Standalone Training of Context-Dependent Deep Neural Network Acoustic Models

Chao Zhang & Phil Woodland

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

University of Cambridge

11 November 2013

Conventional Training of CD-DNN-HMMs

- CD-DNN-HMMs rely on GMM-HMMs in two aspects:
  - Training labels — state-to-frame alignments
  - Tied CD state targets — GMM-HMM based decision tree state tying
- Is it possible to build CD-DNN-HMMs independently from any GMM-HMMs?
- Standalone training of CD-DNN-HMMs

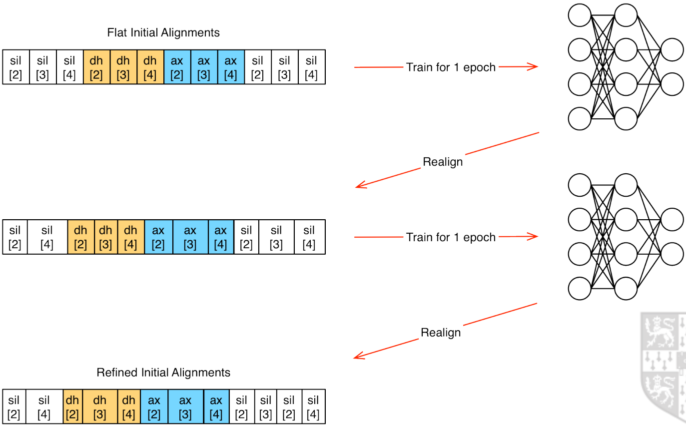
Standalone Training of CD-DNN-HMMs

- The standalone training strategy can be divided into two parts:
  - Alignments — by CI- (monophone state) DNN-HMMs trained in a standalone fashion
  - Targets — by DNN-HMM based decision tree target clustering

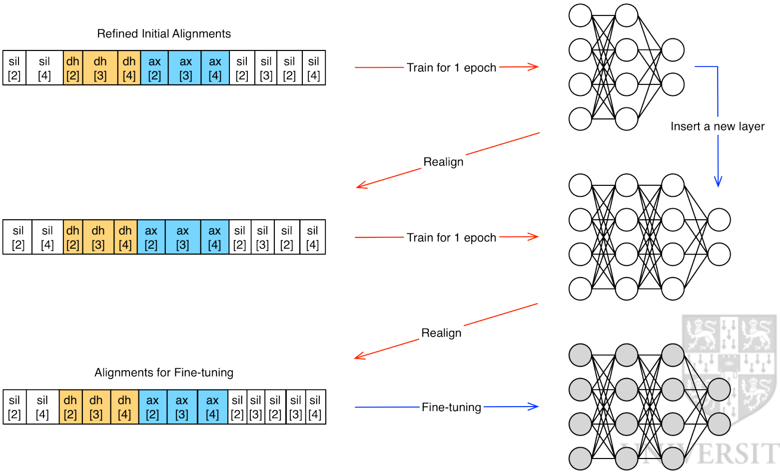
Standalone Training of CI-DNN-HMMs

- The standalone CI-DNN-HMMs are trained with *flat initial alignments* (with averaged CI state duration)
- CI-DNN-HMMs training include:
  - Refine initial alignments in an iterative fashion
  - Train a CI-DNN-HMMs using *discriminative pre-training with realignment* and standard fine-tuning

Initial Alignment Refinement



Discriminative Pre-training with Realignment



DNN-HMM based Target Clustering

Natural Speech Technology

- Assume the output distribution for each target is Gaussian with common covariance matrix, i.e., $p(\mathbf{z} \mid \mathcal{C}_k) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$
  - the kth target
  - sigmoidal activation vector from the last hidden layer
- $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ are estimated based on maximum likelihood criterion
  - the features are de-correlated with state-specific rotation
  - the left clustering process is the same as the original approach
- Next, we investigate the link between the Gaussian distributions and the DNN output layer

UNIVERSITY OF CAMBRIDGE

DNN-HMM based Target Clustering

Natural Speech Technology

- From Bayes' theorem,

$$p(\mathcal{C}_k|\mathbf{z}) = \frac{p(\mathbf{z}|\mathcal{C}_k)P(\mathcal{C}_k)}{\sum_{k'} p(\mathbf{z}|\mathcal{C}_{k'})P(\mathcal{C}_{k'})}$$

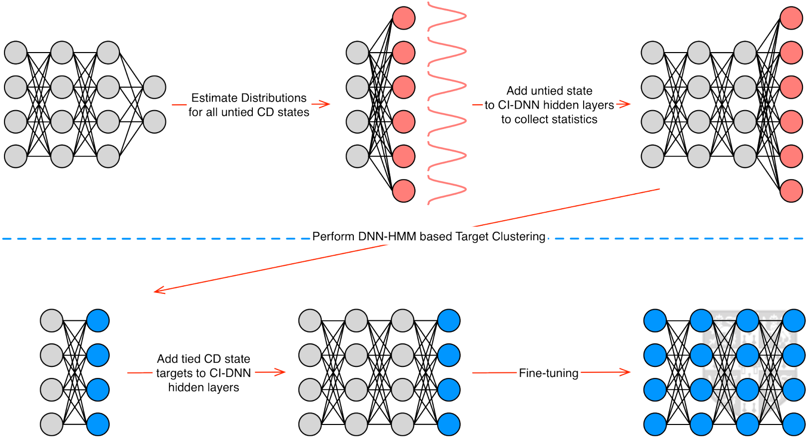
$$= \frac{\exp\{\boldsymbol{\mu}_k^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z} - \frac{1}{2}\boldsymbol{\mu}_k^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln P(\mathcal{C}_k)\}}{\sum_{k'} \exp\{\boldsymbol{\mu}_{k'}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z} - \frac{1}{2}\boldsymbol{\mu}_{k'}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{k'} + \ln P(\mathcal{C}_{k'})\}}$$

- According to *softmax* output activation function,

$$p(\mathcal{C}_k|\mathbf{z}) = \frac{\exp\{\mathbf{w}_k^{\mathsf{T}}\mathbf{z} + b_k\}}{\sum_{k'} \exp\{\mathbf{w}_{k'}^{\mathsf{T}}\mathbf{z} + b_{k'}\}}$$

UNIVERSITY OF CAMBRIDGE

Procedure of Building CD-DNN-HMMs



Experiments

- Wall Street Journal training set (SI-284), along with 1994 H1-dev (Dev) and Nov'94 H1-eval (Eval) testing sets were used.
  - utterance level CMN and global CVN
- MPE GMM-HMMs have 5981 tied triphone states and 12 Gaussian components per state
  - MPE GMM-HMMs were with $((\text{13PLP})_{\text{D\_A\_T\_Z}})_{\text{HLDA}}$
- Every DNN had 5 hidden layers with 1000 nodes per layer
  - All DNN-HMMs were with $9 \times (\text{13PLP})_{\text{D\_A\_Z}}$
  - sigmoid/softmax hidden/output activation function
  - cross-entropy training criterion
- 65k dictionary and trigram language model

CI-DNN-HMM Results

Table : Baseline CI-DNN-HMM Results ($351 \times 1000^5 \times 138$).

ID	Type	DNN Alignments	WER% Dev	WER% Eval
G2	MPE GMM-HMMs	—	8.0	8.7
I1	CI-DNN-HMMs	G2	10.5	12.0

Table : Different CI-DNN-HMMs trained in a standalone fashion.

ID	Training Route	WER% Dev	WER% Eval
I3	Realigned	12.2	14.3
I4	Realigned+Conventional	11.7	13.8
I5	Conventional	12.2	15.0
I6	Conventional+Conventional	12.0	14.6

CD-DNN-HMM Results

- Baseline CD-DNN-HMMs (D1) were trained with G2 alignments. The WER on Dev and Eval are 6.7 and 8.0, respectively.
- CD-DNN-HMMs with different clustered targets were listed in the table. The hidden layer and alignments were from I4.

Table : CD-DNN-HMM based state tying results ($351 \times 1000^5 \times 6000$).

ID	Clustering	BP Layers	WER%	
			Dev	Eval
G3	GMM-HMM	Final Layer	7.6	9.0
G4		All Layers	6.8	7.9
D2	DNN-HMM	Final Layer	7.7	8.7
D3		All Layers	6.8	7.8

- The CD-DNN-HMMs (D3) trained without relying on any GMM-HMMs is comparable to baseline D1.

Conclusions

- We accomplish training CD-DNN-HMMs without relying on any pre-existing system
  - train CI-DNN-HMMs by updating the model parameters and the reference labels in an interleaved fashion
  - adapt decision tree tying to the sigmoidal activation vector space of a CI-DNN
- The experiments on WSJ SI-284 have shown
  - the proposed training procedure gives state-of-the-art performance
  - the methods are very efficient