A General ANN Extension for HTK  
Chao Zhang & Phil Woodland  
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 3.5 in 2015.  

Design Principles  
- To accommodate new models and methods, HTK-ANN should be designed should be 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 code and tools.  
  - To simplify the transfer of many technologies.  
- HTK-ANN should be "research friendly".  

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 now.  

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 (for CE & MMSE), feature files (for autoencoder), and lattice files (for MMI & MPE).  
- Only standard EBP with SGD is available at present.  
  - Gradient refinement: momentum, gradient clipping, weight decay, etc.  
  - Learning rate schedulers: List, Exponential Decay, AdaGrad, 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.  

Other Features  
- Math kernels: CPU, MKL, and CUDA based new kernels for ANNs.  
- Input transforms: compatible with HTK SI/SD input transforms.  
- Speaker adaptation: an ANN parameter unit online replacement.  
- Model Edit (using HHEd)  
  - Insert/Remove/Initialise an ANN layer  
  - Add/Delete a feature element to a feature mixture  
  - Associate an ANN model to HMMs  
- Decoders  
  - HVite: tandem/hybrid system 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 (based on the same decision tree).  

Building Hybrid SI System  
- Steps of building CE based SI CD-DNN-HMMs using HTK  
  - Produce desired tied state GMM-HMMs by decision tree tying (HHEd).  
  - Generate ANN-HMMs by replacing GMMs with an ANN (HHEd).  
  - Generate frame-to-state labels with a pre-trained system (HVite).  
  - Train ANN-HMMs based on CE (HTrainSGD).  
- 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 (HTrainSGD).  

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 normalisations are treated as activation functions.  
  - SD parameters are replaceable according to speaker ids.  

Experiments  
- Systems were trained on 200 hours NST MGB Challenge Data and evaluated on BBC 1week development set (manual segmentation).  
- DNNs are with 1k node hidden layers and 6k node output layers.