
Tandem acoustic modeling: Neural nets for mainstream ASR?

Dan Ellis
International Computer Science Institute
Berkeley CA
dpwe@icsi.berkeley.edu

Outline

- 1 Tandem acoustic modeling
- 2 Inside Tandem systems:
What's going on?
- 3 Future directions



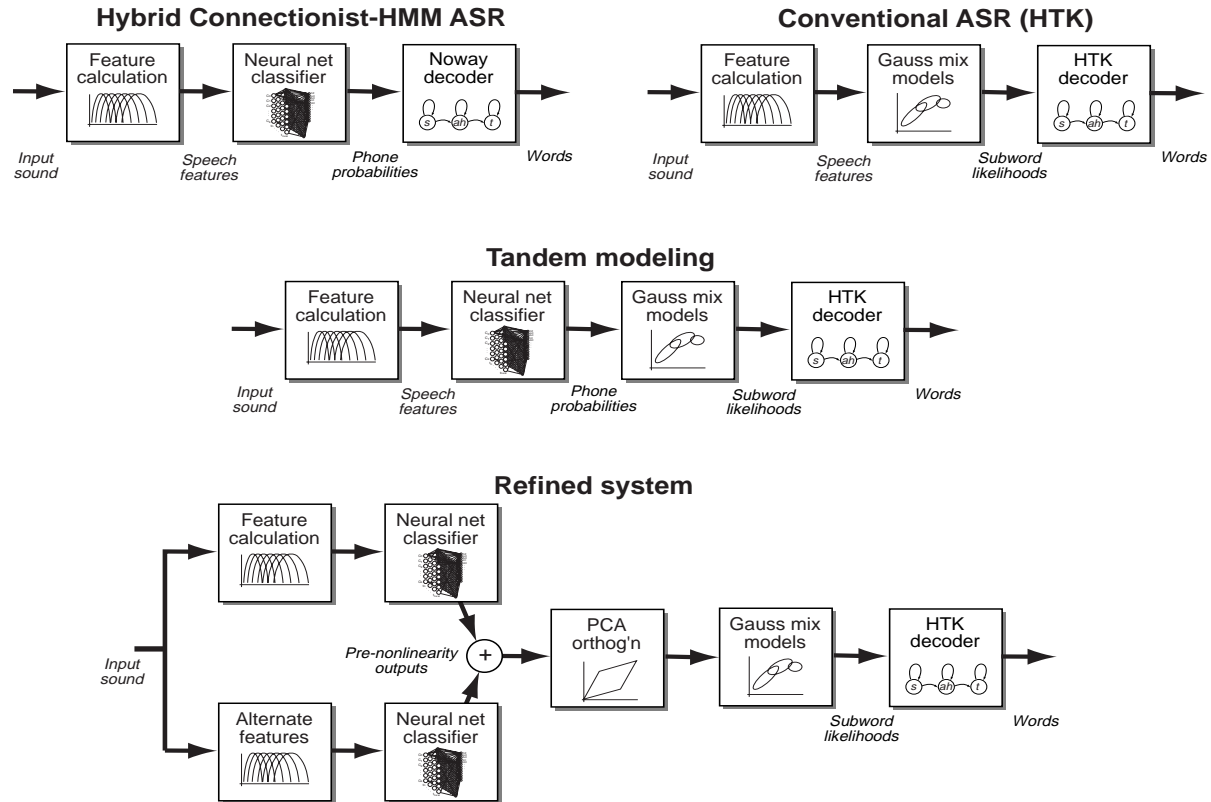
1

Tandem acoustic modeling

- **ETSI Aurora ‘noisy digits’ evaluation**
 - new features (for distributed speech recognition)
 - Gaussian mixture HTK back-end provided
- **How to use hybrid-connectionist tricks?
(multistream posterior combination etc.)**
 - **Use posterior *outputs*
as *features* for HTK...**
 - = ***Tandem* connection of two large statistical
models:
Neural Net (NN) and Gaussian Mixture (GMM)**



The Tandem structure



- **Better results when posteriors are made more 'Gaussian'**
- **Tandem allows posterior combination for HTK**



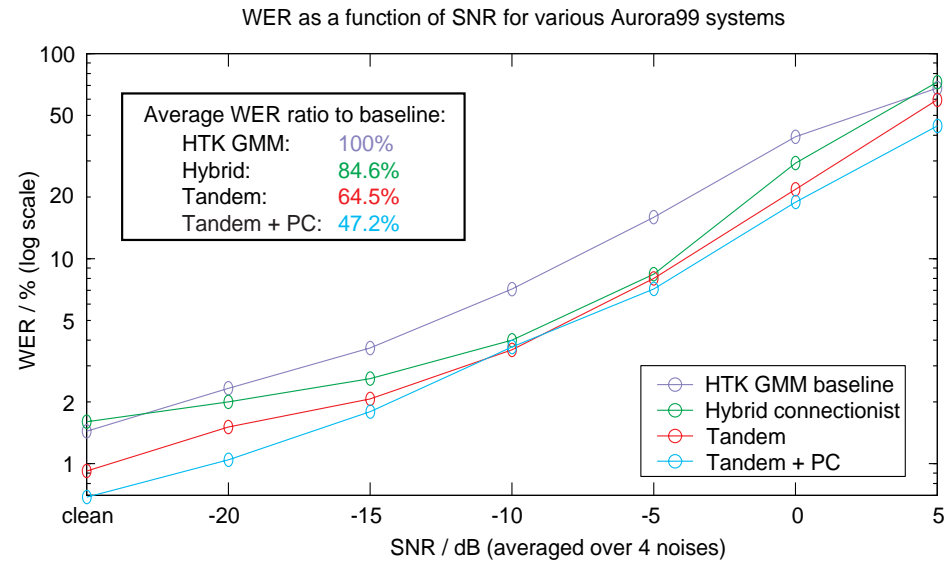
Training a tandem model

- **Tandem modeling uses two feature-spaces**
 - NN estimates phone posteriors (discriminant)
 - GMM models subword likelihoods (distributions)
- **Training procedure**
 - NN trained (backprop) on base features to forced-alignment phone targets
 - GMM trained on modified NN outputs via EM to maximise subword model likelihoods
 - HTK backend knows *nothing* of phone models
- **Decoupled (good) but sequential**
- **Training sets?**
 - can use same for both - learning different info
 - could use different - for cross-task robustness



Tandem system results

- It works very well:



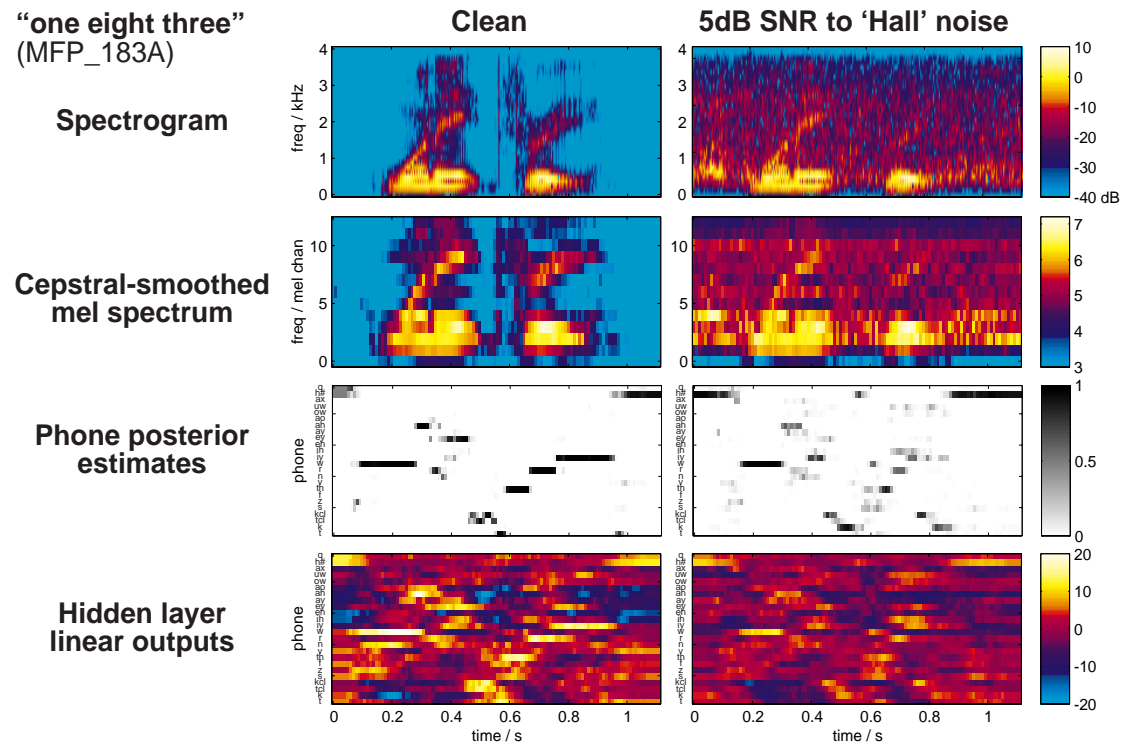
<i>System-features</i>	<i>Avg. WER 20-0 dB</i>	<i>Baseline WER ratio</i>
HTK-mfcc	13.7%	100%
Neural net-mfcc	9.3%	84.5%
Tandem-mfcc	7.4%	64.5%
Tandem-msg+plp	6.4%	47.2%



2

Inside Tandem systems: What's going on?

- Visualizations of the net outputs

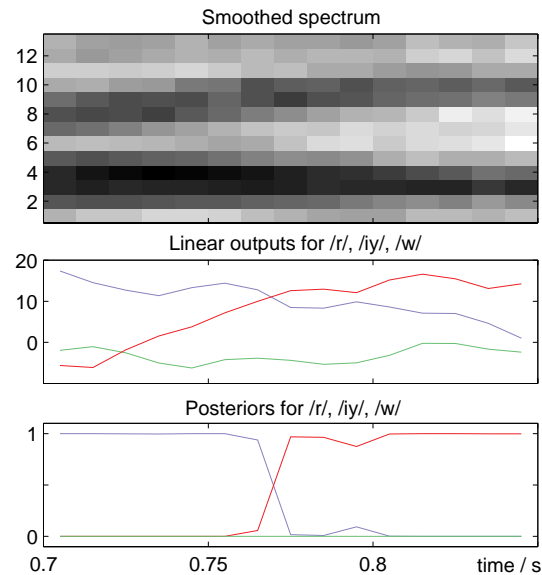


- Neural net normalizes away noise



Feature space 'magnification'

- **Neural net performs a nonlinear remapping of the feature space**

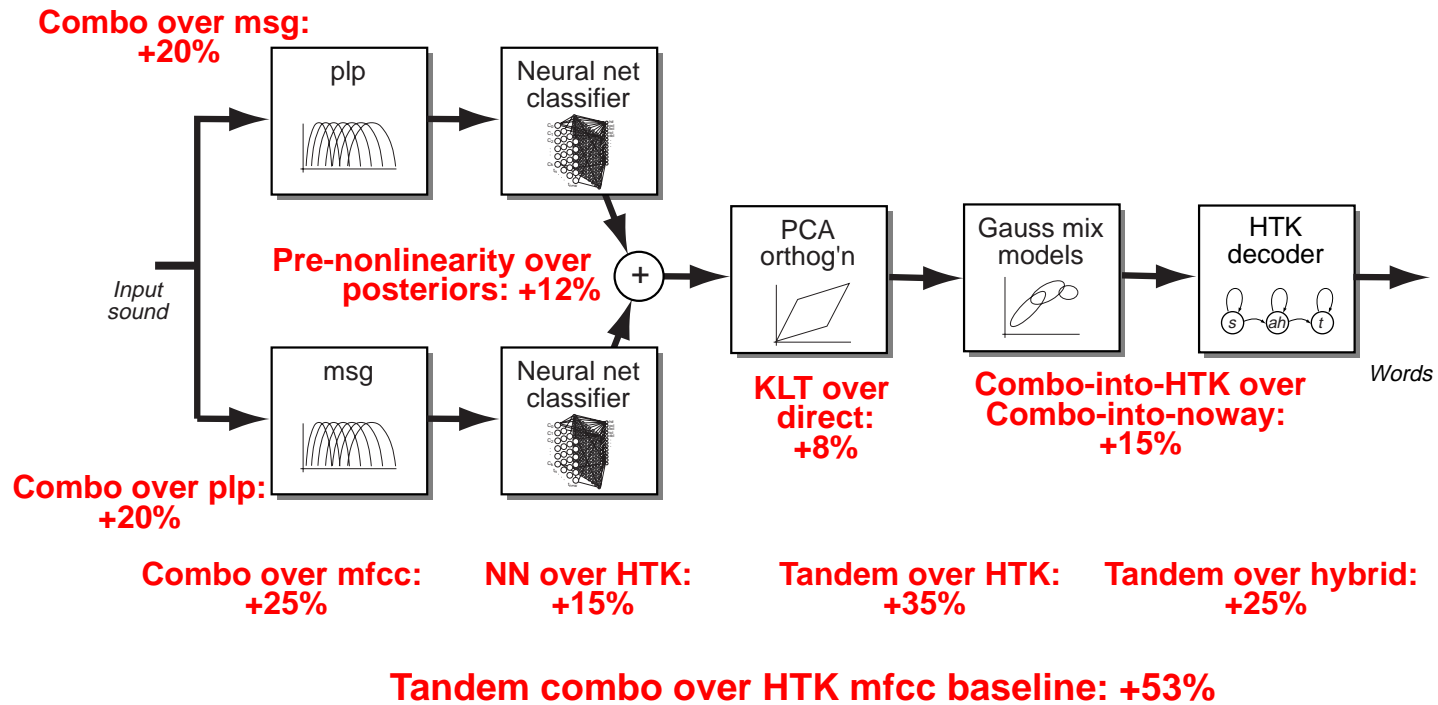


- small changes across critical boundaries result in large output changes



Relative contributions

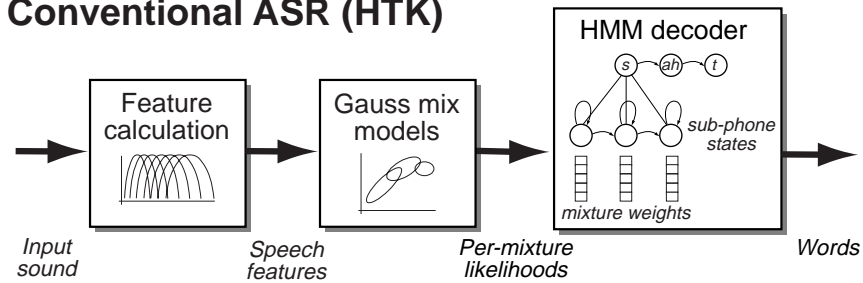
- Approx relative impact on baseline WER ratio for different component:



Omitting the GMMs

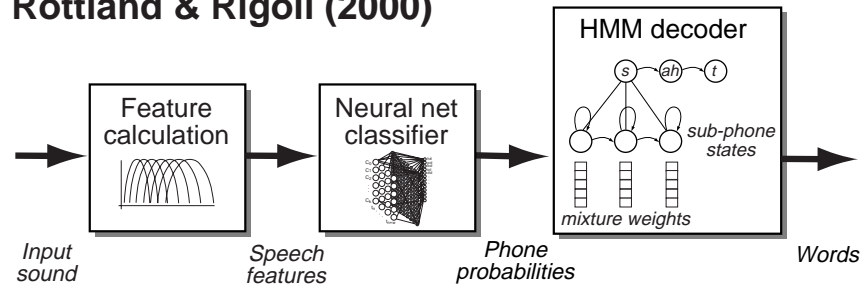
- “Tied posteriors” (Rottland & Rigoll, ICASSP):

Conventional ASR (HTK)



- EM training of GMM and HMM mixture weights

Rottland & Rigoll (2000)



- only mixture weights trained by EM

System	WSJ0 WER
Hybrid baseline	15.8%
“Tied posteriors”	9.4%



3

Discussion

- **Key limitation: task-specific**
 - NN is *not* like features
(it's part of the trained system)
- **Aurora1999 was a 'matched condition' task**
 - same noises added in training and test
 - Aurora2000 has mismatched conditions
 - Tandem modeling works just as well
- **How to relax specificity?**
 - train on alternative task?
 - use articulatory targets



Future developments

- **How to optimize NN for this structure?**
 - integrated training...previous work
 - HMM states as targets?
- **Understanding the gains**
 - better analysis of each piece's contribution
 - strengths of different modeling approaches
 - effects of model/training set size variation
 - "tied posteriors"?
- **Other speech corpora**
 - need both NN and GMM systems...
 - Switchboard is next goal

