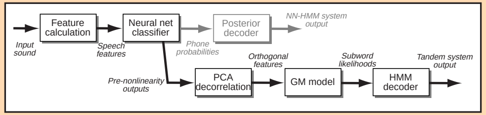
# Tandem acoustic modeling in large-vocabulary recognition

Dan Ellis • Columbia University & ICSI • dpwe@ee.columbia.edu
Rita Singh • Carnegie Mellon University • rsingh@cs.cmu.edu
Sunil Sivadas • Oregon Graduate Institute • sunil@ece.ogi.edu

**Summary:** In tandem acoustic modeling, classification is performed by a neural net followed by a Gaussian mixture model, achieving dramatic improvements on small-vocabulary tasks. For the larger SPINE1 task, much of the benefit disappears when used with context-dependent modeling and MLLR adaptation.

## Introduction

- **Tandem acoustic modeling** refers to using the outputs of a discriminantly-trained **neural network** as the inputs to a conventional **GMM-HMM speech recognizer.** Two acoustic models, neural net and Gaussian mixture, are thus used in tandem:




- When working with the **ETSI Aurora noisy digits task**, the tandem architecture, in conjunction with posterior-level feature stream combination facilitated **WER reductions of over 50%**:

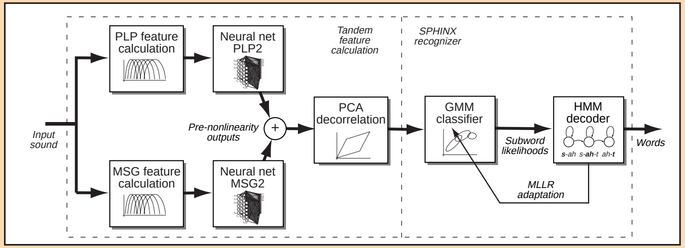
| **Aurora results** | *WER% / SNR* | | | *WER* |
|---|---|---|---|---|
| *Feature* | *Clean* | *15 dB* | *5 dB* | *ratio%* |
| GMM MFC baseline | 1.4 | 3.7 | 15.9 | 100.0 |
| NN MFC baseline | 1.6 | 2.6 | 8.7 | 84.6 |
| Tandem MFC | 0.9 | 2.1 | 8.0 | 64.5 |
| Tandem PLP+MSG | 0.7 | 1.5 | 7.2 | **47.2** |

- We wanted to see if these kinds of improvements could be extended to tasks involving larger vocabularies and more speech variation. We therefore applied the same techniques to the SPINE1 task.

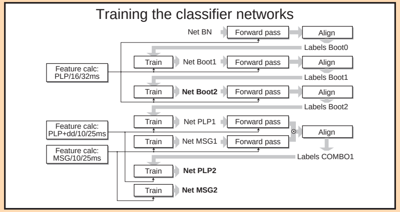
## The SPINE1 task

- The first Speech In Noisy Environments task (SPINE1) was defined by the Naval Research Laboratory (NRL). An evaluation was conducted in August 2000.
- The SPINE1 task consists of **dialogs** between speakers in separate booths engaged in a game of 'Battleships'. Various pre-recorded **noises** are played in the booths to simulate real-world conditions.
- The task has a vocabulary of about 5,000 words, with **natural and informal** grammar and pronunciation.
- About 8 hours of transcribed training material, in a range of background noise conditions, was made available.
- This task is very **challenging**: In the evaluation, the best performance (from a combination of systems) was around 26% WER.

## The Tandem SPINE Recognizer



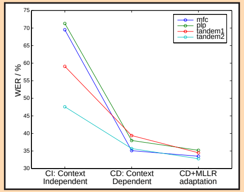

- The **tandem** system consists of a neural net discriminant classifier for context-independent phones followed by a GMM-HMM recognizer
- The neural net system uses two parallel streams based on **different feature representations.**
- Combining conventional PLP features with the more 'sluggish' MSG features gives consistent performance improvements.
- Posterior probabilities estimated by the neural-net classifiers are efficiently **combined** by omitting the net's final nonlinearity and **summing** the output layer activations.
- **Decorrelation** by full-rank Principal Component Analysis improves performance by about 15% relative, presumably because it is a **better fit** to the GM model.
- The output of the neural networks and post-processing is fed as **input into a GMM-HMM** recognizer – the CMU SPHINX-III system.
- The recognizer has no prior knowledge of the specific form of the input features i.e. it is an **unmodified recognizer**, with the net outputs used as features
- The GM model can employ **context-dependent** modeling and MLLR-style **adaptation,** enhancements not normally possible in a neural net system.
- We used CMU's SPINE1 setup, optimized for MFC features, with 2600 context-dependent senones and a single iteration of one-class MLLR adaptation




## Training

- Tandem modeling first **trains a discriminant network**, then separately trains a GMM system on network outputs.
- Network trainings are based on earlier **forced alignments** to context-independent phone labels (Viterbi training).
- * Starting from a Broadcast News net, we trained networks based on two feature streams for the new SPINE task.
- The SPHINX GMM-HMM system was then trained via **conventional EM** on the outputs of the networks as if they were normal features.

## Results




- We compared 4 feature sets:

**mfc** - standard MFC features
**plp** - comparable PLP features
**tandem1** - Tandem based on PLP
**tandem2** - Tandem with PLP+MSG

in 3 HMM model conditions:

**CI** - 39 context-indep. phone states
**CD** - 2600 context-dep. senone states
**CD+MLLR** - added MLLR adaptation

- For the **Context Independent** models, the tandem2 features reduced the baseline WER by 31%.
- Moving to **Context Dependent** models effects much larger improvements on the regular features (mfc, plp) than on the tandem features, bringing all results close together.
- Adding **MLLR adaptation** benefits the tandem systems slightly more, making the tandem2 system the best by a small margin.

## Discussion

- Neural nets (discriminant) followed by GMMs (distribution models) work well for modeling **context-independent phones** even for natural, unconstrained speech.
- Tandem features **interact poorly with context-dependent** state models. Perhaps the context-independent network outputs are confounding the contextual cues within each class.
- **MLLR benefits tandem** CD systems more than conventional features: contextual information may be more variable (but still present) in tandem features.

## Future work

- Would a larger set of **context-dependent discriminant classes** (perhaps a factored network) work better?
- How does performance depend on **training set** size? Should the nets and GMMs be trained on separate data?
- What is the effect of additional processing (normalization, deltas) in the **posterior-features domain**?
- Would it help to **train the net** to a more directly relevant criterion?

for ICASSP-2001 Salt Lake City • 2001-05-05 dpwe@ee.columbia.edu