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

Introduction

• In a hybrid connectionist-HMM speech recognizer, a neural net estimates the posterior probability of a context-independent phone label \( p(t|X) \), given a window of acoustic features \( X \).

• 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).

• 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.

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.

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.

• The neural network outputs are trained to choose between phones even from high noise levels.

• The HTK GMM system is not informed about the phones used in the relationship between the phone estimates and the utterances.

Discussion & Future work

• Feeding posterior 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:

Acknowledgments

Work at ISI was supported by the European Union under ESPRIT project RESPIRE (28143). Work at OGI was supported by DOD under MDA904-96-1-0211, NSF under IRI-9712572, and by an industrial grant from Intel Corporation.

© ICASSP 2000 Tutorial 2000-05-30 dpwe@icsi.berkeley.edu