ERROR VISUALIZATION FOR TANDEM ACOUSTIC MODELING ON THE AURORA TASK

Manuel J. Reyes-Gomez and Daniel P.W. Ellis

Dept. of Electrical Engineering, Columbia University, New York, NY 10027

ABSTRACT

Tandem acoustic modeling consists of taking the outputs of a neural network discriminantly trained to estimate the phone-class posterior probabilities of speech, and using them as the input features of a conventional Gaussian mixture model speech recognizer, thereby employing two acoustic models in tandem. This structure reduces the error rate on the Aurora 2 noisy English digits task by more than 50% compared to the HTK baseline. Even though there are some reasonable hypothesis to explain this improvement, the origins are still unclear. This paper introduces the use of visualization tools for error analysis of some variants of the tandem system. The error behavior is first analyzed using word-level confusion matrices. Posteriorgrams (displays of the variation in time of per-phone posterior probabilities) provide for further analysis. The results corroborate our previous hypothesis that the gains from tandem modeling arise from the very different training and modeling schemes of the two acoustic models.

1. INTRODUCTION

Tandem acoustic modeling combines the conventional Gaussian mixture/hidden Markov model (GMM/HMM) approach with the more unusual hybrid connectionist-HMM approach. After a common feature extraction stage, the traditional GMM/HMM approach associates particular feature vectors with speech units by modeling their distributions under each class with a set of Gaussian mixture models (GMMs). The estimation of the GMM parameters is done according to a maximum-likelihood criteria via the EM algorithm. These distributions are then used to calculate the likelihoods of different speech sounds used in the hidden Markov Model (HMM) decoder. In the hybrid-connectionist approach [1], the GMMs are replaced by a neural network discriminantly trained to estimate the posterior probabilities of the phone classes given the feature observations. These posterior probabilities are then converted into scaled likelihoods before being used in the HMM decoder.

In the tandem acoustic modeling, the outputs of the same discriminantly-trained neural network are used as the input features of a conventional GMM/HMM speech recognizer. The original development of tandem systems showed an improvement in the performance on the Aurora task [2] of these systems compared with the HTK baseline of more than 30% [3]. Further variations of the original system have increased the performance even more [4]. Even though the origin of the improvement is still unclear, some hypothesis were proposed in the previous work.

In this paper we analyze in more depth the performance of Tandem systems particularly in terms of their error behavior. Since the Aurora task includes a large number test conditions composed of different noise types and levels, error analysis is quite complex. We have found visualization tools very useful for this task because they can provide a concise form to analyze several aspects of the system behavior, and thereby reveal important patterns.

To achieve a better understanding of the error behavior, we analyzed several variants of the original tandem system, which are introduced in the next section. The "GMM Posteriors" system helps to clarify the difference between tandem modeling and conventional GMM systems, and the "Standard + deltas + normalization" system has performance significantly improved over the standard tandem system. In section 3, we use confusion matrices and their differences to analyze the variations in the error behavior between these systems. In section 4, we further analyze the behavior of the original tandem system and the GMM Posteriors system using posteriorgrams, visualizations of the posterior probability variations in time. In section 5 we draw conclusions concerning the success of tandem modeling.

2. STANDARD TANDEM SYSTEM AND ITS VARIANTS

Tandem acoustic modeling consists of two acoustic models, a neural net (NN) and a Gaussian mixture model (GMM), in tandem feeding a hidden Markov model (HMM). This system can be divided into three stages:

1. Phone-posteriors estimation (first acoustic model)
2. Posteriors processing, in which the outputs of the first model are conditioned to better suit the second model
3. Conventional GMM/HMM speech recognizer (second acoustic model and HMM decoder).

Figure 1 shows the block diagrams of the four systems compared in this paper. First is the simple baseline (which we refer to as the PLP baseline or "Base"), where the basic 12th order PLP cepstra are fed directly into the GMM/HMM recognizer defined for the Aurora task using the HTK package. (The PLP features are for most purposes very similar to MFCCs). Both of these block are used in all the system variants.

The second system in figure 1 is the original tandem system [3], (hereafter, the standard system or "Std"), in which the phone-posterior estimation module is implemented by a neural network. The second stage consists of the removal of the net's final nonlinearity to approximate the log probabilities and a decorrelation stage implemented with PCA.

To investigate the importance of the neural network in the tandem approach, we investigated a version in which phone-posteriors were calculated by a GMM instead. A phone-based GMM/HMM system was directly trained on the PLP features, including deltas, then the phone posterior probabilities are calculated via Bayes' rule from the class likelihoods modeled by the GMM.

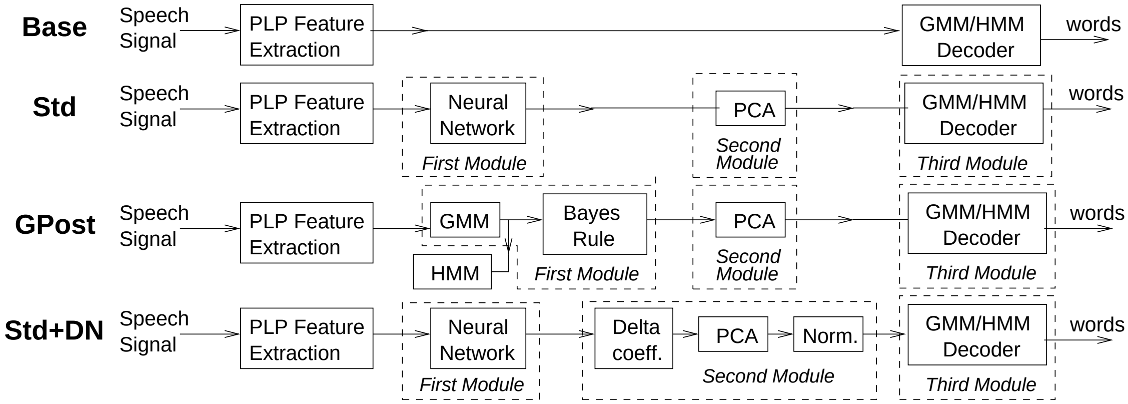

Fig. 1. Block diagrams of the recognizer variants discussed in this paper.

We call this the GMM Posteriors system, or “GPost”. In this case, the second module (posterior processing) again consists of a PCA decorrelation stage.

Finally, we have found that the standard tandem system can be improved by expanding the second module to incorporate delta calculation before decorrelation, and per-utterance feature normalization after the decorrelation to give each dimension zero-mean and unit-variance. This system is known as Standard plus deltas plus normalization, or “Std+DN”.

The performance of these systems on the Aurora 2 task is summarized in table 1. We see that the Base system performs very much on a par with the standard Aurora multicondition reference, whereas the GPost system is slightly worse. The standard tandem system shows an improvement of 30% relative to the Aurora reference, and Std+DN increases this by a significant margin.

	WER %			Ref.
System	test A	test B	test C	impr.
PLP baseline: Base	12.21	14.49	16.14	-2.2
Standard tandem: Std	7.69	11.12	9.98	30.0
GMM posteriors: GPost	13.88	15.25	16.87	-10.8
Std+deltas+norm: Std+DN	6.82	8.32	7.21	44.5

Table 1. System word-error-rates (WERs) for the three Aurora test sets (averaged over 20..0 dB SNR), and the percentage improvement over the Aurora multicondition reference system (“Ref. impr.”, the standard figure of merit for the Aurora task).

3. CONFUSION MATRICES AND DIFFERENCE CONFUSION MATRICES

The first visualization tool we use in our analysis is the confusion matrix. This is obtained by comparing the recognized words with the actual words spoken in each utterance, keeping a record for each word in the vocabulary. The results are then depicted in a graphical form, shown in the top row of figure 2. Within each image, the rows represent the actual words present in the utterances, excepting the last row which represent the number of insertions present in the final transcription. The columns represent the words reported by the recognizer, with the exception of the last column which counts the deletions committed. The leading-diagonal values represent the correctly recognized words and all the off-diagonal values represent different sources of error. Numerical values of the overall average word-error rates for the included conditions are shown superimposed on each matrix.

Confusion matrices can be calculated for different systems over different subsets of the noise conditions and levels defined in the Aurora task; figure 2 shows confusion matrix images for the four systems we are considering. Since recognition performance is near-perfect for high-SNR cases, we focus on the low-SNR cases (SNRs of 5, 0 and -5 dB), and show results averaged over only these cases for every Aurora noise type. (Averaging only within the separate test sets A, B and C revealed intriguing differences, but we have yet to analyze these in detail). We note that the WERs in this low-SNR case are rather large.

For example, in the low-SNR case for the PLP baseline system (top left of figure 2), we see that the significant error contributions come from deletions, (particularly for “six”, less so for “five”, “seven”, “nine” and “oh”) and from insertions of “oh”. The worst confusions come from words misrecognized as “oh” (particularly “four” and “zero”), although these may be accounted for as combined deletions and “oh” insertions. We should mention here that we experimented with changing the word insertion penalty of the HTK decoder in order to balance the number of deletions and insertions, but this only worsened the overall word-error rate.

Comparing the confusion matrices of two systems can give us an insight into the origin of differences between their bottom-line error rates. However, rather than looking at two images, we can simply subtract the matrices, and display the result. This gives us the second row of figure 2, the *confusion differences*, i.e. the differences between the adjacent pairs of confusion matrix images.

Here, darker colors in the leading diagonal mean that the right-hand system has a better performance recognizing that particular word. Darker colors in the off-diagonal values represent more errors (confusions, deletions or insertions depending on the position in the matrix) for the right-hand system. Note that zero difference is indicated by 50% gray in these figures. In all matrix rows other than insertions, increases in correct recognition must balance decreases in error, both logically (since we are describing a fixed set of word examples), and because the sums of the underlying confu-

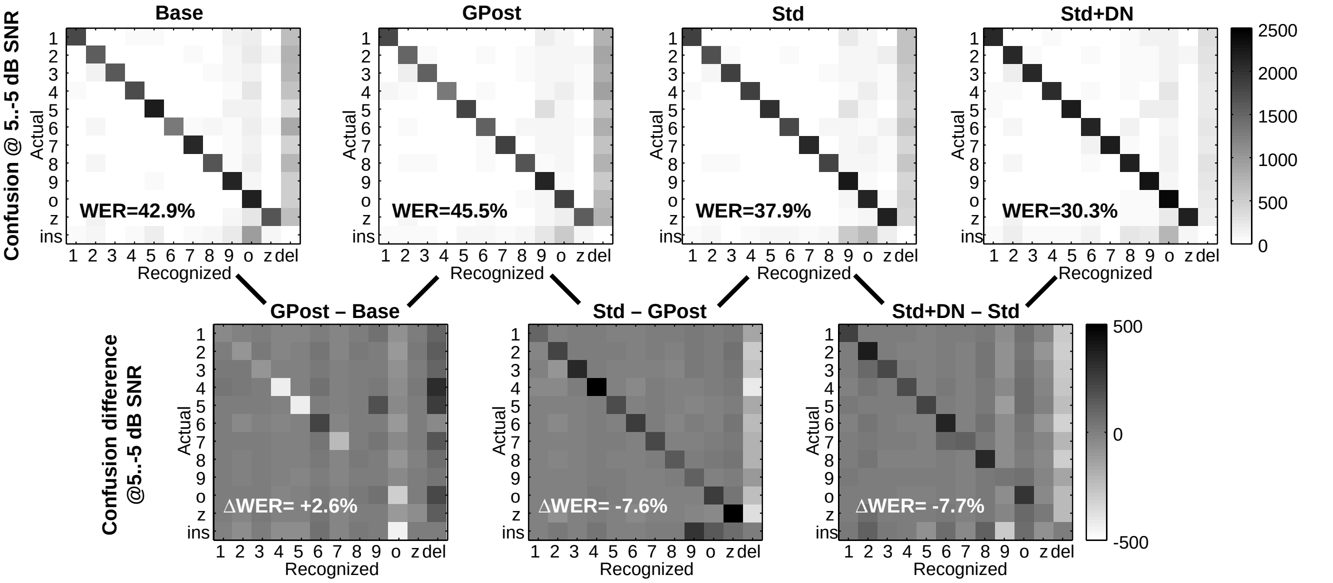

Fig. 2. Confusion matrices for all four systems in the low-SNR condition, and three confusion differences showing the differences between the adjacent matrices in the top row.

sion matrix rows are equal.

Looking at the first panel in the lower row of figure 2, the difference between the GMM posterior system and the PLP baseline, we notice that the performance of these two systems is quite similar (as table 1 corroborates) although the GPost system shows significantly great numbers of deletions for “four”, “five”, “seven” “oh” and “zero”, fewer insertions of “oh”, and rather more “five”-to-“nine” confusions. Even though the GPost system has a tandem structure, we interpret its performance similarity to the PLP baseline as indicating that the second GMM (within the HTK recognizer) does not have anything new to learn from the posteriors generated from the first GMM. (This was in fact our original expectation of *all* tandem systems.)

The confusion difference between the GPost system and the Standard system over the low SNR conditions (middle panel) reveals that the better performance of the Standard system is due primarily to a smaller number of deletions (lighter grays in the final column). We can also observe that the standard system has a *larger* number of insertions of the word “nine” i.e. it is worse than the GPost system, and by extension the Base system, in this particular aspect.

When comparing the performance of the standard system and the Std+DN for low SNR signals, we can observe that the inclusion of deltas and normalization increases the performance of the latter system by further diminishing the number of deletions and largely eliminating the insertions of the word “nine”. After looking at a range of confusion statistics, this emerged as a clear trend: the NN-based systems were particularly vulnerable to “nine” insertions, but including the deltas of the tandem posteriors as part of the feature vector corrected this. Seemingly, the dynamics of the transition between the /n/ and /ay/ posteriors was additionally characteristic for this word.

4. POSTERIORGRAMS

Since the GPost and Std systems differ only in the basis for phone-posterior estimation (distribution GMMs or discriminant NN), the differences summarized in the middle confusion difference image can be further examined by comparing the posteriors themselves. The phone-posteriors estimation module generates 24 posterior probabilities for each frame, and these can be rendered as rows in an image to give a visualization reminiscent of a spectrogram–in that it shows intensity variation over time–but for a range of phone class probabilities rather than frequency band energies. We call this representation a *posteriorgram*, and some examples are shown in figure 3. Ideally, at each frame, the posterior that corresponds to the phone being spoken should be the one with the highest value (darkest). The pink boxes (which appear gray in monochrome) outline the ‘true’ targets derived by forced alignment.

In the figure, the GMM-based posteriors are shown alongside those from the neural network for the same utterance both with and without added noise. The posteriors obtained from the neural network show a closer resemblance to the targets, especially in the time interval 0.5-1.0 sec. More importantly, in both signals they have a lower entropy (less ambiguity) than the GMM posteriors. This difference in posterior entropy is not by chance, but is a direct result of the discriminant training of the neural network, which updates its weights to assure an increment of the correct posterior probability while decreasing the incorrect ones. On other hand, the GMM learns its parameters by maximizing the likelihood of the correct phone class but without assuring the minimization of the remainder. The smaller entropy in the Std system explains its lower number of confusions in comparison with the GPost system.

Notice than in contrast to the clean case and in contrast to the NN, the GMM posteriorgram for the noisy signal shows a significant, spurious shift of probability to the background phone (sil). We consider this failure of the GMM in discriminating between noisy speech and background to be reflected in the high deletion

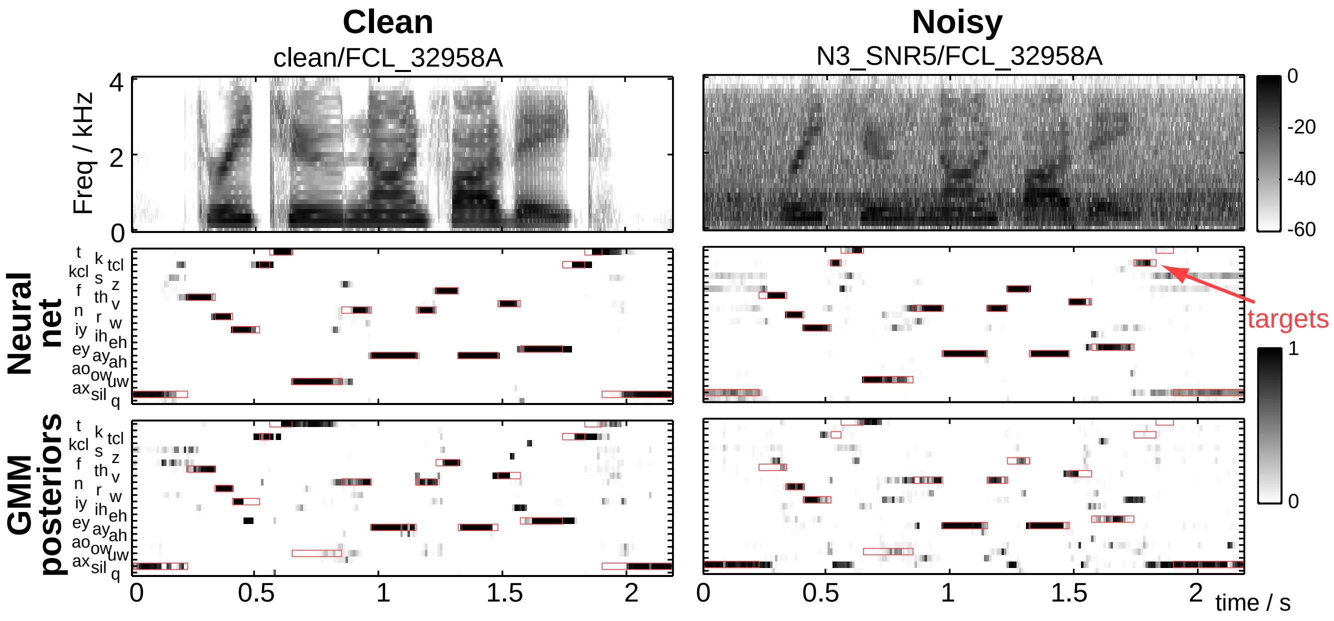

Fig. 3. Posteriorgrams of the same utterance in clean and noisy conditions, based on both neural net and GMM.

rate for the GPost system seen in the confusion matrices. Moreover, as the Base system behaves rather like the GPost, and since they are based on similar GMM distribution models, we speculate that the deletions plaguing the Base system at low SNR have a similar origin.

Other features visible in the posteriorgrams are the confusion of /sil/ as /f/ or /s/, both for GMM in clean, and for the NN in noise, although for the GMM in noise, the overall bias to /sil/ has corrected this.

The most prominent difference between target and actual posteriors is the near complete failure of the GMM posteriors to register the /uw/ from “two” around 0.8 sec. This may reflect the nature of the EM training used for the neural network, which uses the label-sequence constraints on the training data however it sees fit. In this case, because /uw/ occurs only in “two”, and because /t/ must be shared only with “eight”, the training has apparently incorporated the vowel from “two” at the end of the /t/ model. It would be interesting to investigate the wider effects of such phenomena.

5. CONCLUSIONS

We have previously concluded [4] that the most likely explanation for the success of tandem modeling lies in the complementary nature of the two models, both in terms of representation (hidden units describing class boundaries versus Gaussians covering distribution peaks) and in terms of training scheme (back-propagation to Viterbi targets versus full EM). The differences in the posteriorgrams show how the two acoustic models can learn different aspects of the acoustic information. It seems that the global remapping of the feature space performed by the neural network does not eliminate the information to be learned by the GMM, and that moreover it permits the GMM to work in a more successful distribution-modeling space, as evidenced by the performance improvement in general, and by specific improvements such as the reduction of certain deletions, as shown in our current analyses.

Prior to this investigation of error visualization tools, our insight into the differences between the various systems was limited to comparing the monotonic changes in WER as a function of noise level. Confusion matrix images and posteriorgrams have revealed considerably more specific information concerning the strengths and weaknesses of different modeling techniques. Curiosities such as the interaction of posterior-deltas and the insertion of “nine” have a remedial value in developing improved systems.

6. ACKNOWLEDGMENTS

This work was supported in part by the European Commission under LTR project RESPITE (28149).

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] H.G. Hirsch and D. Pearce, “The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions,” *ISCA ITRW ASR2000*, Paris, September 2000.

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

[4] D. Ellis and M. Reyes, “Investigations into tandem acoustic modeling for the Aurora task,” *Proc. Eurospeech*, Aalborg, September 2001.