A General ANN Extension for HTK

Natural Speech Technology

Chao Zhang and Phil Woodland

Cambridge University Engineering Department

UNIVERSITY OF CAMBRIDGE

Abstract

- HTK-ANN enables ANNs with a general structure for acoustic modelling and feature extraction in HTK.
- Include recent ANN techniques, e.g., sequence training, stacking, speaker adaptation, and parameterised activation functions.
- Fully integrated to HTK, to reuse existing GMM-HMM methods for ANN-HMMs.
- HTK-ANN has been tested at CUED on data sets ranging from 3 to 1,000 hours and will be released as part of HTK V3.5 in 2015.

Design Principles

- To accommodate new models and methods, HTK-ANN should be designed as generic as possible
  - Flexible input feature configurations.
  - Generic ANN model architectures.
- HTK-ANN should be compatible with existing HTK functions
  - To minimise the effort to reuse previous source codes and tools.
  - To simplify the transfer of many technologies.
- HTK-ANN should be "research friendly".

ANN Training Facilities

- HTK-ANN has both frame level (CE, MMSE) and sequence level (MMI, MPE) training criteria.
- ANN labels come from frame-to-label alignment (CE & MMSE), feature files (autoencoder), and lattice files (MMI & MPE).
- Only standard EBP with SGD is available at present.
  - Gradient refinement: momentum, graident clipping, L2 norm, *etc.*
  - Learning rate schedulers: List, Exponential Decay, Ada Grad, modified New Bob, *etc.*

Data Cache

- Frame based shuffling: CE/MMSE for DNN and (unfolded) RNN.
- Utterance based shuffling: MMI, MPE, and MWE training.
- Batch of utterance level shuffling: RNN, ASGD.

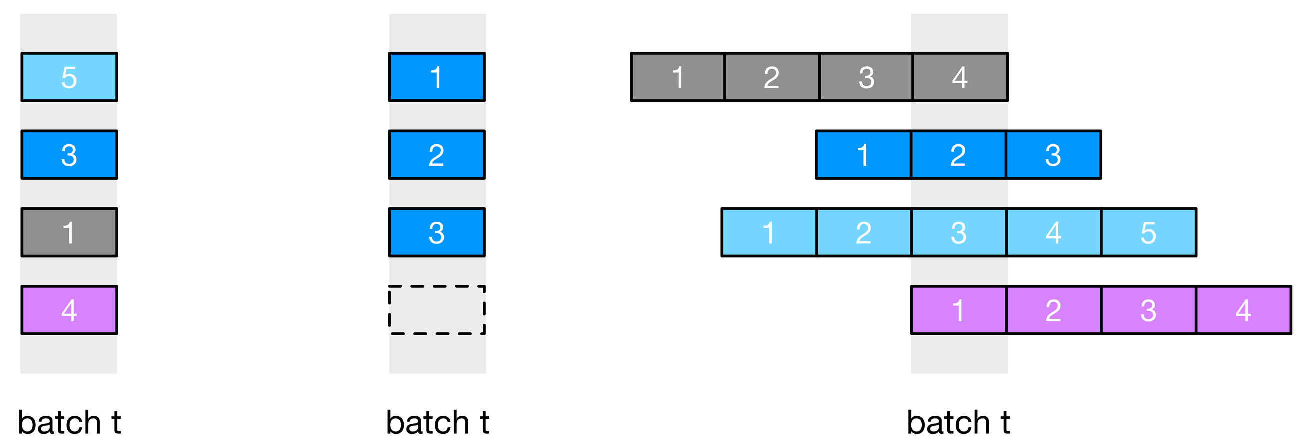

Figure 1: Examples of different types of data shuffling

Generic ANN Support

- Each ANN can have any number of layers.
  - The input vector to an ANN layer is defined by a *feature mixture*.
  - Each feature mixture has any number of *feature elements*.
  - A feature element defines a fragment of the input vector by *source* (acoustic features or ANN layers) and *context shift set* (integers for time difference).
- ANNs can be any directed cyclic graph (recurrent ANNs) but only directed acyclic graphs (feedforward ANNs) can be trained.

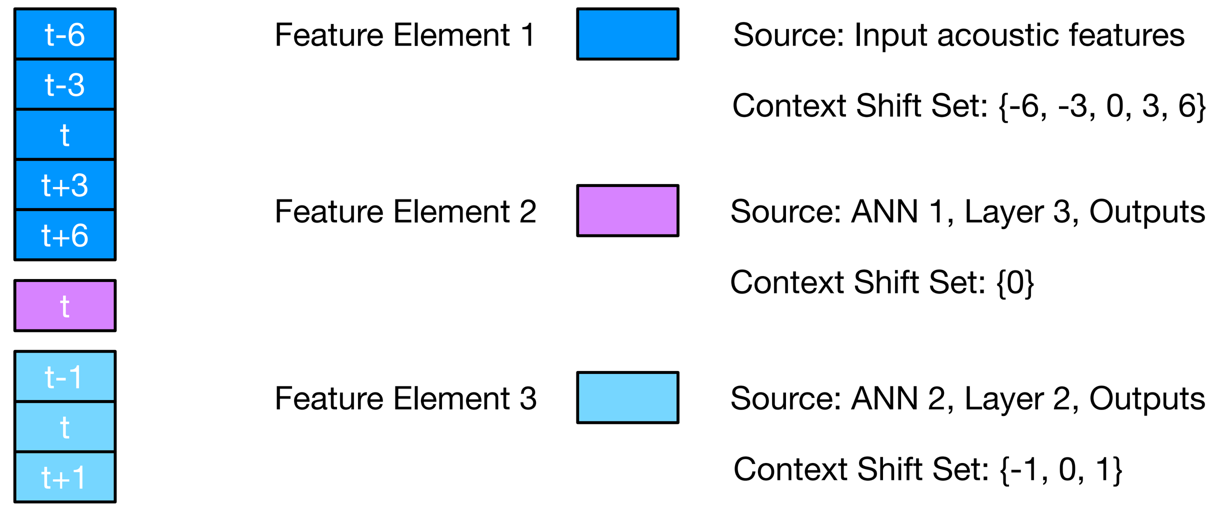

Figure 2: An example of a feature mixture

Other Features

- Math kernels: new CPU, MKL, and CUDA kernels for ANNs.
- Input transforms: compatible with HTK SI/SD input transforms.
- Speaker adaptation: ANN parameters replacement.
- Model Edit (using HHEd)
  - Insert/Remove/Initialise ANN layers
  - Add/Delete a feature element to/from a feature mixture
  - Associate an ANN model with HMMs
- Decoders
  - HVite: tandem/hybrid decoding/alignment/model marking
  - HDecode: tandem/hybrid system LVCSR decoding
  - HDecode.mod: tandem/hybrid system model marking
  - Joint decoder: log-linear combination of HTK systems.

Building Hybrid SI System

- Steps to build CE based SI CD-DNN-HMMs using HTK
  - Produce tied state GMM-HMMs by decision tree tying (HHEd).
  - Generate ANN-HMMs by replacing GMMs with an ANN (HHEd).
  - Generate alignments with a pre-trained system (HVite).
  - Train ANN-HMMs based on CE (HNTrainSGD).
- Steps for CD-DNN-HMM MPE training
  - Generate num. & den. lattices (HLRescore & HDecode).
  - Phone mark num. & den. lattices (HVite or HDecode.mod).
  - Perform MPE sequence training (HNTrainSGD).

ANN Front-ends for GMM-HMMs

- ANNs can be used as GMM-HMM front-ends by using a feature mixture to define the composition of the GMM-HMM input vector.
- HTK can accommodate a tandem SAT system as a single system.
  - Mean & variance norm are treated as affine activation functions.
  - SD parameters are replaceable according to speaker ids.

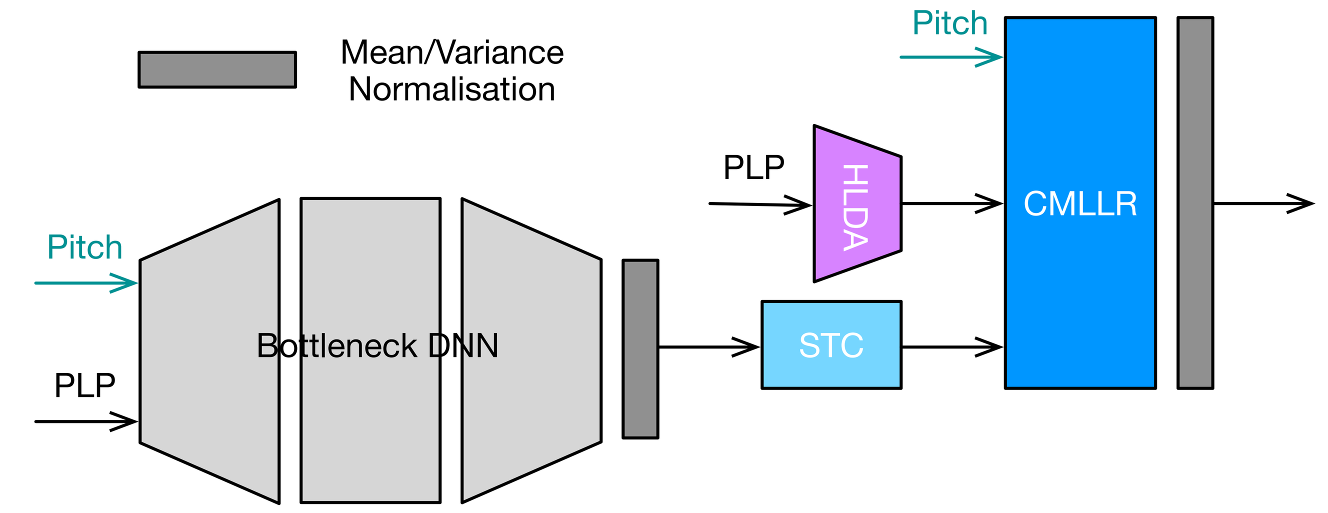

Figure 3: A composite ANN as a tandem SAT system front-end

Experiments

- Systems were trained on 300 hours Mandarin CTS data and evaluated on 2014 DARPA BOLT project development set.
- DNNs with 2k node hidden layers and 12k node output layer.
- Joint decoding was with system dependent weights (1.0, 0.2).

System	Criterion	%WER
Hybrid SI	CE	34.5
Hybrid SI	MPE	31.6
Tandem SAT	MPE	33.2
Hybrid SI $\oplus$ Tandem SAT	MPE	31.0

Table 1: *BOLT tandem, hybrid, and joint decoding performance.*

- Systems also built using 700 hours English broadcast data selected from 7 weeks of BBC programmes.
- Evaluations on BBC 1 week development set.
- DNNs have 1k hidden nodes and 9.5k output nodes.
- System dependent weights for hybrid and tandem joint decoding were (1.0, 0.4).

System	Criterion	%WER
Hybrid SI	CE	28.4
Hybrid SI	MPE	25.9
Tandem SI	MPE	27.0
Hybrid SI $\oplus$ Tandem SI	MPE	24.6

Table 2: *Multi-Genre Broadcast (MGB) Challenge developing system performance. Results were with manual segmentation, 64k vocabulary, and fg language model.*

E-Mail: cz277@cam.ac.uk

Website: http://mi.eng.cam.ac.uk/~cz277