Investigations into Tandem Acoustic Modeling for the Aurora Task
Dan Ellis & Manuel Reyes • Columbia University • {dpwe,mjr59}@ee.columbia.edu

Summary: In Tandem acoustic modeling, a neural net classifier and a Gaussian mixture distribution model are used in series to improve recognition accuracy. Here, we investigate several possible sources for this improvement. We conclude that the net and GMM are simply complementary classifiers.

Introduction
- Tandem modeling arose from an effort to find a way to use neural-network acoustic classifiers with a GMM-based HTK back-end.
- A neural network is trained (discriminantly) to estimate posterior probabilities of each subword unit.
- These posterior probabilities are lightly conditioned then used as input features to train a standard GMM-HMM via EM.
- Using the neural net allows multiple feature streams to be used via posterior combination.
- This combination of approaches led to 50% reduction in WER on the 1999 Aurora task.
- Our baseline result (single PLP feature stream) achieves an average improvement of 30% over the HTK reference system for the Aurora 2000 multicondition task.

Different training data
- Since there are two acoustic models in a Tandem system, there is a question of whether to use the same or different data to train each one. We want to train the GMM to learn the behavior of the neural net on unseen data.
- Given a finite training set, we have to split the set into two halves called T1 and T2. New model has half as much training data, which will impair its performance.
- Using different halves for each training (T1:T2) is marginally better than training both models on the same half-set (T1:T1). However, both are significantly worse than the baseline system which uses the (same) whole training set for both trainings.
- We conclude that the net and GMM trainings are extracting different complementary information from the same training data.

Why does Tandem processing help?
- The neural network is trained discriminantly, for better modeling of class boundaries.
- The GMM system models class-conditional distributions, but is trained via full EM, leading to more appropriate subword units.
- The tandem configuration appears to combine these advantages, but how?
- Is it better to use different training data to train the two models (net and GMM)?
- Is it important that the net uses phone units while the GMM has whole-word models?
- How well would it work to use a GMM in place of the neural net to estimate phone posterior probabilities (via Bayes’ rule)?
- In addition to the PCA orthogonalization, is there other conditioning that we can usefully apply to the net outputs?

Different subword units
- The original Aurora system used 24 context-independent phones as the network output targets and 181 whole-word models in the GMM-HMM system.
- To test if this was a source of modeling benefits, we conducted several variants:
  - using 24 phonemes in both neural net and GMM-HMM (24-24)
  - training the neural net to estimate posteriors of all 181 whole-word substates (181-181)
- Because the 181 output net gave a very large input feature vector for the GMM, we also tried:
  - rank reduction of the 181 state posterior to the top 40 PCA dimensions (181-40)
- All variants performed similarly to the baseline, indicating that subword unit choice is not a significant factor in Tandem performance.
- Curiously, the 181-output net performs significantly better on test B (unmatched noise).

Adding multiple feature streams
- The best tandem systems exploit the phone-posterior representation to combine two or more feature streams via log-averaging.
- All the systems described so far use only a single stream of plp features.
- When we tried to add a stream of msg features to our best system from above, the improvements did not carry through. The second feature stream’s posteriors behave differently, particularly under delta calculation.
- Our best results came from simply normalizing the combined post-PLP PCA net outputs.