Investigations into Tandem Acoustic Modeling for the Aurora Task

Dan Ellis & Manuel Reyes • Columbia University • {dpwe,mjr59}@ee.columbia.edu

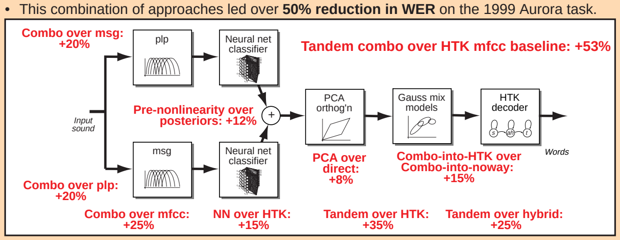
Columbia University

Lab ROSA

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 over **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:

System	*Accuracy%, 20-0 dB*			*Rel. improvement %*			***Avg.***
	test A	*test B*	*test C*	*test A*	*test B*	*test C*	***impr.%***
Baseline PLP Tandem	92.3	88.9	90.0	36.8	19.0	38.5	**30.0**

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, Iis there **other conditioning** that we can usefully apply to the net outputs?

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. Now each 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.

System	*Accuracy%, 20-0 dB*			*Rel. improvement %*			***Avg.***
	test A	*test B*	*test C*	*test A*	*test B*	*test C*	***impr.%***
Same half T1:T1	91.4	88.3	88.9	29.7	14.7	31.8	**24.2**
Different halves T1:T2	91.5	88.7	89.2	29.8	17.9	33.6	**25.9**

Varying the subword units

- The original Aurora system used 24 **context independent phones** as the network output targets and 181 **whole-word model states** in the GMM-HMM system.
- To test if this was a source of modeling benefits, we constructed 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 posteriors 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** (mismatched noise).

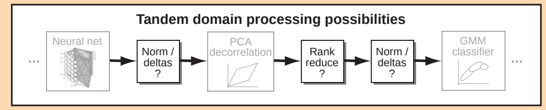
System	*Accuracy%, 20-0 dB*			*Rel. improvement %*			***Avg.***
	test A	*test B*	*test C*	*test A*	*test B*	*test C*	***impr.%***
Phone:Phone 24:24	92.2	88.3	89.5	35.6	14.4	35.3	**27.0**
WW:WW 181:181	91.5	90.1	90.2	30.2	27.7	39.3	**31.3**
WW reduced 181:40	91.9	90.2	90.4	33.8	28.3	40.8	**33.2**

GMM-derived posteriors

- To test the importance of the neural net, we trained a **GMM** directly on the PLP features, then used **Bayes' rule** to convert the likelihoods of each phone state into **posterior probabilities**.
- Substituting these for the net outputs gave a kind of Tandem system where both acoustic models are GMMs. However, it performed **worse than the single GMM** HTK baseline.

Tandem domain processing

- The standard Tandem setup uses **PCA orthogonalization** on the outputs of the neural network before feeding them to the GMM. This gives about 8% relative improvement.
- We were curious to know if **other kinds of processing** at this point might give further improvements. Good candidates include common feature processing such as **rank reduction** (helpful in the subword unit experiments), **normalization,** and **delta** calculation.



- We experimented with a large number of configurations, summarized in the results below:
  - **Rank reduction** on the original 24 dimension feature vector does not help; even dropping to 21 dimensions hurts performance (Top 21 PCA)
  - **Delta calculation** helps most when applied after the PCA rotation (PCA+deltas)
  - **Normalization** (non-causal per-dimension mean and variance normalization in this case) helps most when applied after PCA (PCA+normalize)
  - The best configuration we found for using both deltas and normalization is to calculate **deltas before the PCA**, then **normalize after the PCA** (deltas+PCA+normalize)

System	*Accuracy%, 20-0 dB*			*Rel. improvement %*			***Avg.***
	test A	*test B*	*test C*	*test A*	*test B*	*test C*	***impr.%***
Top 21 PCA dim'ns	92.2	88.8	89.8	35.6	18.5	37.0	**29.0**
PCA+deltas	92.6	90.4	91.1	39.4	30.1	45.0	**37.0**
PCA+normalize	93.0	91.0	92.4	42.4	34.4	53.1	**41.7**
deltas+PCA+normalize	93.2	91.7	92.8	44.0	39.4	55.5	**44.9**

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-PCA** net outputs:

System	*Accuracy%, 20-0 dB*			*Rel. improvement %*			***Avg.***
	test A	*test B*	*test C*	*test A*	*test B*	*test C*	***impr.%***
plp-msg Tdm baseline	93.2	91.4	91.9	44.4	37.1	50.0	**42.8**
plp-msg w/ PCA+norm	93.8	92.1	93.7	48.8	42.3	61.3	**49.1**

for Eurospeech Aurora ESE, Aalborg • 2001-08-29 dpwe@ee.columbia.edu