# TANDEM ACOUSTIC MODELING IN LARGE-VOCABULARY RECOGNITION

*Daniel P.W. Ellis*[1]*, Rita Singh*[2]*, and Sunil Sivadas*[3]

[1]Dept. of Electrical Eng., Columbia University, New York NY, USA
[2]School of Computer Science, Carnegie Mellon University, Pittsburgh PA, USA
[3]Dept. of Electrical and Computer Eng., Oregon Graduate Institute, Portland OR, USA

## ABSTRACT

In the tandem approach to modeling the acoustic signal, a neural-net preprocessor is first discriminatively trained to estimate posterior probabilities across a phone set. These are then used as feature inputs for a conventional hidden Markov model (HMM) based speech recognizer, which relearns the associations to subword units. In this paper, we apply the tandem approach to the data provided for the first Speech in Noisy Environments (SPINE1) evaluation conducted by the Naval Research Laboratory (NRL) in August 2000. In our previous experience with the ETSI Aurora noisy digits (a small-vocabulary, high-noise task) the tandem approach achieved error-rate reductions of over 50% relative to the HMM baseline. For SPINE1, a larger task involving more spontaneous speech, we find that, when context-independent models are used, the tandem features continue to result in large reductions in word-error rates relative to those achieved by systems using standard MFC or PLP features. However, these improvements do not carry over to context-dependent models. This may be attributable to several factors which are discussed in the paper.

## 1. INTRODUCTION

Neural networks (NNs) have several qualities making them attractive as feature classifiers in speech recognition systems [1]: When used to estimate the posterior probabilities of a closed set of subword units, they allow discriminative training in a natural and efficient manner. They also make few assumptions about the statistics of input features, and have been found well able to cope with highly correlated and unevenly distributed features - such as spectral energy features from several adjacent frames [2]. These qualities distinguish NNs from Gaussian mixture models (GMMs), which are often used to build independent distribution models for each subword (i.e. they are not discriminative), and which work best when supplied with low-dimensional, decorrelated input features.

In small tasks, the so-called 'hybrid' NN-HMM systems have performed as well as or better than GMM-HMM systems. However, in large-vocabulary tasks such as DARPA/NIST Broadcast News [3], GMM-HMM systems have performed significantly better, in part due to extensive speaker and environment adaptation. Equivalent adaptation is much more difficult for NN-based systems.

Combining NN and GMM modeling within a single system holds the potential of combining the advantages of both, and several groups have pursued variants of this theme [4, 5]. We recently developed a particularly simple variant, which we have termed 'tandem acoustic modeling' [6], in which an NN classifier is first trained to estimate context-independent phone posterior probabilities. The probability vectors are then treated as normal feature vectors and used as the input for a conventional GMM-HMM system, which is not given any knowledge of the special information represented by its input features.

A system based on this tandem approach performed best in the 1999 ETSI Aurora evaluation [7], which involved recognizing continuous digit strings in a wide range of noisy backgrounds. We were interested in whether this result would generalize to larger tasks, in which GMM systems can outperform NNs by exploiting a much larger repertoire of subword units than is practical with a network. Would the use of an NN feature preprocessor continue to confer an advantage in larger tasks involving more contextual variability? Also, would model adaptation schemes such as MLLR [8] be effective in the new feature space defined by the network outputs?

To address these questions, we developed a tandem system for the NRL SPINE1 task [9]. This task involves a medium-sized vocabulary of about 5000 words, and the utterances are predominantly noisy, with signal-to-noise ratios ranging from 5 dB to 20 dB. The data consists of human-human dialogs between two individuals seated in separate sound booths, involved in a battleship game. The individuals communicate through push-button communication devices similar to those used by military personnel in the field. Pre-recorded noises from real military settings are played out in each booth, thus simulating noisy communication during military action. The recognition task for this kind of data is very challenging. Whereas word error rates (WERs) in the noisy digits task were at 1% or below for the best cases, the very best systems in the recent SPINE1 evaluation achieved about 25% WER [9].

In the next section we describe the tandem system for SPINE1 and how it was trained. Section 3 presents our experimental results which compare context-independent and context-dependent systems, with and without the NN preprocessor and MLLR. In section 4 we discuss the results, followed by our conclusions.

## 2. THE SPINE1 TANDEM SYSTEM

The SPINE1 tandem recognizer is illustrated in figure 1. Input speech is fed to two feature extraction blocks, one generating the widely-used Perceptual Linear Prediction (PLP) features, the other calculating Modulation-filtered Spectrogram (MSG) features [2]. In our previous experiments with multiple feature streams, we have consistently found that combining these two representations can lead to significant error reduction [10]. The base PLP feature vector is a 13-element cepstrum, usually augmented by deltas and double-deltas. The MSG features consist of two banks of 14 spectral energy features each; the first bank is filtered to contain

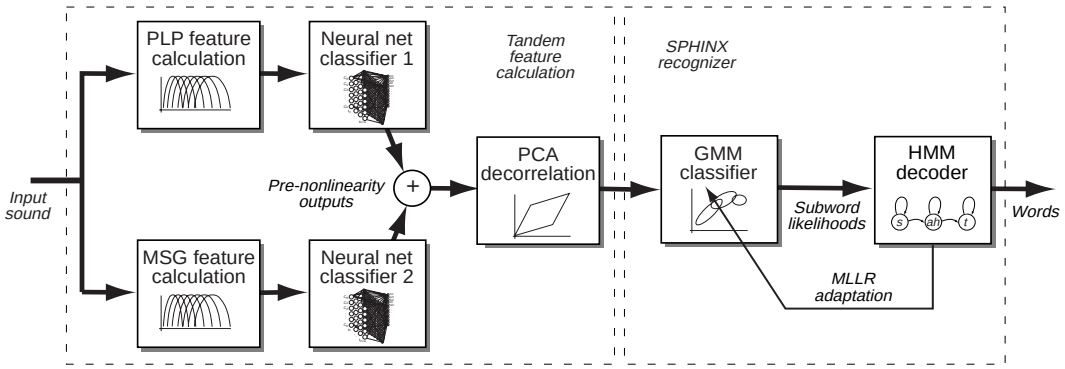


**Fig. 1**. Block diagram of the tandem recognizer used for the SPINE1 task.

modulation frequencies between 0 and 8 Hz, while second covers 8 to 16 Hz.

Each feature stream feeds its own neural network classifier, a multi-layer perceptron with a single hidden layer. The input to the network is a window of successive feature vectors (in this case, 9 frames), that provides the classifier with temporal context. The MSG network has 252 input units ($9\times28$) and the PLP network has 351 ($9\times39$) input units. The output layer consists of 56 nodes, each associated with a particular context-independent phone class. Input and output are connected by a fully-connected hidden layer of 1000 units.

Because each network is trained to estimate the same phone targets, the activation of corresponding output units can be combined to 'pool the expertise' of the two nets. In multistream hybrid-NN-HMM systems, we find that the geometric mean of the posteriors (i.e. averaging in the log domain) consistently performs best (or close to best) among simple combination schemes [10]. In the tandem context, however, we use the neural network activations without their final 'softmax' linearity to improve the symmetry and Gaussianness of their distributions. This is equivalent, modulo a scale factor, to the log of the posterior outputs, and thus we find that a simple sum of the corresponding activations in the two nets is an effective way to combine the feature streams.

This summed phone-activation vector is then passed through a static decorrelation matrix obtained from PCA analysis over the training data. This gave a modest performance gain in our small vocabulary work, presumably by improving the match between the feature distributions and the assumptions of the GM models. There is no data reduction associated with this step. The output of this stage is a 56 element feature vector at each time frame; these are then used in place of the usual feature vectors in an otherwise unmodified GMM-HMM system, in our case the CMU SPHINX-III recognizer.

### 2.1. System training

The training of each of the several classifiers in a tandem system is of course the main factor determining overall performance. In essence the scheme is very simple: the neural network stage is trained according to the normal procedure used for a hybrid NN-HMM system, then the features extracted from the network are fed to the GMM-HMM system which is trained according to a standard EM procedure.

Training the network for a hybrid system, however, requires

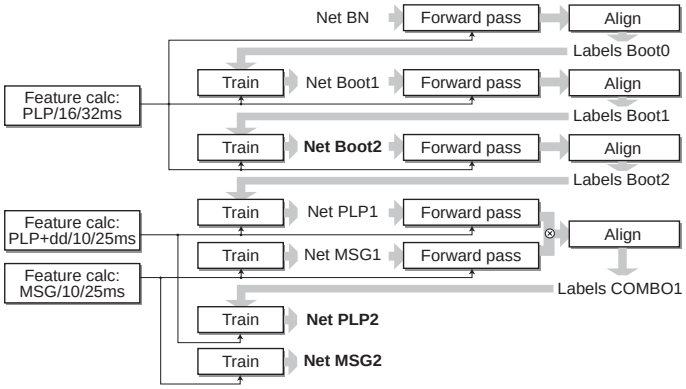


**Fig. 2**. Training path for the neural networks.

aligned target labels for all the training data, which are used in back-propagation training with a minimum-cross-entropy criterion. The procedure for generating these labels for the SPINE1 training data is illustrated in figure 2. To bootstrap, we used a net developed for the DARPA Broadcast News task [3] (denoted "Net BN" in the figure), a versatile starting point for a variety of tasks.

We calculated features for the SPINE1 data suitable for this network (which expects 12th order PLP cepstra without deltas, computed over a 32ms window every 16ms), and performed a forward-pass of the network to calculate the phone posterior estimates for every frame of the data. These were then Viterbi-aligned to the word transcripts of the training data, also based on pronunciations from our Broadcast News system. This generated a first set of aligned phone labels ("Labels Boot0" in the figure) suitable for training a new network based on the SPINE1 data.

This process of forced alignment was repeated twice, to let the labels converge for the new SPINE1 data. The output of this second realignment, "Labels Boot2", was used to train two new networks, based on PLP features including deltas and double-deltas, and MSG features (as described above). These two parallel networks were then applied to the data, and their posterior probability estimates combined by log-domain averaging as the basis for a final stage of realignment. Labels from this final alignment were use to train the final pair of networks, denoted "Net PLP2" and "Net MSG2" in the figure.

In the next section, we report results for two tandem systems.

| Type of feature (dimensionality) | CI models | CD models 2600 senones | MLLR with CD models |
|---|---|---|---|
| MFC with delta and double-delta (39) | 69.5% | 35.1% | 33.5% |
| PLP with delta and double-delta (39) | 71.3% | 38.0% | 35.2% |
| Tandem 1 (54) | 59.1% | 39.4% | 34.5% |
| Tandem 2 (56) | 47.6% | 35.7% | 32.8% |

**Table 1**. Word error rates (%) obtained with the SPHINX-III recognition system for various feature sets. The dimensionality of each feature set is indicated in parentheses. "CI" stands for context-independent, and "CD" stands for context-dependent.

The first (which was submitted the official NRL SPINE1 evaluation) is based on "Net Boot2", the second-generation network based on the Broadcast News-style features and frame rates. This single network was used in a tandem configuration without feature combination to process the entire SPINE1 training set, and these results were used to train a GMM-HMM system.

The second tandem system used two feature streams as illustrated in figure 1 and was based on the nets "Net PLP2" and "Net MSG2". The additional training required for these nets was not completed in time for the SPINE1 evaluation.

In both cases, the SPHINX-III GMM-HMM system was trained via a conventional EM procedure. The full system used 3 state per HMM context-dependent triphone models with 2600 tied states, each modeled by a mixture of 8 Gaussians. The means and variances of the tandem features were not normalized as this was observed to deteriorate the recognition performance. Difference and double difference features were not used by the SPHINX-III system for the tandem features. The models were adapted in an unsupervised manner to the evaluation data after an initial decoding pass, using a single iteration of single-class MLLR. More details of the CMU SPHINX-III system used for the SPINE1 task are given in [9].

## 3. EXPERIMENTAL RESULTS

The CMU SPHINX-III system was trained with the SPINE1 data, which consisted of about 8 hours of recordings. Models were trained for the two different tandem features as well as for standard MFC and PLP features. For the MFC and PLP features cepstral mean normalization was performed and difference and double-difference cepstra were used for recognition. Recognition was performed on the SPINE1 evaluation data, which consisted of about 9 hours of recordings. The word error rates obtained for all the systems are shown in table 1.

We note from table 1 that the word error rates obtained with both tandem systems are much lower than those obtained by the MFC and PLP systems when context-independent models are used for recognition. Specifically, the WER for the Tandem 1 system is 15% lower relative to that of the MFC system, while the WER for the Tandem 2 system is 31% lower. Moreover, for context-independent (CI) models, we observe that the WER of the Tandem 2 system is 19% lower than Tandem 1. This improvement clearly results from the post-classifier combination with the second feature stream, made possible by the tandem structure.

When recognition is performed with context-dependent models the tandem systems are not seen to be better than the MFC-based system. However, the WERs obtained with the Tandem 2 features are lower than those obtained with the PLP features that are used to derive the tandem features. This leads us to conjecture that had the tandem preprocessors used the MFC features, the final tandem systems may have performed better than the MFC-based system. This hypothesis remains to be investigated. The Tandem 2 system is observed to be better than the Tandem 1 system at the context-dependent (CD) stage as well, though the relative improvement observed is lower than that observed with CI models.

The final column in the table shows the word error rates after a single pass of single-class MLLR performed on the CD models using the SPINE1 evaluation data. The relative improvement in the performance of the CD models after MLLR is 4.6% for MFC, 7.4% for PLP, 12.4% for Tandem 1 and 8.1% for Tandem 2.

Unlike the MFC and PLP based systems, the SPHINX-III recognizer was not explicitly optimized for the tandem systems. With such optimization, we hope to improve the word error rates for the tandem systems by an additional 1-2% absolute.

## 4. DISCUSSION

The principal question motivating this work was whether the 50% reduction in WER obtained by tandem modeling in a small-vocabulary task [7] would extend to a larger system.

In the current work, we find that using multistream tandem features on a larger-vocabulary task with a state-of-the-art GMM-HMM recognizer, gives approximately a 31% improvement over baseline MFC and PLP features with the context-independent models. However, this large improvement is mainly eliminated for the context-dependent models. On the other hand, model adaptation using MLLR results in a greater improvement for the tandem features than for the standard MFC and PLP features, making all features roughly comparable for a context-dependent, MLLR-adapted system.

We interpret the discriminatively-trained neural net in the tandem systems as performing a remapping of the feature space that magnifies regions around key phonetic boundaries and compressing regions that correspond to a single phone, minimizing the effects of non-phonetic variations such as speaker characteristics and noise. This soft remapping retains some information from the original signal that the subsequent Gaussian mixture model in the GMM-HMM recognizer can usefully exploit. We believe the gains of the tandem approach arise from the combination of *discriminative* modeling (in this case via the NN) which marshalls parameters to focus on 'critical' regions, and, within the more uniform feature space created by the discriminative models, *distribution* modeling by the GMMs, which are better suited to modeling a large number of classes.

However, in the modeling of context-dependent classes, the advantages gained by the net's feature space remapping appear to be largely nullified. This may be because the NN optimizes the separability between a different (and smaller) set of classes (context-independent phones) than those modeled by the GMM-HMM (context-dependent phones). Thus, while the context-independent phones themselves become more separable, this may be achieved at the cost of increased overlap and confusion of the various context-dependent versions of the same phone. This problem could be reduced if the NN were trained to discriminate between context-dependent units, but constructing a net with such a large number of often-similar outputs is a major challenge. The

resulting probability vectors would also have a very high dimensionality. These problems can, however, be tackled by grouping of context-dependent units, and the usage of dimensionality reduction techniques such as LDA.

MLLR adaptation improves the fit between GMM models and test data. We see that the systems performing less well with unadapted CD models show a greater benefit from MLLR: It is possible that MLLR is able to mitigate the within-class confusion suggested above.

A simpler explanation for the lower performance of CD tandem systems could be that the 8 hours of SPINE1 training data were insufficient to train the large number of parameters needed by CD models for the high-dimensional tandem features.

The adaptation related results are instructive from another perspective: the assumption underlying schemes such as MLLR is that acoustic and speaker variations can be modeled by a low-order transformation of feature space. This works for conventional features, for which noise and pronunciation variations may conceivably be modeled as simple shifts. However, the feature space at the output of the neural network in a tandem system bears a highly complex nonlinear relationship to the original feature space, and we have little understanding of how these features behave when faced with variable input data quality. The current results seem to indicate that the linear-shift assumption is equally applicable to the tandem features.

An important advantage associated with the dual models in the tandem system is that system enhancements specific to each model can be included at the appropriate stage. This has been demonstrated in the current system, which uses posterior combination in the NN stage (something made particularly easy by the compact, context-independent representation), in conjunction with MLLR adaptation and tied mixture weights, schemes developed specifically for GMM-HMM systems. (Mixture weights have been used with success directly on the outputs of a posterior-estimation neural network in [11]. However, by omitting the Gaussian models altogether, that approach would not be able to take advantage of MLLR-style adaptation.)

## 5. CONCLUSIONS

We have shown that the tandem approach of using a combination of neural networks, trained to estimate posterior probabilities of context-independent phone classes, as feature preprocessors for an otherwise unmodified state-of-the-art Gaussian-mixture hidden Markov model speech recognizer can achieve significant and worthwhile reductions in word-error rate when CI models are used. Further work needs to be done to extend the benefits to CD models. Also, the training of the NN components is fundamental to the success of the overall system; the somewhat-involved process for the current task could probably be improved. The gains shown by MLLR adaptation of the network output features suggest a significant systematic variation of features with acoustic or speaker changes; we are interested in better characterizing this variation, perhaps to compensate for it more directly and effectively.

We conclude that for systems based on a relatively small number of subword classes, a category including many commercial applications, tandem-style neural network feature preprocessors are likely to offer considerable advantages. In particular, when the background noise level is high, as with the SPINE1 data, the nets have shown themselves able to effect very significant error reductions in context-independent systems.

## 6. ACKNOWLEDGMENTS

Ellis was supported by the European Union under the ESPRIT LTR project Respite (28149) through the International Computer Science Institute, where much of the work was performed. Singh was supported by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. Sivadas was supported by the DoD under MDA904-98-1-0521, by the NSF under IRI-9712579, and by an industrial grant from Intel Corporation. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

## 7. REFERENCES

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

[2] 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.

[3] 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.

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

[5] 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.

[6] H. Hermansky,D. Ellis and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems,” Proc. ICASSP, Istanbul, June 2000.

[7] 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. ICASSP, Istanbul, II-1117-1120, June 2000.

[8] 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.

[9] R. Singh, M. Seltzer, B. Raj and R. Stern, “Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination,” submitted to ICASSP-2001, Salt Lake City, May 2001.

[10] A. Janin, D. Ellis and N. Morgan, “Multi-stream speech recognition: Ready for prime time?” Proc. Eurospeech-99, II-591-594, Budapest, September 1999.

[11] J. Rottland and G. Rigoll, “Tied posteriors: an approach for effective introduction of context dependency in hybrid NN/HMM LVCSR,” Proc. ICASSP, Istanbul, June 2000.