Unsupervised Task Discovery in Multi-Task Acoustic Modeling

Josh Meyer[*]
* University of Arizona

THE UNIVERSITY OF ARIZONA

Abstract

- *Multi-Task Learning works (esp. in low-resource)*
- *However, tasks are hard to make*
- *Better to discover tasks automatically*
- *Experiment with k-means on MFCCs*
- *Initial results*

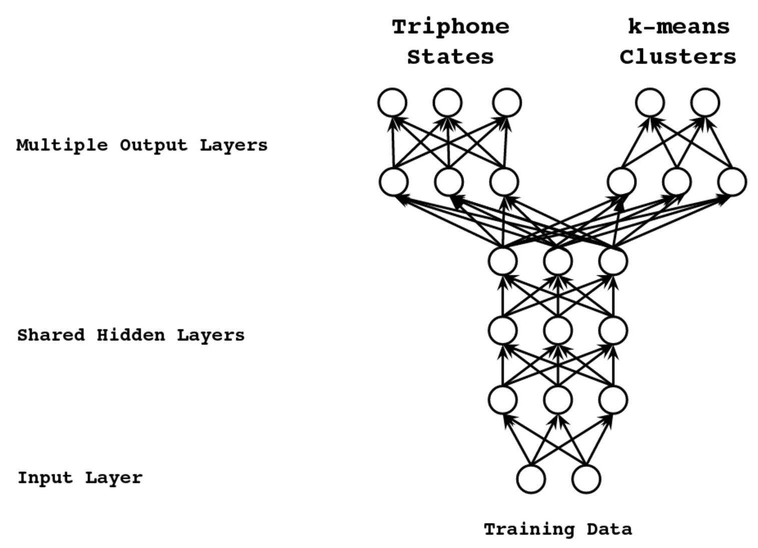

Figure 1: Multi-Task Learning Architecture

1. Background

- Multi-Task Learning in Acoustic Modeling
  - Multilingual
    - new language == new task
  - Monolingual
    - new linguistic encoding == new task
    - Monophones vs. Triphones

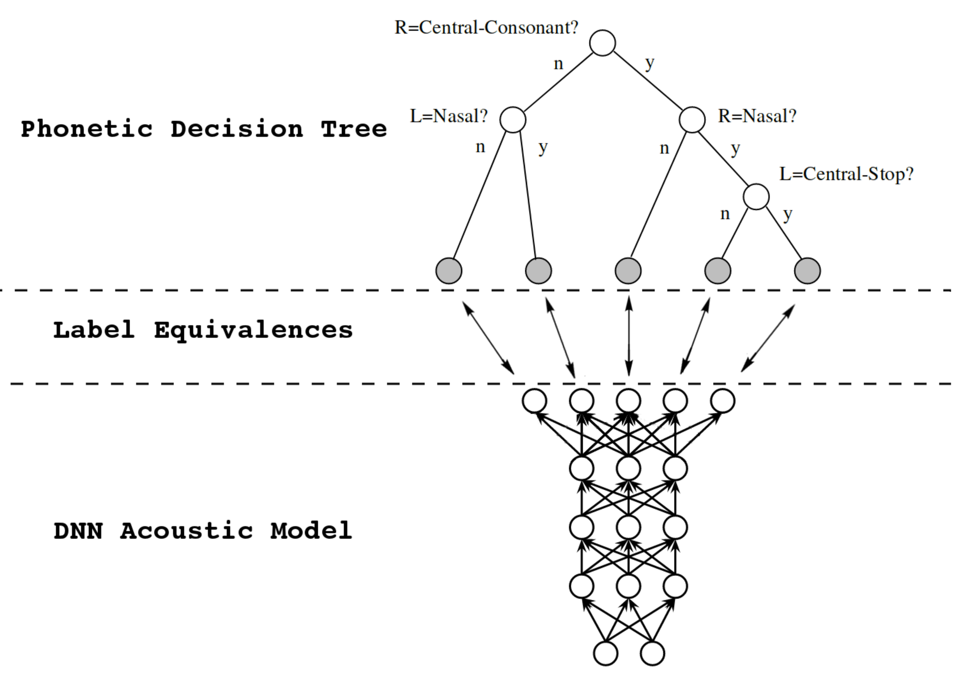

Figure 2: Label Correspondance of Decision Tree / DNN

2. Alignment

- Feature Extraction
  - 13 PLP features, 25ms Hamming windows, 10ms shift, 16 frame left-context & 12 frame right-context, CMVN
- GMM Alignment
  - Monophones: 1,000 Gaussians, 25 iterations EM // Triphones: 2,000 leaves & 5,000 Gaussians, 25 iterations EM

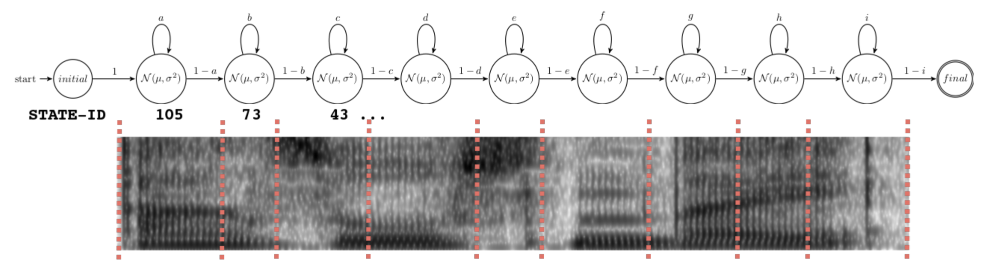

Figure 3: GMM-aligned training examples

3. Clustering

- k-means Clustering
  - A set number of clusters is discovered via TensorFlow's standard k-means clustering.

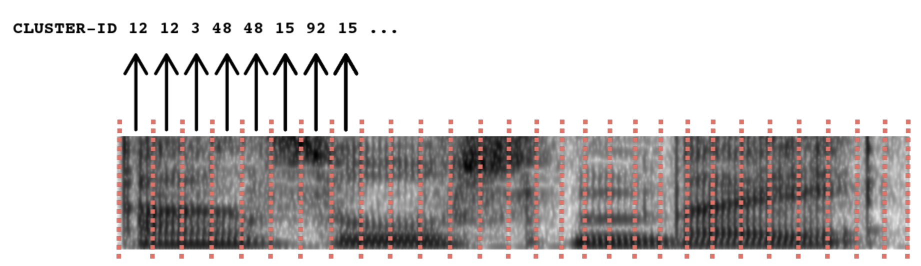

Figure 4: k-means clustered training examples

4. Mapping Triphone States → Clusters

- Mapping triphone states → k-means clusters
  - All training examples aligned to triphone state are mapped to most common k-means cluster.

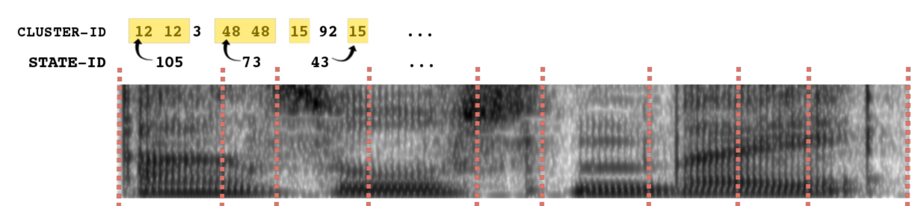

Figure 5: GMM-aligned training examples

5. Cluster Contents

- 672 leaves in Kaldi and 1024 clusters in TF
- 185 new labels after mapping
  - 123 / 185 are interpretable
- 101 of new labels contain mixed phonemes
  - 39 / 101 contained either only vowels or only consonants
- 84 of new labels contain one phoneme
  - 9 / 84 contained more than one triphone of phoneme

Table 1: Discovered intelligible Phoneme Clusters

Vowels		Consonants	
a j	a u	k r	g n m
a o	a ih	k p	s sh ch
e j	e ih	r ng	t k s p
e y	o u	d ch	m ng
u ih y	u ih	t k	t k h
i e y	o ih	d z	t k s
a e oe j ih	j ih	l z	t ch d
a ih o u y		n p	t k zh b
			t g b s sh z zh

6. Multi-Task DNN Training Set-up

- DNN Acoustic model training
  - Multi-Task Time-Delay Neural Network
  - 5-epochs, 11 hidden layers, *ReLU* activations
  - $\alpha_{initial} = 0.0015 \rightarrow \alpha_{final} = 0.00015$
  - Each task has penultimate + ultimate output layer

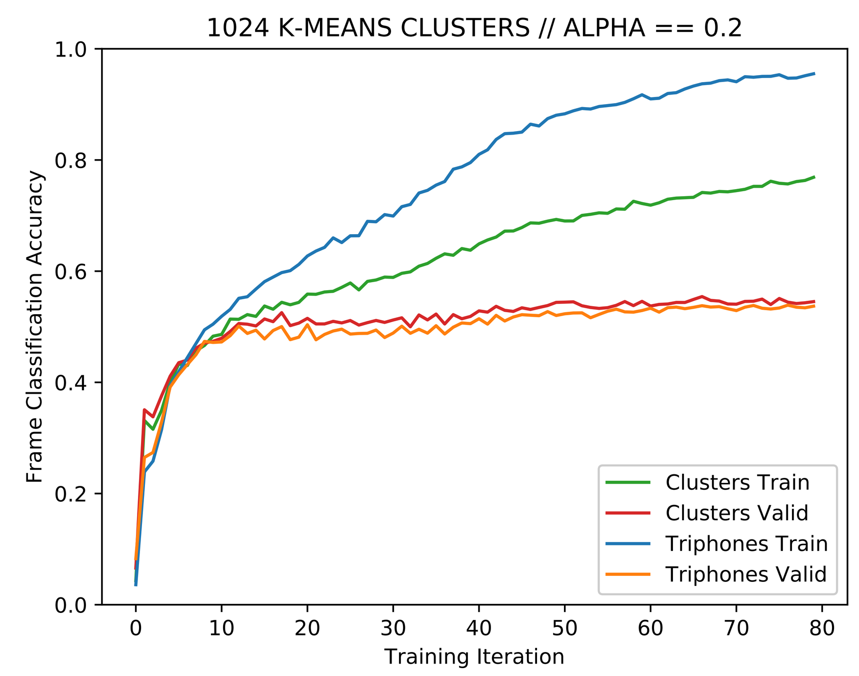

Figure 6: Model Accuracy During Training

7. Testing Setup

- k-folds cross-validation ($k == 6$)
  - 511 utterances for train
  - 100 utterances for test
- Decoded with 1-gram LM

8. Results: Traditional Weighting Scheme

- Loss $= ((1 - \alpha) * MAIN + \alpha * AUX)$
- WER not better than Baseline

Table 2: WER% for Traditional Weighting Scheme

	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
Single Task Baseline		57.10 ± 3.25	
+ 256 k-means cluster targets	57.71 ± 1.59	57.27 ± 1.60	57.89 ± 1.29
+ 1024 k-means cluster targets	57.74 ± 3.17	**57.08** ± 2.62	57.77 ± 0.79
+ 4096 k-means cluster targets	57.13 ± 2.45	57.76 ± 1.61	57.72 ± 0.64

9. Results: Simple Weighting Scheme

- Loss $= (MAIN + \alpha * AUX)$
- WER better than Traditional Loss
- WER marginally better than Baseline (in some cases)

Table 3: WER% for Simple Weighting Scheme

	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
Single Task Baseline		57.10 ± 3.25	
+ 256 k-means cluster targets	57.28 ± 2.09	57.92 ± 1.78	**56.96** $\pm .70$
+ 1024 k-means cluster targets	57.58 ± 2.68	**56.86** ± 1.11	57.19 ± 1.31
+ 4096 k-means cluster targets	57.78 ± 2.36	57.51 ± 2.65	**57.03** ± 1.48

10. Discussion

- Good auxiliary tasks exist (we just need to find them)
- Initial Results show small improvements, given good hyper-parameters
- Clustering in high-dimensional feature space isn't great
  - Find better projections: LDA, source DNN activations (from well-resourced lang)
- Big net overfits to both tasks
  - add more tasks
  - use smaller net

11. Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE-1746060). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.