TANDEM CONNECTIONIST FEATURE EXTRACTION FOR CONVENTIONAL HMM SYSTEMS

Hynek Hermansky[1,2], *Daniel P.W. Ellis* [2] *and Sangita Sharma*[1] *

[1]Oregon Graduate Institute of Science and Technology, Portland, Oregon, USA
[2]International Computer Science Institute, Berkeley, California, USA

ABSTRACT

Hidden Markov model speech recognition systems typically use Gaussian mixture models to estimate the distributions of decorrelated acoustic feature vectors that correspond to individual subword units. By contrast, hybrid connectionist-HMM systems use discriminatively-trained neural networks to estimate the probability distribution among subword units given the acoustic observations. In this work we show a large improvement in word recognition performance by combining neural-net discriminative feature processing with Gaussian-mixture distribution modeling. By training the network to generate the subword probability posteriors, then using transformations of these estimates as the base features for a conventionally-trained Gaussian-mixture based system, we achieve relative error rate reductions of 35% or more on the multi-condition Aurora noisy continuous digits task.

1. INTRODUCTION

The standard structure of current speech recognition systems consists of three main stages. First, the sound waveform is passed through feature extraction to generate relatively compact feature vectors at a frame rate of around 100 Hz. Secondly, these feature vectors are fed to an acoustic model which has been trained to associate particular vectors with particular speech units; commonly, this is realized as a set of Gaussian mixtures models (GMMs) of the distributions of feature vectors corresponding to context-dependent phones. Finally, the output of these models provides the relative likelihoods for the different speech sounds needed for a hidden Markov model (HMM) decoder, which searches for the most likely allowable word sequence.

The acoustic model is trained using a corpus of examples that have been manually or automatically labeled. For distribution Gaussian-mixture models, this can be done according to a maximum-likelihood criteria via the EM algorithm. However, this is not optimal: typically, we would rather have a discriminative criteria that optimized the ability to distinguish different classes, rather than just the match within each class. The hybrid connectionist-HMM framework [1] replaces the GMM acoustic model with a neural network (NN), discriminatively trained to estimate the posterior probabilities of each subword class given the data. Hybrid systems have been shown to have comparable performance to GMM-based systems for many corpora, and are argued to give simpler systems and training procedures.

Because of the different probabilistic basis (likelihoods versus posteriors) and different representations for the acoustic models (means and variances of mixture components versus network weights), techniques developed for one domain are often difficult to transfer to the other. The relative dominance of likelihood-based systems has resulted in the availability of very sophisticated tools such as HTK [2] offering advanced, mature, and integrated system parameter estimation procedures. On the other hand, discriminative acoustic model training and certain combination strategies facilitated by the posterior representation are much more easily implemented within the connectionist framework.

In this paper we successfully combine these two approaches by using the output of a neural network classifier as the input features for the Gaussian mixture models of a conventional speech recognizer. The resulting system, which effectively has two acoustic models in tandem - first a neural-net then a GMM - performs significantly better than either the hybrid or conventional baselines on the Aurora noisy digits task [3], achieving an average 35% relative error rate reduction over the multiple test conditions when based on the same mel-cepstral features. By exploiting the combination schemes available for connectionist models, systems based on multiple features streams can also be constructed, with even more dramatic reductions in error rate.

The next section describes this tandem structure in more detail. Section 3 describes our results on the Aurora task, and section 4 discusses the implications and interpretation of these results, which we summarize in the final section.

2. APPROACH

The overall system is illustrated in figure 1. The training procedure is rather simple. First, a hybrid connectionist-HMM system is trained, which amounts to training the neural network acoustic model (a conventional multi-layer perceptron (MLP) structure with one hidden layer) to estimate the posterior probabilities of each possible subword unit (in our case, context-independent phones). The network is trained by backpropagation with a minimum-cross-entropy criterion to 'one-hot' targets obtained from either hand labeling or a forced alignment of the training data generated using an earlier acoustic model. (For the results below, the entire training and realignment process was repeated several times to stabilize the labels). The input to the network is a context window of several successive frames of the feature vector; we typically use a context window of 9 frames, corresponding to 90 ms of audio at a 10 ms frame rate.

The output of the neural network is a vector of posterior probabilities, with one element for each phone; one such vector is generated for context windows centered on each input feature vector. Conventionally, these would go directly to an HMM decoder to find the word sequence, but instead we use them as the 'feature' inputs for a Gaussian-mixture-based HTK system. Typically, the

* Joint first authors appear in random order.

0-7803-6293-4/00/$10.00©2000 IEEE

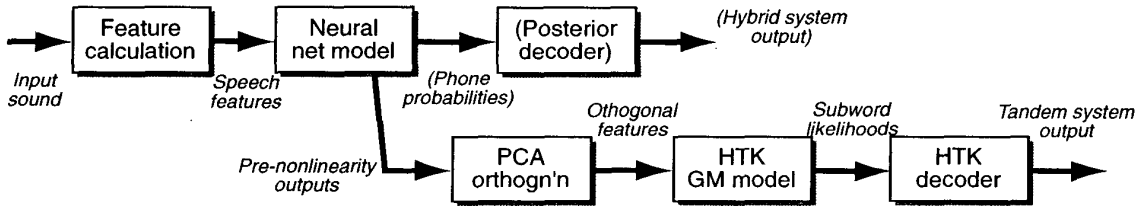

Figure 1: Block diagram of the tandem speech recognition system, in which a neural network, trained to phone targets, is used to generate input features fed to a conventional GMM HTK recognizer. Items in parentheses correspond to the conventional hybrid recognition system.

number of phones is between 30 and 50, so the total dimensionality of the feature space is much the same as with normal features augmented by deltas and double-deltas.

Because the posterior probabilities have a very skewed distribution, we find it advantageous to warp them into a different domain, for instance by taking their logs. An alternative to this is to omit the final nonlinearity in the output layer of the neural network. We use the "softmax" nonlinearity in this position (exponentials normalized to sum to 1), so simply skipping it is very close to taking the log of the subsequent probabilities.

The 'features' constituted by the log-posteriors have the rather unusual property of tending to contain one large value (corresponding to the current phone) with all other values much smaller. We find that applying a global decorrelation via the Karhunen-Loeve (KL) transform improves system performance, presumably by improving the match of these features to the Gaussian mixture models.

For our experiments below with the Aurora task, our GMM-based HTK system was the baseline defined for that task: 11 whole-word models of 18 states each, and 3 mixture components per state, plus a 5 state, 6 mixture silence model.

The Gaussian mixture system must of course be retrained with the new features. This can be done on the same training set as was used to train the neural network, although better results should be possible by using a second set of utterances held out from the original training, to make the features truly representative of the behavior of the net on unseen data. This has to be balanced against the impact of reducing the training data available to each stage, which we did not investigate.

A more radical approach is to use a completely separate task to train the neural network; some results and discussion relating to this option are presented below.

3. RESULTS

Our results are summarized in table 1. The Aurora task consists of connected digit strings (from TIdigits) mixed with four different kinds of background noise at 7 different signal-to-noise ratios (SNRs) from clean to -5 dB, for a total of 28 test conditions. For clarity, we report just three word-error-rate (WER) figures for each system, averaged across the four noise conditions and spanning several SNR levels. For a single figure-of-merit, we report the average per-condition ratio of word-error rate to the baseline HTK system using the plain mel-frequency cepstral coefficient (MFCC) features plus deltas and double-deltas. All the other systems in table 1 are also based on the 14 MFCC features plus deltas and double-deltas, and use the same neural network model which took a context window of 9 frames for a total of 378 input units feeding a hidden layer of 480 units then an output layer of 24 units, one for each of the phones used in our pronunciation models for this task.

System	WER% / SNR Clean	15dB	5dB	Baseline ratio %
HTK baseline	1.4	3.7	15.9	100.0
Hybrid baseline	1.6	2.6	8.7	84.6
Tandem logp	0.9	2.2	9.0	69.1
Tandem lino	1.2	2.5	9.3	81.4
Tandem logp+KL	1.1	2.1	9.3	71.0
Tandem lino+KL	0.9	2.1	8.0	64.5

Table 1: Word error rates and average WER-to-baseline ratios for different systems, all based on MFCC features. The first line is the standard HTK GMM baseline defined for the Aurora task. The second line is a conventional hybrid system, based on the posterior estimates generated by the neural-net acoustic model. The remaining four lines are the results of tandem systems, feeding versions of the posteriors into the HTK system; "logp" indicates that the log of the posteriors are taken, whereas "lino" systems use the neural-net outputs directly, before the final nonlinearity is applied. "+KL" indicates that a full-rank Karhunen-Loeve orthogonalization was applied before passing the values to the HTK system.

From the table, we see that the baseline hybrid system has, on average, only about 85% of the word errors of the HTK baseline, although it performs insignificantly worse in the clean condition. (Since there are 13,159 words in each per-SNR test, 5% significance requires a difference of about 0.25% in WER). The best tandem system, using the pre-nonlinearity outputs of the network plus Karhunen-Loeve orthogonalization, manages to reduce the baseline word error rate by more than a third overall, for a very significantly improved performance.

3.1. Feature stream combination

One approach that has shown itself to be beneficial time and again in hybrid systems is feature stream combination via simple averaging of the log posterior probabilities from several independent acoustic model networks [4, 5]. Since this generates a posterior probability stream comparable to the output of a single network, we can similarly use it as input to an HTK system, either by taking log probabilities followed by orthogonalization, or by simply summing the pre-nonlinearity network outputs, which is mathematically very similar as discussed above. To illustrate, table 2 gives some example results of comparing two different feature streams,

System	WER% / SNR Clean	15dB	5dB	Baseline ratio %
Hybrid PLP	2.6	2.8	10.6	89.6
Hybrid MSG	2.1	2.9	11.6	87.1
Hybrid PLP+MSG	1.3	1.9	8.5	60.6
Tandem PLP+MSG	0.7	1.5	7.2	47.2

Table 2: Word error rates and average WER-to-baseline ratios for different Aurora systems based on PLP and MSG features. "Hybrid", and "Tandem" have the same meanings as the previous table. Feature combination for the hybrid system was by averaging log posteriors; for the tandem system, it was by summing pre-nonlinearity outputs. The tandem system also uses KL orthogonalization.

System	WER% / SNR Clean	20dB	15dB	Baseline ratio %
HTK MFCC baseline	1.0	5.3	11.0	100.0
Tandem Aurora-PLP	1.0	2.1	4.4	72.1
Tandem Stories-PLP	1.3	10.3	16.3	144.6

Table 3: Word error rates and average WER-to-baseline ratios comparing tandem systems based on neural networks based on PLP features and trained over the full multi-noise Aurora set (Aurora-PLP) or the separate, clean OGI Stories corpus (Stories-PLP). All GM models were trained on clean data only, although the Aurora-PLP net was trained on the full range of conditions.

PLP [6] and modulation-filtered spectrogram (MSG, [7]). These are two components of our full Aurora system, which is described in more detail in [8]. (Note that the quoted results are for an earlier system that used per-utterance normalization of all features; this tends to hurt the clean case, but help the high-noise cases. More important, however, are the comparative results between the systems in the table).

These results show again the great benefits obtained in combining features as distinct as PLP and MSG. These gains not only carry over into the tandem system, but we again see a relative improvement of approximately 20% in the overall ratio-to-baseline figure by adding the GMM stage onto the base neural network.

3.2. Cross-corpus results

All tandem systems presented so far have used the same Aurora multi-condition training set for both the neural network and the GMM training. For a number of reasons, including the effort required to build a system for a new task, it would be desirable to have "task-independent" version of the neural-net stage, constituting a single "black-box" feature extractor, similar to MFCCs, which could be applied in a wide range of circumstances. To investigate this possibility, we trained a network of the same size as our previous examples on the large-vocabulary OGI Stories corpus (which we have used previously as a source of general-purpose acoustic models [9]). Since the Stories corpus contains only clean speech, we tested on a modified Aurora task using only the clean utterances for training the HTK system, and testing only on clean and SNRs of 20 and 15 dB. These results are shown in table 3.

Note that although all the HTK trainings and the Stories network were based on clean data only, the comparison Aurora-based network was trained over the full range of noise conditions, and hence permits much better performance for the non-clean conditions. Unfortunately, we see that the Stories-based network performs much worse, and significantly worse than the MFCC-based HTK baseline. This result is discussed in the next section.

4. DISCUSSION

This is far from being the first time that neural networks have been proposed as feature preprocessors for speech recognition. Bengio [10] suggested using them to increase state likelihoods in HMM systems, and Rigoll and Willet [11] showed significant improvements from an MLP inserted as a feature preprocessor into a previously-trained Gaussian-model HMM system, again training the net based on the HMM state. Fontaine et al. [12] use the first 3 layers of a four-layer net as a form of "non-linear discriminant analysis" (NLDA), to emphasize the relationship to the better-known linear discriminant analysis (LDA). They achieved a 20-25% relative error reduction for the Phonebook large-vocabulary isolated-word corpus. However, training of four layer networks is typically rather demanding, and the unknown structure of the representation employed in the hidden layer precludes the kinds of combinations described above.

Given the quite dramatic gains shown by the tandem architecture, it is worth spending a little time discussing what is actually going on. Our view of these neural network classifiers is that they focus their modeling power on the small patches of feature space that lie on the boundaries between phones, since these represent the most difficult cases in the training set [13]. Thus, we can imagine the neural network performing some kind of global remapping of feature space in which these boundaries are vastly magnified and sensitively mapped, whereas all the mid-class regions are discounted and only coarsely reflected in the output. This is exactly what you want from a feature space: that it emphasizes significant variation and minimizes or removes irrelevant detail. However, this usually comes at the price of task specificity, as discussed below.

Although it is easy to accept that this discriminative transformation confers performance advantages, this doesn't explain why the tandem arrangement of GMM after neural-net should perform better than the the conventional hybrid system using the neural network outputs directly. We can only speculate about the cause of this additional gain, but we note (a) the Gaussian models do introduce a large number of additional parameters, evidently in a useful way; (b) in these experiments, the HTK system used whole-word models rather than phone models, which may have provided an alternative, advantageous perspective on the training data; and (c) the HTK system does a full re-estimation of all HMM parameters, whereas the simpler hybrid system used fixed transition penalties, state durations etc. In the past we have seen little or no cost to this simpler structure, but it may have been poorly matched to this task.

As an alternative to conventional feature calculation, the features-plus-neural-net has several distinctive characteristics:

Since the net is trained to discriminate specific phone targets, it is intrinsically language dependent; indeed, in the small-vocabulary Aurora task, the net will not really be learning 'context-independent' phones at all, but rather the phones in the few specific

contexts in which they are observed. We attempted to evaluate to what extent we were taking advantage if this property in the current work by training the feature network on the larger variety of speech material in the OGI Stories corpus, and, as shown in the table 3, observed a large increase in the error rate for the noisy conditions.

It may be of an interest that in our earlier work [9] we also compared corpus-dependent and corpus-independent nets and observed only about 25-30% increase in the error rate from the use of the corpus-independent feature net. Both corpora contained relatively clean telephone speech and differed only in the vocabulary (i.e. the development and test sets were collected under identical conditions). Such an increase is consistent with the current result for the clean condition (the first column in the table 3). Thus, we still believe that a corpus-independent version of this approach should be feasible but would require better understanding of issues involved in training of the feature mapping net.

The amount of additional calculation involved is also rather large compared to 'conventional' feature extraction. Our neural networks have in the region of 200,000 parameters, which corresponds the number of multiply/adds required to generate each feature vector. This is probably an order of magnitude larger than the FFT and summations involved in calculating MFCCs. It is, however, comparable to the calculations performed in the acoustic classifier (since it is in fact an acoustic classifier), so the overall impact on the entire recognition process is not enormous. In [8], we show positive results with this approach but using much smaller nets of about 28,000 weights in total.

One of our motivations in pursuing this work is to develop a framework able to exploit the advantages of both GMM and neural-net based systems. Although that is demonstrated in our results, there are other techniques that may not combine so well. For instance, condition and speaker adaptation is often achieved in GMM systems by adjusting the means (and perhaps variances) of the mixture components, treating these variations as low-dimensional transformations of feature space. If, however, these variations shift the basic acoustic features away from the regions of 'high magnification' provided by the neural network, no amount of mixture-shifting in the subsequent GMMs will be able to bring the overall system back to optimal performance.

5. CONCLUSION

Historically, the use of neural-net acoustic models in the hybrid connectionist-HMM speech recognition framework has been seen as a rival alternative to the mainstream GMM-HMM approach. In this work, however, we have shown a simple scheme that combines both modeling approaches to achieve gratifying benefits. This work is at an early stage: it begs many questions over exactly where the benefits are coming from, and how broadly they may be applicable. However, given this very encouraging start, we are strongly encouraged to continue investigating such tandem NN-GMM acoustic modeling.

6. ACKNOWLEDGMENTS

This research was supported by DoD under MDA904-98-1-0521, NSF under IRI-9712579, and by industrial grant from Intel Inc. Author Ellis was supported by the European Union under the ESPRIT LTR project Respite (28149).

REFERENCES

[1] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach.* Kluwer Academic Publishers, Boston, 1994.

[2] S. Young, J. Odell, D. Ollason, V. Valtchev and P. Woodland, "Hidden Markov Model Toolkit V2.1 reference manual," Technical report, Speech group, Cambridge University Engineering Department, March 1997.

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

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

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

[6] H. Hermansky, "Perceptual linear predictive (plp) analysis for speech," *The Journal of The Acoustical Society of America*, 87:1738-1752, April 1990.

[7] 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.

[8] 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," in preparation.

[9] H. Hermansky, S. Sharma and P. Jain, "Data-derived nonlinear mapping for feature extraction in HMM," *Proc. ASRU-99*, Keystone, December 1999.

[10] Y.R. Bengio, R. De Mori, G. Flammia and R. Kompe, "Global optimization of a neural-hidden markov model hybrid," *IEEE Trans. on Neural Networks*, 3:252-258, 1992.

[11] G. Rigoll and D. Willett, "A NN/HMM hybrid for continuous speech recognition with a discriminant nonlinear feature extraction," *Proc. ICASSP-98*, Seattle, April 1998.

[12] V. Fontaine, C. Ris and J.M. Boite, "Nonlinear Discriminant Analysis for improved speech recognition", *Proc. Eurospeech-97*, Rhodes, 4:2071-2074, 1997.

[13] G. Williams and D. Ellis, "Speech/music discrimination based on posterior probability features," *Proc. Eurospeech-99*, Budapest, September 1999.