Improved recognition by combining different features and different systems

Daniel P.W. Ellis

International Computer Science Institute
1947 Center St. #600, Berkeley CA 94704-1198
(510) 666-2940

<dpwe@icsi.berkeley.edu>

Submitted to AVIOS 2000
2000-04-14

Improved recognition by combining different features and different systems

Daniel P.W. Ellis
International Compute Science Institute, 1947 Center St. #600, Berkeley CA
dpwe@icsi.berkeley.edu

Abstract

Combining multiple estimators to obtain a more accurate final result is a well-known technique in statistics. In the domain of speech recognition, there are many ways in which this general principle can be applied. We have looked at several ways for combining the information from different feature representations, and used these results in the best-performing system in last year's Aurora evaluation: Our entry combined feature streams after the acoustic classification stage, then used a combination of neural networks and Gaussian mixtures for more accurate modeling. These and other approaches to combination are described and compared, and some more general questions arising from the combination of information streams are considered.

Introduction

Despite the successful deployment of speech recognition applications, there are circumstances that present severe challenges to current recognizers – for instance, background noise, reverberation, fast or slow speech, unusual accents etc. In the huge body of published research there are many reports of success in mitigating individual problems, but fewer techniques that are of help in multiple different conditions. What is needed is a way to combine the strengths of several different approaches into a single system.

One thread of research at ICSI has been the development of novel representations of the speech signal to use as features for recognition. Often these are related to aspects of the auditory system, such as the short-term adaptation of RASTA (Hermansky & Morgan 1994) and the 2-16 Hz modulation frequency sensitivity of MSG (Kingsbury 1998). We typically find that each feature type has particular circumstances in which it excels, and this has motivated our investigations into methods for combining separate feature streams into a single speech recognition system.

A related question arises when comparing different basic recognition architectures. ICSI has pioneered the "hybrid connectionist" approach to speech recognition, using neural networks in place of the conventional Gaussian mixture estimators as the recognizer's acoustic model (Morgan & Bourlard 1995). Neural networks have many attractive properties such as their discriminative ability to focus on the most critical regions of feature space, and their wide tolerance of correlated or non-Gaussian feature statistics. However, many tools and techniques have been developed for Gaussian-mixture-based systems that cannot easily be transferred to the connectionist approach. We were therefore also interested in techniques for combining architectures, to create a single system that could exploit the benefits of both the connectionist- and Gaussian-mixture-based systems.

Combination techniques also offer various practical advantages. International collaborations are central to ICSI's charter, and good mechanisms for combining relatively independent systems have made it possible for us to build single recognition systems that combine acoustic models trained at ICSI with those from our European collaborators. At a smaller scale, being able to con-

struct systems from relatively-independent pieces without having to retrain the entire assembly can significantly increase the overall complexity of the recognizers we can practically produce.

This paper briefly reviews some of the theoretical attractions of combination, then surveys several practical realizations of these advantages, many arising from projects conducted at ICSI. We conclude with a discussion of some of the outstanding research issues in combination-based speech recognition systems.

The justification for combinations

Combination is a well-known technique in statistics. For instance, if we have several different ways to obtain an estimate of the same underlying value, a better estimate can usually be obtained by combining them all, for instance by averaging. The key requirement for this to be beneficial is that the 'noise' in each estimator (i.e. the difference between the estimated and true values) should be uncorrelated. If this is true, for instance if the estimators are based on different measurement techniques subject to different sources of error, then on average the errors will tend to cancel more often than they reinforce, so an optimal combination will improve accuracy.

An example of this principle in action is the system of Billa et al. (1999). They combined three nearly identical sets of speech features, the only difference being that the analysis frame rate varied between 80 and 125 samples per second. Although all the feature sets were using the same representation, the slight differences in how the signal was framed were enough to introduce some decorrelation between the errors in the streams, and the combined system performed significantly better than any of the component streams. For the neural network models used at ICSI, another way to get different estimators from a single feature representation is to train the networks based on different random starting conditions, and we have seen some small benefits from this approach (Janin et al. 1999). However, using a pair of networks based on the same features did not perform as well as training a single network of twice the size. In speech recognition, we can almost always get a marginal improvement simply by increasing the number of model parameters, so combination schemes need to offer some advantages (in terms of performance or practicalities) in comparison to this simplest approach.

As mentioned above, our experience in practice is that certain processing will perform particularly well on certain subsets of the data. If we could figure out when this is the case, either because the estimator has some measure of how well it is doing, or because we have a separate classifier telling us which model is likely to work best, then we would expect to be able to make a more successful combination of the information. This "mixture of experts" approach has also been widely investigated, but for speech recognition it often proves as difficult to classify the data into different domains of expertise as to make the overall classification into speech sounds [Morris 2000]. However, a combination system that somehow tends to de-emphasize the poorly-performing models will be preferable to unweighted averaging.

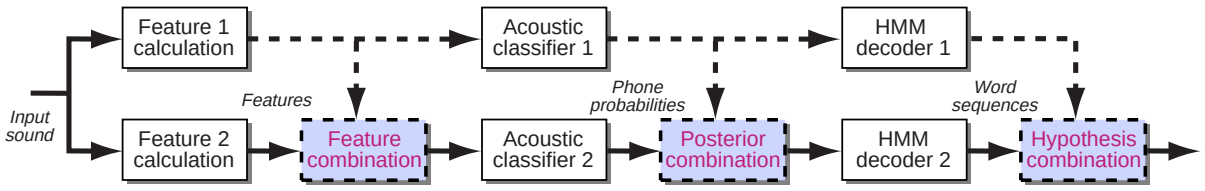

Figure 1: Alternative combination strategies for a recognition system based on two feature streams. The information may be combined after feature calculation, after classification, after decoding, or by some combination of these strategies.

Figure 1 illustrates some of the ways in which two feature streams might be combined in a recognizer. If we break the speech recognition into feature calculation, acoustic classification and HMM decoding, the streams could be combined after any of these stages. Specific examples of each of these three possibilities are discussed below.

Feature combinations

Our first forays into feature combination came during the early developments of MSG features (Kingsbury & Morgan 1997). A set of novel features gave a word error rate (WER) twice as large as the standard baseline features, yet when the two systems were combined by simply multiplying the posterior probability estimates for each phone, combining with the weaker features affected a 15% relative WER reduction on our baseline features. Such “posterior combination” via multiplication remains the most successful combination scheme we have found at this level; a possible explanation for this success is that it carries an element of the “mixture of experts” approach mentioned above: If a poorly-performing classifier becomes ‘equivocal’ so that all possible classes are given roughly equal probabilities, then it will have little or no net effect when combined via multiplication with a more confident model; the weaker model will be discounted by the combination.

We have since used posterior combination in a variety of situations, including the 1998 DARPA/NIST Broadcast News evaluation system from the SPRACH project, a collaboration with Cambridge and Sheffield universities (Cook et al. 1999). Most recently, we applied the approach to the 1999 Aurora task (Pearce 1998); this was one of the key elements in the best-performing system in that evaluation, a collaboration between ICSI, the Oregon Graduate Institute, and Qualcomm (Sharma et al. 2000).

Our results in this task were obtained via posterior combination (PC) – using separate acoustic classifier models for each feature stream and combining the posterior probability outputs. For completeness, we made extensive comparisons with the simpler approach of feature combination (FC) – i.e. merging the two feature streams and classifying them jointly within a single, larger model. We found a complex pattern of results, with the best approach depending on which feature streams were to be combined (Ellis 2000); some of the results are shown in table 1. We argued that feature combination was better suited to feature streams that showed informative co-dependence, and posterior combination was more appropriate when the feature spaces were closer to independent. (In fact, posterior combination by multiplication is very close to the optimal strategy for two streams that are conditionally mutually independent).

System configuration	Average word error rate (varying SNR)
Baseline Aurora recognizer	13.7%
Feature combination (4 streams into 1 net)	8.1%
Posterior combination (4 nets)	9.2%
Best mixture (2 nets of 2 streams)	7.1%

Table 1: Comparative performance of different Aurora recognition systems. Apart from the baseline, all results are based on the same four feature streams. The word error rate is an average over signal-to-noise ratios between 20 and 0 dB.

System combinations

The final scheme indicated in figure 1 is hypothesis combination (HC), where complete, independent recognizers are run based on each feature stream, then their final word hypotheses merged to give a combined result. This has been implemented by the ROVER software (Fiscus 1997), developed by NIST. ROVER takes as input as many different word-string hypotheses as there are independent systems available, then generates a best 'consensus' word-string, respecting any confidence values attached to the words by the individual recognizers. When used with the 9 Broadcast News systems in the 1998 NIST evaluation, ROVER managed to improve the 14% WER of the best single system to almost 10%, a most significant gain (Pallett et al. 1999).

ROVER was also used within the SPRACH Broadcast News system. Feature streams were merged by posterior combination in various ways to generate thee word-hypothesis streams for ROVER to combine. Thus, based on extensive empirical investigations, this system employed two of figure 1's combination strategies at the same time, showing that different combination strategies are exploiting different kinds of information within the underlying feature streams, and can thus themselves be combined.

A different approach to system combination is illustrated in figure 2. The Aurora task mentioned above is specifically concerned with defining *features* that might for instance be calculated within future mobile telephone handsets prior to transmission to speech recognition systems at a central location. We were interested to see if our neural network processing could be beneficial in such an application, so we experimented with training a conventional Gaussian-mixture-based speech recognition system (using the HTK scripts provided with the Aurora tasks) taking as input features the posterior probabilities for each phone estimated by our connectionist models. Thus, we combined in tandem two different pattern-recognition approaches, neural networks and Gaussian mixtures, to obtain more detailed modeling of the acoustic signal. Somewhat to our surprise, this was very successful, resulting in recognizers that were significantly superior to either connectionist or Gaussian-mixture-based systems alone. Our best system used four different feature streams (developed both by us and by our partners at OGI), classified by independent neural networks then merged by posterior combination. Simple transformations were applied to this probability stream to make it a better match to Gaussian assumptions, and it was then passed as features to the Gaussian mixture recognizer. This system reduced the average word error rate by some 60% compared to the standard Gaussian-only baseline, the best performance in the evaluation by a significant margin (Hermansky et al. 2000).

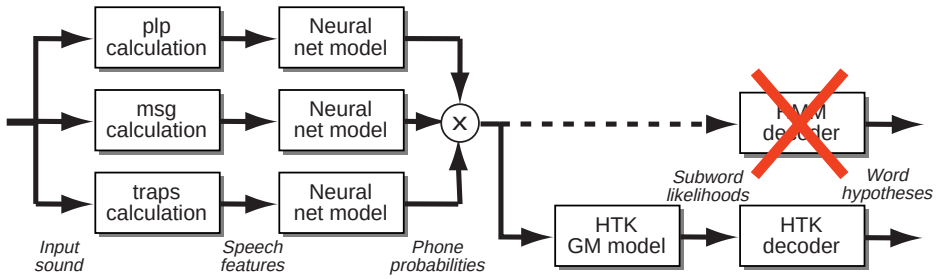

Figure 2: The 'tandem' system configuration. Several feature streams are separately modeled then combined at the posterior probability level. Rather than feeding these probabilities directly to an HMM decoder, they are used as features for a Gaussian mixture based recognizer (implemented within the HTK package).

Conclusions

It should come as no surprise that combining systems can lead to performance improvements, but it is more impressive how very significant these improvements can be, as shown in the examples cited above. This implies that the different features and approaches are not only extracting important information that is being missed by their peers, but also that our current implementations have contrived to weight the streams beneficially. However, we recognize that our combination schemes are rather arbitrary and there must surely be ways to optimize them more precisely to realize even greater gains from combination.

Questions we would like to answer include identifying which streams will combine most successfully, and which combination method (or methods) should be used. Combination relies on the idea that the streams behave differently, but at the same time that difference should not be at the cost of reduced accuracy in the underlying classification task. We are investigating information-theoretic measures such as condition mutual information, both between the feature streams and between the subsequent classifier outputs, to see if we can predict which streams to combine without having to test every possibility.

The tandem combination of connectionist and Gaussian-mixture recognition strategies opens a wide range of interesting issues. Ideally, by incorporating both kinds of statistical model, we can build a system that can benefit from the relative advantages of both approaches. We have already shown that posterior combination, one of the attractions of connectionist modeling, is useful in this setting; we would like to see if recognition-time adaptation schemes that have been used with great effect in Gaussian mixture systems – such as MLLR (Leggetter& Woodland 1995) – can continue to be useful in this highly nonlinear domain.

Combination schemes have shown themselves to be very useful despite the fact that the origin of their benefits is somewhat unclear. A clearer understanding of when and how combination serves to reduce errors would allow us to improve our schemes and perhaps even to design new features specifically intended for combination, rather than being concerned about their performance when used alone.

References

J. Billa, T. Colhurst, A. El-Jaroudi, R. Iyer, K. Ma, S. Matsoukas, C. Quillen, F. Richardson, M. Siu, G. Zavaliagkos, H. Gish, "Recent Experiments in Large Vocabulary Conversational Speech Recognition," *Proc. IEEE Int. Conf. on Acous., Speech & Sig. Proc.*, Phoenix, May 1999.

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.

D. Ellis, "Feature stream combination before and/or after the acoustic model," ICSI Technical Report TR-00-007.

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.

H. Hermansky, D. Ellis and S. Sharma, "Connectionist feature extraction for conventional HMM systems," *Proc. IEEE Int. Conf. on Acous., Speech & Sig. Proc.*, Istanbul, June 2000.

H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech & Audio Proc.* 2(4), 1994.

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

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

B. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," *Proc. IEEE Int. Conf. on Acous., Speech & Sig. Proc.*, Munich, 2:1259-1262, April 1997.

C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language* 9, 171-186, 1995.

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

A. Morris, personal communication concerning the characterization of feature statistics at different signal-to-noise ratios, January 2000.

D. Pallett, J. Fiscus, J. Garofolo, A. Martin, M. Przybocki, "1998 Broadcast News Benchmark Test Results," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Herndon VA, 1999.

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

S. Sharma, D. Ellis, S. Kajarekar, P. Jain and H. Hermansky, "Feature extraction using non-linear transformation for robust speech recognition on the Aurora database," *Proc. IEEE Int. Conf. on Acous., Speech & Sig. Proc.*, Istanbul, June 2000.