# Standalone Training of Context-Dependent Deep Neural Network Acoustic Models

Natural Speech Technology

Chao Zhang and Phil Woodland

UNIVERSITY OF CAMBRIDGE

Cambridge University Engineering Department

## Introduction

**Standard CD-DNN training:**

- A well-trained GMM-HMM system has to be used for CD-DNN training for
  - state-to-frame alignments;
  - defining a set of tied context-dependent (CD) states.
- State-to-frame alignments serve as the training labels of CD-DNNs.
- DNN targets are derived from decision-tree based tied-state GMM-HMM system.

**Standalone CD-DNN training:**

- We propose to train CD-DNNs independently from any existing system by
  - training CI-DNNs using discriminative pre-training with integrated realignments;
  - modifying standard decision tree state tying to cluster explicitly estimated (approximately equivalent terms to) CD-DNN output distributions.
- Proposed technique gives comparable WERs to GMM-HMM dependent CD-DNNs.

## Proposed Training Procedure for CI-DNNs

**Initial Alignment Refinement:**

- The initial alignments are transcriptions with uniformly segmented CI states.
- The alignments are repeatedly refined for a number of iterations by

1. training a 3-layer MLP from scratch for 1 epoch with current alignments;
2. using the resulting MLP to realign the training set.

**Discriminative Pre-training with Realignment:**

- Aim is to interleave training label refinement with adding hidden layers to the DNN.
- The pre-training steps are

1. train a 3-layer MLP for 1 epoch and use it to realign the data;
2. replace current output layer with a hidden layer along with a new output layer;
3. train the modified MLP with the latest alignments for 1 epoch;
4. use the MLP to realign the reference transcriptions;
5. repeat steps 2-5 until required DNN structure is realised.

## DNN Class-Conditional Distributions for Decision Tree Tying

- To use decision tree tying, DNN class-conditional distributions are needed.
- If $\mathbf{z}_t$ is the input to the final layer. For a DNN output target class $\mathcal{C}_k$, assume $p(\mathbf{z}_t|\mathcal{C}_k) = \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_k, \Sigma)$, we have

$$p(\mathcal{C}_k|\mathbf{z}_t) = \frac{\exp\left\{-0.5 \cdot \mathbf{z}_t\Sigma^{-1}\mathbf{z}_t + \boldsymbol{\mu}_k^{\mathrm{T}}\Sigma^{-1}\mathbf{z}_t - 0.5 \cdot \boldsymbol{\mu}_k^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_k + \ln P(\mathcal{C}_k)\right\}}{\sum_{k'}\exp\left\{-0.5 \cdot \mathbf{z}_t\Sigma^{-1}\mathbf{z}_t + \boldsymbol{\mu}_{k'}^{\mathrm{T}}\Sigma^{-1}\mathbf{z}_t - 0.5 \cdot \boldsymbol{\mu}_{k'}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_{k'} + \ln P(\mathcal{C}_{k'})\right\}}.$$

- The relation between $\mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_k, \Sigma)$ and the final layer is

$$\eta\,\mathbf{w}_k^{\mathrm{T}} = \boldsymbol{\mu}_k^{\mathrm{T}}\Sigma^{-1}$$
$$\eta\, b_k = -0.5\boldsymbol{\mu}_k^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_k + \ln P(\mathcal{C}_k),$$

where $\mathbf{w}_k$ and $b_k$ are the weights and bias of $\mathcal{C}_k$, $\eta$ is a scaling factor.

## DNN-HMM based Decision Tree Target Clustering

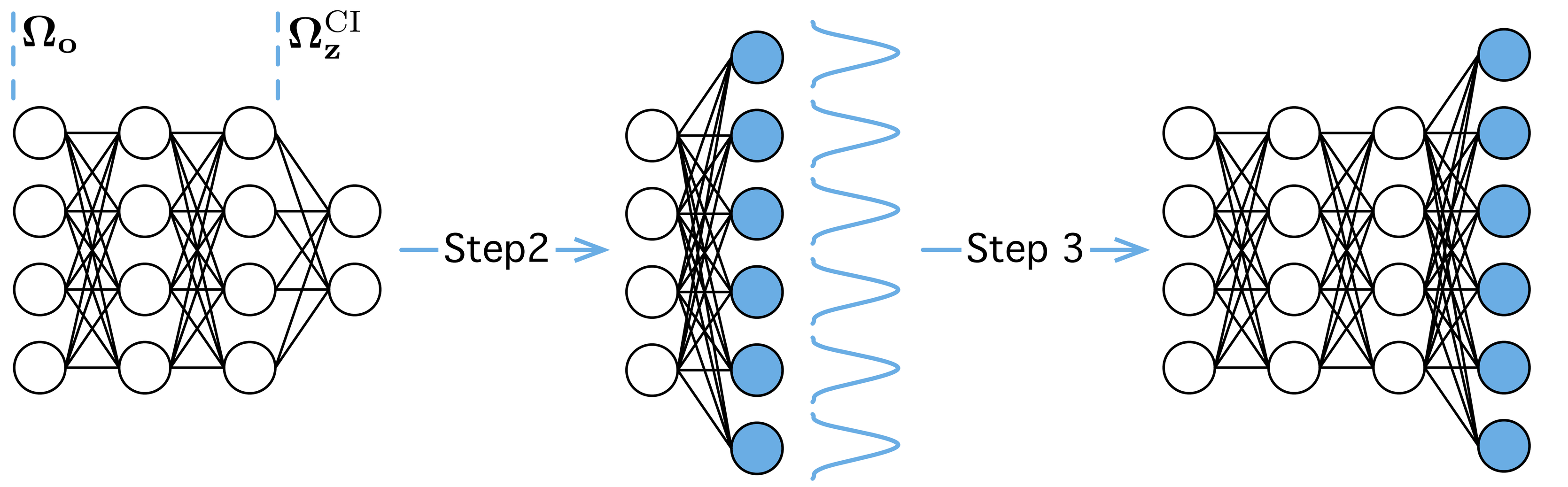
**Adapting GMM-HMM decision tree tying to DNN-HMMs:**

- Estimate $\mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_k, \Sigma)$ → convert to a DNN output layer → collect $\sum_t \gamma_k(t)$ using a modified DNN → do decision tree tying with $\mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_k, \Sigma)$ and $\sum_t \gamma_k(t)$.
- $\mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_k, \Sigma)$ are estimated based on maximum likelihood criterion.
- PCA is used to simplify the computation of $|\Sigma|$ for decision tree tying.
- GMM/DNN based decision tree clusters in $\Omega_\mathbf{o}/\Omega_\mathbf{z}$, the space of $\mathbf{o}/\mathbf{z}$.

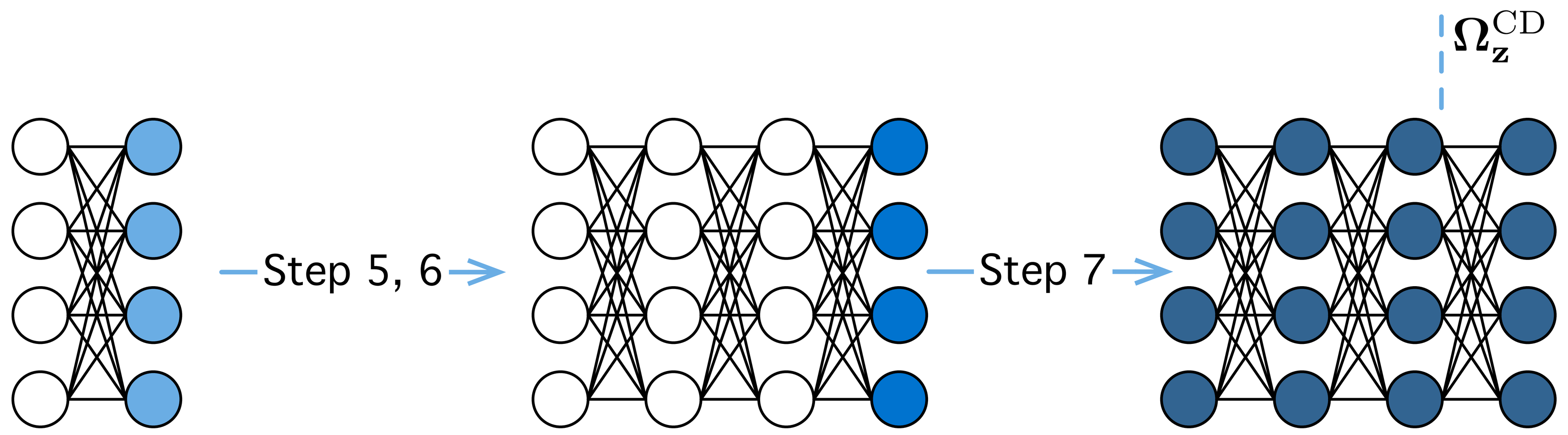
**Training CD-DNN-HMMs based on CI-DNN-HMMs:**

- The steps of training CD-DNN-HMMs using existing CI-DNN-HMMs are

1. realign the training set with CI-DNN-HMMs;
2. estimate $p(\mathbf{z}_t|\mathcal{C}_k)$ for all seen untied CD states with CI-DNN-HMM hidden layers;
3. convert $p(\mathbf{z}_t|\mathcal{C}_k)$ to an output layer with untied CD states to collect $\sum_t \gamma_k(t)$;




4. perform DNN decision tree tying to generate the clustered CD state targets;
5. add a new output layer with the clustered targets to CI-DNN-HMM hidden layers;
6. train the output layer only and realign the training set with the resulting model;
7. perform fine-tuning according to the new alignments.




- This method make predictions about the best targets of $\Omega_\mathbf{z}^{\mathrm{CD}}$ in $\Omega_\mathbf{z}^{\mathrm{CI}}$.

## Experimental Setup

- Wall Street Journal training set (SI-284), along with 1994 H1-dev (Dev) and Nov'94 H1-eval (Eval) testing sets were used.
- MPE GMM-HMMs with $((\text{13PLP})_{\text{D\_A\_T\_Z}})_{\text{HLDA}}$ had 5981 tied triphone states.
- Every DNN was with $9 \times (\text{13PLP})_{\text{D\_A\_Z}}$ and had $5 \times 1000$ hidden layers.
- All experiments were with a 65k dictionary and a trigram language model.

## Experimental Results

**Baseline System Performance:**

- I1 performed better than I2 since system G2 had HLDA and $\Delta\Delta\Delta$ features.

| ID | Type | Alignments | Dev WER% | Eval WER% |
|---|---|---|---|---|
| G2 | MPE GMM-HMMs | — | 8.0 | 8.7 |
| I1 | CI-DNN-HMMs | G2 | 10.5 | 12.0 |
| I2 | CI-DNN-HMMs | I1 | 10.7 | 13.7 |
| D1 | CD-DNN-HMMs | G2 | 6.7 | 8.0 |

**CI-DNN-HMM Standalone Training:**

- Discriminative pre-training with realignment (I3 and I4) reduced WERs.

| ID | Training Route | Dev WER% | Eval WER% |
|---|---|---|---|
| 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 |

**DNN-HMM based Target Clustering:**

- D2 outperformed G3 → clustering in $\Omega_\mathbf{z}^{\mathrm{CI}}$ (of I4) matches I4 hidden layers better.
- Fine-tuning reduced the WER difference between different clustering methods.
- The standalone CD-DNN-HMM system, D3, is comparable to D1, in terms of WER.

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

## Conclusion

- We accomplished training CD-DNN-HMMs without relying on any existing system.
  - Training CI-DNNs interleaves reference state alignment and adding new hidden layers.
  - Modified decision tree state tying to cluster Gaussian distributions with a common covariance matrix for every untied CD state based on $\mathbf{z}$ of the CI-DNN.
- The proposed training procedure gives state-of-the-art hybrid system performance on the standard SI-284 training setup for the Wall Street Journal corpus.

## Acknowledgements

Chao Zhang is supported by Cambridge International Scholarship from the Cambridge Commonwealth, European & International Trust and by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).


E-Mail: cz277@cam.ac.uk

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