are msg3a and msg3b. The two feature stream pairs have previously been combined with PC [5, 10].

A single classifier may be trained on any number of concatenated streams, corresponding to what we are terming FC. Any number of such classifiers can then have their outputs combined via PC to form a complete system. Four streams gives us 15 possible FC classifiers (4 with one input stream, 6 with two, 4 with three and one with all four), which then offer us 2^{15}-1 or 32767 possible PC configurations. However, most of these use the same feature stream multiple times, grossly violating our independence assumptions. If we limit ourselves to systems in which each stream is used exactly once, and impose the further condition that all classifiers in a given system should have the same number of input streams, there are just five configurations to consider: pure FC, where all streams go into a single classifier; pure PC, combining four separate per-stream classifiers; and the three possible arrangements of streams to build a pair of two-stream FC classifiers which are then combined by PC. These five alternatives for combining the four streams are the main focus of the results section.

2.2. Task and recognizer

The experiments were conducted on the Aurora noisy digits task [11]. This consists of continuous digit strings mixed with four kinds of background noise at several different signal-to-noise ratios (SNRs). Both the training and test data consist of a mix of noise conditions, making this a 'matched multicondition' task.

In every case, the classifiers were multilayer perceptron neural networks, trained with 480 hidden units and 24 output units for the 24 phones used in the vocabulary. Each network took a context window of nine consecutive feature vectors to give input layers varying between 117 and 486 units. The networks were trained by backpropagation, using a minimum-cross-entropy criterion, to hard targets derived from a previous forced-alignment of the training material. A pilot experiment showed that doubling the hidden layer size for a two-stream net improved performance by only 3% relative, indicating that this is not a serious limit to system performance. The final posterior estimates were converted into word hypotheses by the standard hidden Markov model decoder we use [12].

2.3. Stream dependence

To measure the statistical dependence between feature streams, we draw upon [13], which investigated the selection of individual feature elements based on mutual information criteria. The argument in the introduction implied that FC should be preferable to PC when elements in the different streams have structure in their joint distributions relevant to the classification problem. This corresponds to a relatively large conditional mutual information (CMI) between the streams – that is, given the correct class, knowledge of one stream reduces our ignorance of another stream by a certain number of bits; equivalently, it imposes constraints on the distribution of the second stream. By contrast, the conditional statistical independence between streams suggested by the averaging of log posteriors in PC would correspond to an inter-stream conditional mutual information of zero.

Estimating the conditional mutual information is typically both complex and data-intensive. Here we make a series of approximations: Firstly, we discard conditioning and assume that CMI will vary as the unconditioned mutual information (MI) between the streams. Secondly, we approximate the MI between two streams (vectors) by taking the average, for each element within a stream, of the maximum MI across all elements in the second stream. (This makes certain assumptions about the MI within stream elements, so we decorrelate the msg features with a discrete cosine transform in an effort to balance the two stream-pairs in this regard. Also, because this measure is asymmetric, we take the average of the measure in both directions.) Having reduced the problem to the calculation of MI between pairs of feature elements, we again follow [13] in building 5-component Gaussian mixture models of the joint distribution between the elements, then estimating the MI of this distribution numerically.

3. RESULTS

The Aurora task defines 28 different test conditions, varying noise type and SNR. To provide a single figure-of-merit for each system, we calculate the ratio of the word error rate (WER) of the test system to the standard HTK baseline provided with Aurora, and average this across all conditions to get a mean improvement on the baseline. These are the figures presented in table 1.

The first two blocks in table 1 correspond to the pairs of related basic feature streams. Individual feature streams all perform a little worse than the baseline, varying from 5.9% more errors for the direct plp12 features to a 41.6% error increase for the msg3b bank. (The baseline employs deltas and double deltas, so it is forgivable for these individual streams to do worse). When we combine within these basic pairs by FC (denoted $\diamond$ in the table) – i.e. the way they are most commonly used – we see dramatic improvements; there is a 15% improvement over the better of the two plp streams, and a 25% improvement for the msg streams. Combining the pairs by PC instead (denoted by $\oplus$) is far worse in both

Feature combination	Parameters	Baseline %
plp12	68k	105.9
dplp12 (deltas)	68k	125.6
plp12 $\oplus$ dplp12	136k	97.6
plp12 $\diamond$ dplp12	124k	89.6
msg3a (0-8 Hz)	73k	112.7
msg3b (8-16 Hz)	73k	141.6
msg3a $\oplus$ msg3b	145k	101.1
msg3a $\diamond$ msg3b	133k	85.8
plp12 $\oplus$ msg3a	141k	88.3
plp12 $\oplus$ msg3b	141k	86.3
dplp12 $\oplus$ msg3a	141k	89.7
dplp12 $\oplus$ msg3b	141k	89.9
plp12 $\diamond$ msg3a	129k	86.4
plp12 $\diamond$ msg3b	129k	78.1
dplp12 $\diamond$ msg3a	129k	87.5
dplp12 $\diamond$ msg3b	129k	82.6
plp12 $\oplus$ dplp12 $\oplus$ msg3a $\oplus$ msg3b	281k	76.5
plp12 $\diamond$ dplp12 $\diamond$ msg3a $\diamond$ msg3b	245k	74.1
plp12 $\diamond$ msg3b $\oplus$ dplp12 $\diamond$ msg3a	257k	70.1
plp12 $\diamond$ msg3a $\oplus$ dplp12 $\diamond$ msg3b	257k	68.1
plp12 $\diamond$ dplp12 $\oplus$ msg3a $\diamond$ msg3b	257k	63.0

Table 1: Parameter counts and average per-condition ratios of word error rates (WERs) to the baseline system for different combinations of the four feature streams plp12, dplp12, msg3a and msg3b. All features were normalized within each utterance. In the feature descriptions, $\diamond$ indicates streams combined by FC, and $\oplus$ indicates systems combined by PC. $\diamond$ binds more tightly than $\oplus$.

cases, yielding system performances approximately midway between the better of the two individual streams and FC. (Since these PC systems consist of the two individual nets, they have precisely double the parameter count of the individual stream systems. FC uses slightly fewer parameters because the two streams share their hidden-to-output layer.)

The next two blocks show the four remaining 2-stream systems possible with these streams, first combined by PC then by FC. We note that these PC systems, which all perform very similarly, are much better than either the within-plp or within-msg PC systems. This is in line with our experience that, in choosing a second stream for PC, it is better to choose the 'most different' one.

However, our interpretation of the difference between FC and PC is confounded by the results for the crossed FC systems, which outperform PC in every case; the two systems involving msg3a are marginally better, with the two msg3b systems showing 10% relative improvements. At 78.1% of the baseline WER, the FC combination of plp12 and msg3b is particularly good, even though msg3b was by far the worst performing individual stream.

The final block presents the five alternative structures for combining all four streams, as described in the previous section. These are ordered by overall performance, with pure PC bringing up the rear, barely better than the best 2-stream system. Pure FC, in which all four streams are fed to a single network, is a slight improvement. The best schemes, however, combine pairs of streams with FC and the resulting posterior estimates with PC, with the very best configuration being to use FC on the two plp-based streams

	plp12	dplp12	msg3a	msg3b
plp12	-	0.04	0.21	0.10
dplp12		-	0.08	0.06
msg3a			-	0.22
msg3b				-

Table 2: Average maximum element-to-element mutual information (MI) between feature streams (in bits), as an indication of the statistical dependence of the streams. Independent streams would have an MI of zero.

and on the two msg-based streams. At 63% of baseline errors, this is a large improvement on any of the individual streams or common FC pairs. It is also a 15% relative improvement over the pure-FC four-stream system, which forms a kind of baseline for using all four streams. This result at least is inline with our expectations that FC is most useful for related streams, and PC better for more independent streams.

To check that our assumptions about the relative independence of the streams are correct, table 2 gives the mutual information values calculated as described above pairwise for each stream. These results are not quite as expected: the plp deltas (dplp12) share rather little mutual information with any of the other streams, and in fact the smallest MI is between the plp deltas and their direct features, even though a context window of 5 successive direct features would completely determine the deltas – the mutual information, however, compares only simultaneous frames, probably a key weakness in this measure. The 0-8Hz MSG bank (msg3a) is comparatively highly informative with both the direct plp12 features and the second msg3b bank. MI between plp12 and msg3b is intermediate.

Comparing these MI values with the performance of the 2-stream systems in table 1, we fail to see the predicted correlation between large MIs and the advantage of FC. While the largest single MI between msg3a and msg3b corresponds to the pair showing the biggest gain of FC over PC, the next most informative pair, plp12 and msg3a, showed the *smallest* gain in switching to FC. Other results are similarly scattered.

The four-stream combination systems in table 1 are more closely in line with our original hypothesis: The best-performing system is the one that uses FC for the most highly-informative stream pair, msg3a $\diamond$ msg3b, followed by the system including FC for the next highest MI pair, plp12 $\diamond$ msg3a. However, the second pairings in each of these systems corresponds to very low MI values; it's not clear how we would expect this to affect the overall system.

4. DISCUSSION

Our basic thesis, that FC should be preferable to PC for streams that have higher mutual dependence, is only weakly supported by our results. In part this may reflect shortcomings of our method for evaluating stream dependence via the average maximum elementwise estimated mutual information. But even if we had access to the 'true' conditional mutual information values between each stream, that still wouldn't be all the relevant information for predicting the best combination strategies: there is also the influence of the underlying utility of the stream to the basic speech recog-

nition task. For instance, if a new feature stream has a relatively high CMI with some baseline stream, but is also a rather weak basis for phone classification (perhaps because it is a useful predictor of the phonetically irrelevant information remaining in the baseline stream), we may well end up *hurting* the performance of our system by introducing the new stream in combination.

This paper has focussed on the question of how to combine feature streams. The wider question of *whether* it will be advantageous to make a combination, or *which* of several streams should be added, has not been addressed, although our results do provide some relevant information. As shown by the well-performing plp12 $\diamond$ msg3b system, it is not necessarily the best-performing or most (or least) mutually-informative streams that make the best combinations. What matters, rather, is the complementarity of the conditions under which each stream performs better or worse, something that is hard to measure with such global statistics.

It seems important to make some discussion of the statistical significance of the results presented; for instance, is the baseline WER ratio of the four-stream, pure-FC system at 76.5% significantly worse than the 74.1% achieved by the pure-PC variant? For individual test conditions, statistical significance can be evaluated (for instance, in comparison to a simple binomial model of word errors). However, because the baseline error rate varies enormously over the 28 test conditions, this test cannot be applied to the aggregate baseline ratio. As a substitute, we report our informal observation that repeated versions of supposedly equivalent tests (for instance, with slightly different network configurations or starting conditions) yielded results that agreed within 3-5% absolute in the baseline ratio figure; differences within this range are probably not significant.

5. CONCLUSIONS

We have compared two techniques for frame-level combination of feature streams, either by feeding the streams into a single neural-network classifier (FC, feature combination), or by using separate classifiers for each stream then averaging the per-class log posterior probability estimates they emit (posterior combination or PC). By investigating a number of alternative combinations of four feature streams, we demonstrated that different combination strategies can have quite varied success, with the optimal combination dependent on the particular properties of the streams concerned. We argued that PC was most appropriate for streams that are statistically independent (given the class), whereas highly correlated stream should be more advantageously combined with FC. We attempted to measure this dependence with an approximation to the conditional mutual information between streams, but the observed pattern of results was only partially explained by these figures.

Combining multiple feature streams is clearly highly beneficial, giving relative WER reductions of 25-40% in our task. Although we have presented some explanation and interpretation of this benefit, the practical questions of when and how to combine feature streams remain predominantly empirical and in need of considerable further investigation.

6. ACKNOWLEDGMENTS

These ideas are based on discussions with my colleagues at ICSI, most particularly Nelson Morgan. Many thanks to Jeff Bilmes for helpful discussions and for providing the tools to calculate mutual information. This work was supported by the European Union under the ESPRIT LTR project Respite (28149).

REFERENCES

[1] J. Fiscus, "A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

[2] S. Wu, *Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition*, Ph.D. dissertation, Dept. of EECS, UC Berkeley, 1998.

[3] S. Okawa, E. Bocchieri and A. Potamianos, "Multi-band speech recognition in noisy environments," *Proc. ICASSP-98*, Seattle, 2:641-644, May 1998.

[4] N, Morgan, and H, Bourlard, "Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach." *Signal Processing Magazine*, pp 25-42, May 1995.

[5] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP" *Proc. ICASSP-97*, Munich, 2:1259-1262, April 1997.

[6] S. Wu, B. Kingsbury, N. Morgan and S. Greenberg, "Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition," *Proc. ICASSP-98*, Seattle, 2:721-724, May 1998.

[7] A. Janin, D. Ellis and N. Morgan, "Multi-stream speech recognition: Ready for prime time?" *Proc. Eurospeech-99*, Budapest, September 1999.

[8] K. Kirchoff and J. Bilmes, "Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values," *Proc. ICASSP-99*, Phoenix, April 1999.

[9] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," *Proc. ICSLP-96*, Philadelphia, 426-429, October 1996.

[10] B. Kingsbury, *Perceptually-inspired Signal Processing Strategies for Robust Speech Recognition in Reverberant Environments*, Ph.D. dissertation, Dept. of EECS, UC Berkeley, 1998.

[11] D. Pearce, *Aurora Project: Experimental framework for the performance evaluation of distributed speech recognition front-ends*, ETSI working paper, September 1998.

[12] S. Renals and M. Hochberg, "Start-synchronous search for large vocabulary continuous speech recognition," *IEEE Tr. Speech and Audio Proc.* 7:542-553, September 1999.

[13] J. Bilmes, *Natural Statistical Models for Automatic Speech Recognition*, Ph.D. dissertation, Dept. of EECS, U.C. Berkeley, 1999.