Tandem connectionist feature extraction for conventional HMM systems

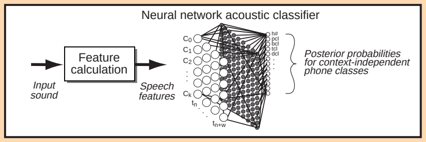
Dan Ellis • International Computer Science Institute, Berkeley • dpwe@icsi.berkeley.edu
Hynek Hermansky & Sangita Sharma* • Oregon Graduate Institute • {hynek,sangita}@ece.ogi.edu

Summary: In hybrid connectionist-HMM speech recognition, the acoustic classifier is a neural network, rather than Gaussian mixture models (GMMs).
Here, we use the outputs of such a network as feature inputs to a conventional GMM-based recognizer, obtaining >30% relative error rate reduction.

OREGON GRADUATE INSTITUTE
OF SCIENCE AND TECHNOLOGY

Introduction

- In a **hybrid connectionist-HMM** speech recognizer, a neural net estimates the posterior probability of a context-independent phone label $p(q_k|X)$, given a window of acoustic features, X:


Neural network acoustic classifier
Input sound
Feature calculation
Speech features
C_0
C_1
C_2
C_k
t_n
t_{n+w}
h#
pcl
bcl
tcl
dcl
Posterior probabilities for context-independent phone classes

- Neural network classifiers have several attractions. The nets are:

  discriminant - the output is trained to choose between phones

  able to learn correlated features and strange distributions - since no distribution assumptions are made

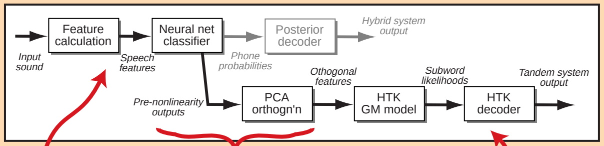
  However, because of the opacity of the model represented in the weights, certain operations (such as adaptation) can be harder than with the more common Gaussian mixture models (GMMs).

- For the ETSI DSR **"Aurora"** evaluation (TIDIGITS with different noises added at various SNRs), we were researching both hybrid recognizers (using our in-house systems) and conventional GMM-based recognizers (using the standard HTK toolkit).
- We wanted to transfer some of the strengths of the connectionist system to the HTK setup, so we tried using the **phone posterior probability estimates as input features** for our HTK training.

Training procedure

- The **tandem** architecture is so named because it uses two statistical models - neural network and Gaussian mixture - in series.
- First, the **neural network** is trained by back-propagation, using maximum cross-entropy, against forced-aligned phone targets.
- Next, the **Gaussian mixtures** are trained by EM to relearn the relationship between the phone estimates and the utterances.
- The HTK GMM system is not informed about the phones used in the neural net training; it independently learns the appropriate patterns.

The Tandem architecture



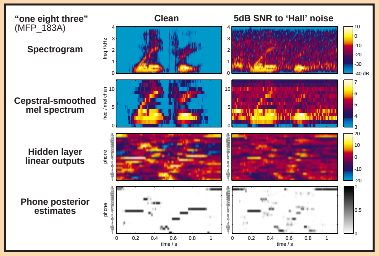
- We start with our normal **hybrid connectionist-HMM system**, trained to estimate posterior probabilities for each of the 24 phones present in TIDIGITS.
- Rather than using this probability stream as input to a posterior-based HMM decoder, we treat the outputs as **features for HTK**.

- **Preprocessing** the posteriors before passing them to HTK greatly improved the results:
- Firstly, we take the network's **linear hidden-layer outputs** before the final 'softmax' nonlinearity. These have a more Gaussian distribution than the very skewed posterior probabilities.
- Secondly, we apply **Principal Component Analysis** (full-rank) to orthogonalize the feature dimensions.

- The features are passed to the **standard HTK recognizer** defined for this task by ETSI.
- The features are modeled by Gaussian mixtures for each of 16 states in 11 whole-word models (1-9, "zero" & "oh").

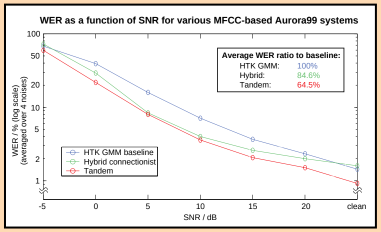
Visualization

- Comparing the spectrum, MFCC feature basis, network linear outputs and posterior estimates for clean and noisy examples illustrates the relative **robustness of the network outputs** to high noise levels.



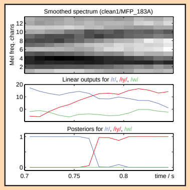
Results

- The ETSI Aurora task defines 4 noise types and 7 SNR levels for a 28 condition test set. (Training has 5 SNR levels.)
- Average word-error-rates (WERs) by SNR show a **consistent 30-40% relative WER advantage** of using tandem modeling.



Why should it work?

- As shown to the right, the neural net does a good job of **magnifying** smooth changes as features cross critical class boundaries.
- Nets model discriminant posteriors via nonlinear weighted sums; GMMs model distributions with parameterized kernels. These very different approaches can extract **complementary information** even from limited training data.



Discussion & Future work

- Feeding posteriors into HTK also allowed us to use **posterior combination** of four feature streams to achieve **under 40% of baseline WER** (see Sharma et al. "Feature extraction .. Aurora database").
- Unlike conventional features, the net is **highly task and language specific**. Training to articulatory targets on a large corpus might help.
- We need to investigate how well GMM techniques such as **MLLR** will work with nonlinearly-transformed features.
- Phone targets were a somewhat arbitrary choice for network training. Targets related to the **HMM states** used in the HTK model might improve results still further.

Related work

- Several researchers have previously investigated neural nets as feature preprocessors in speech recognition. See for example:

Y.R. Bengio, R. De Mori, G. Flammia and R. Kompe, "Global optimization of a neural-hidden markov model hybrid," IEEE Trans. on Neural Networks, 3:252-258, 1992.

V. Fontaine, C. Ris and J.M. Boite, "Nonlinear Discriminant Analysis for improved speech recognition", Proc. Eurospeech-97, Rhodes, 4:2071-2074, 1997.

G. Rigoll and D. Willett, "A NN/HMM hybrid for continuous speech recognition with a discriminant nonlinear feature extraction," Proc. ICASSP-98, Seattle, April 1998.

Acknowledgements

Work at ICSI was supported by the European Union under ESPRIT project RESPITE (28149). Work at OGI was supported by DoD under MDA904-98-1-0521, NSF under IRI-9712579, and by an industrial grant from Intel Corporation.

[*] Sangita Sharma is now with Intel Corporation

for ICASSP-2000 Istanbul • 2000-05-30 dpwe@icsi.berkeley.edu