

Unsupervised Task Discovery for Multi-Task Acoustic Modeling

Josh Meyer

joshua.richard.meyer@gmail.com | @josh_meyer_ | jrmeyer.github.io



Abstract

- ▶ Multi-Task Learning works (good for low-resource languages)
- ▶ However, tasks are hard to make
- ▶ Better to discover tasks automatically
- ▶ Experiment with k-means on MFCCs
- ▶ Data == 1.5 hours of Kyrgyz audio-book
- ▶ Initial Results Promising

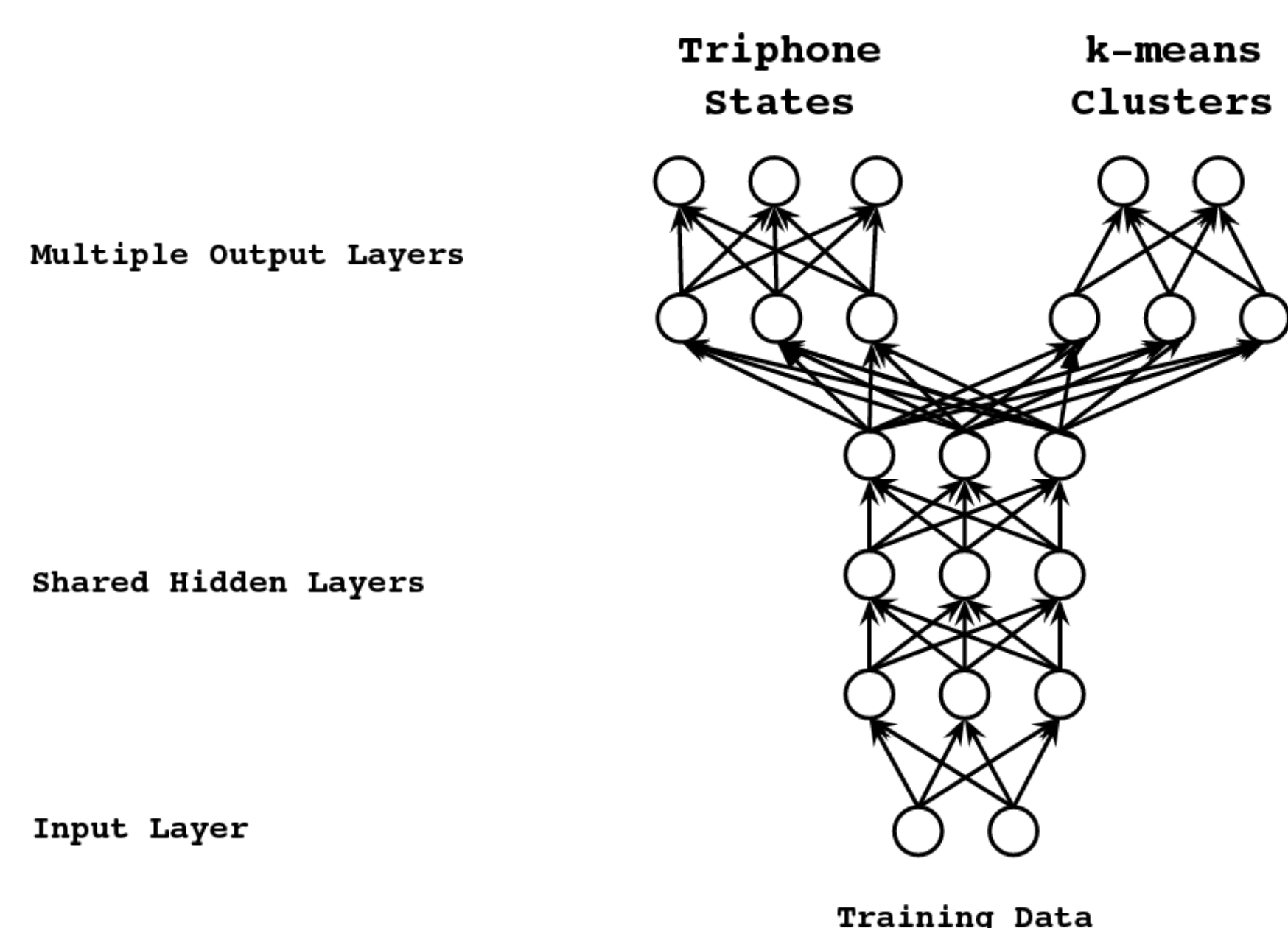


Figure 1: Multi-Task Learning Architecture (Heigold et al. 2013)

1. Background

- ▶ Multi-Task Learning in Acoustic Modeling
 - ▷ Multilingual
 - ▶ new language == new task
 - ▶ e.g. English vs. Kyrgyz
 - ▷ Monolingual
 - ▶ new linguistic encoding == new task
 - ▶ e.g. vowels vs. consonants; monophones vs. triphones

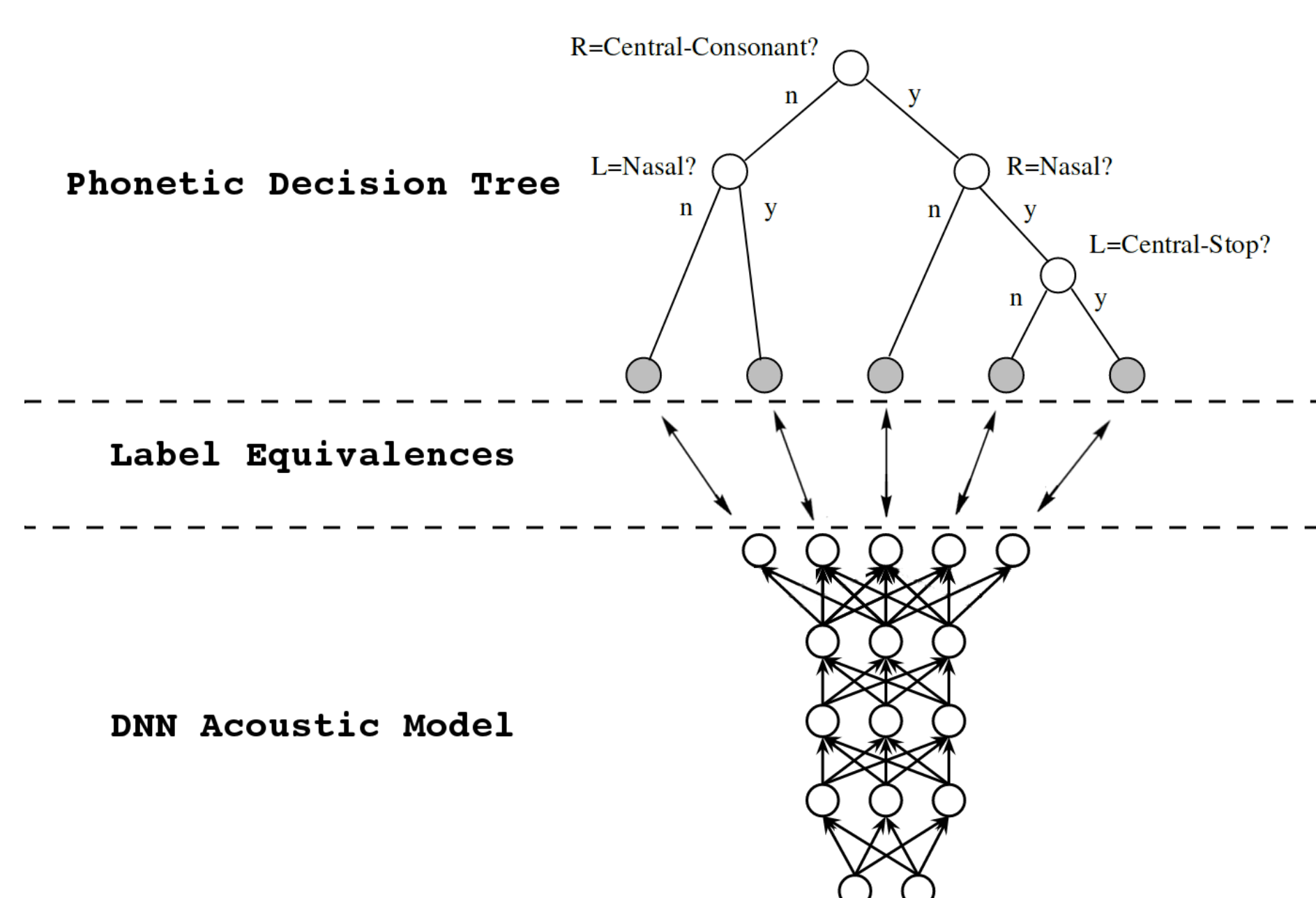


Figure 2: Label Correspondence of Decision Tree / DNN (HTK Book & Heigold et al. 2013)

2. GMM-HMM Alignment

- ▶ All done in Kaldi
- ▶ 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

3. Clustering

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

4. Mapping Triphone States → Clusters

- ▶ All training examples aligned to triphone state are mapped to most common k-means cluster.

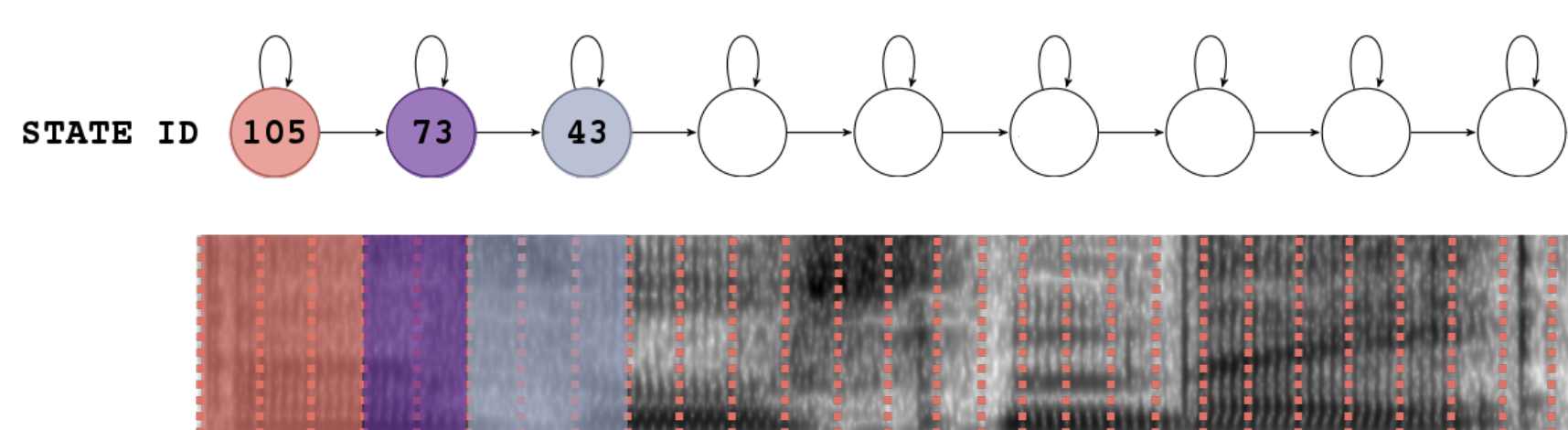


Figure 3: GMM-HMM aligned Triphone States

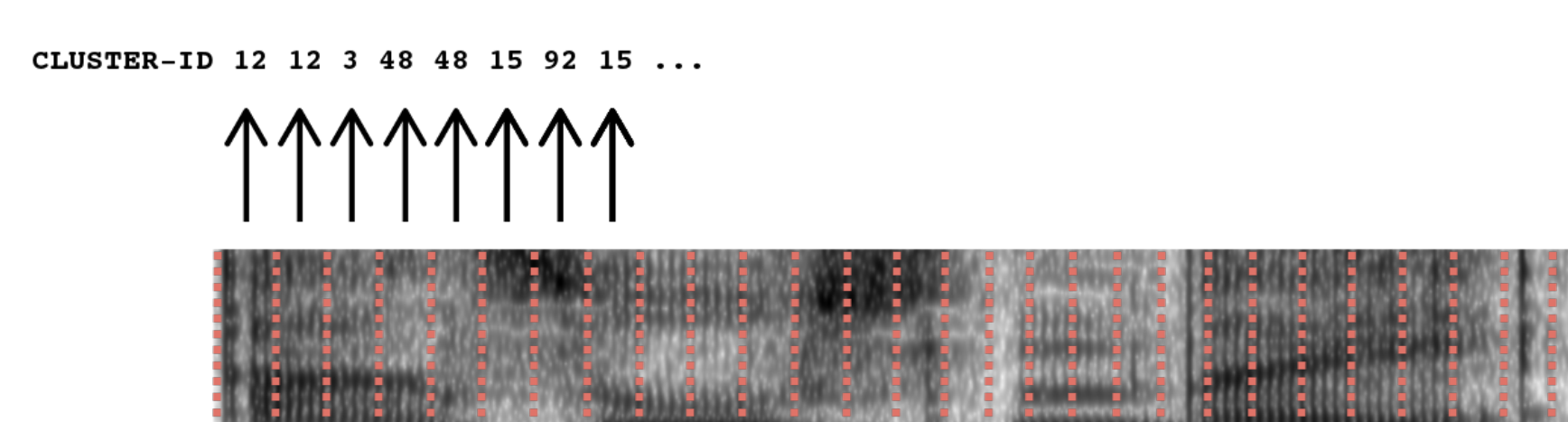


Figure 4: K-Means Discovered Clusters

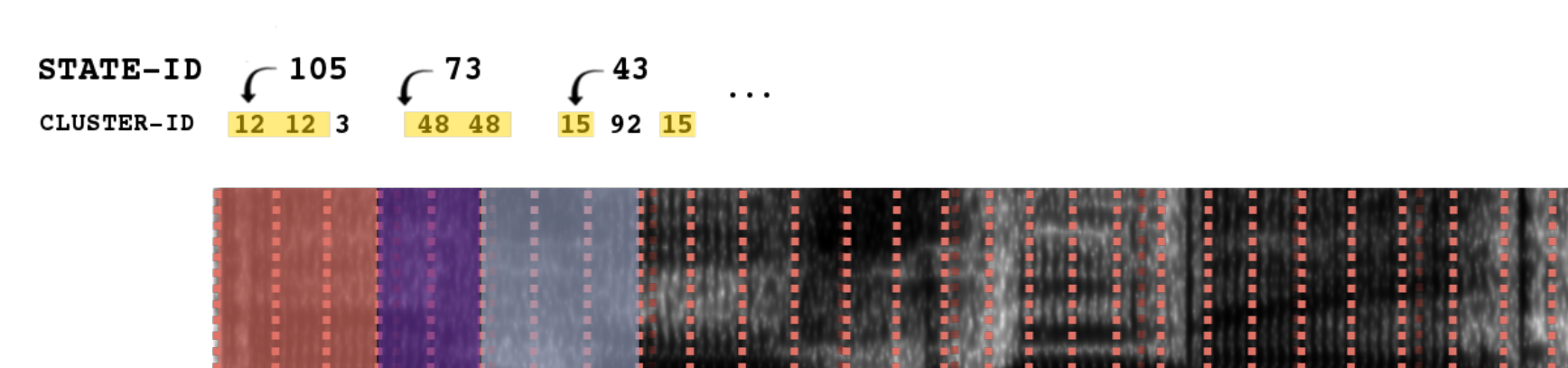


Figure 5: Voting & Mapping of Triphone-States → Clusters

5. Cluster Contents

- ▶ Started with 672 triphone states in Kaldi & 1024 clusters in TensorFlow
- ▶ 185 new labels after mapping
 - ▷ 75 w/ only one triphone state
 - ▶ e.g. { t+a-f }
 - ▷ 9 w/ > 1 triphone of given phoneme
 - ▶ e.g. { t+a-f, g+a-p }
 - ▷ 39 w/ only vowels or only consonants
 - ▶ e.g. { t+a-f, p+o-k }

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	

Table 1: Discovered Clusters of Triphone-States

6. Multi-Task DNN Training Set-up

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

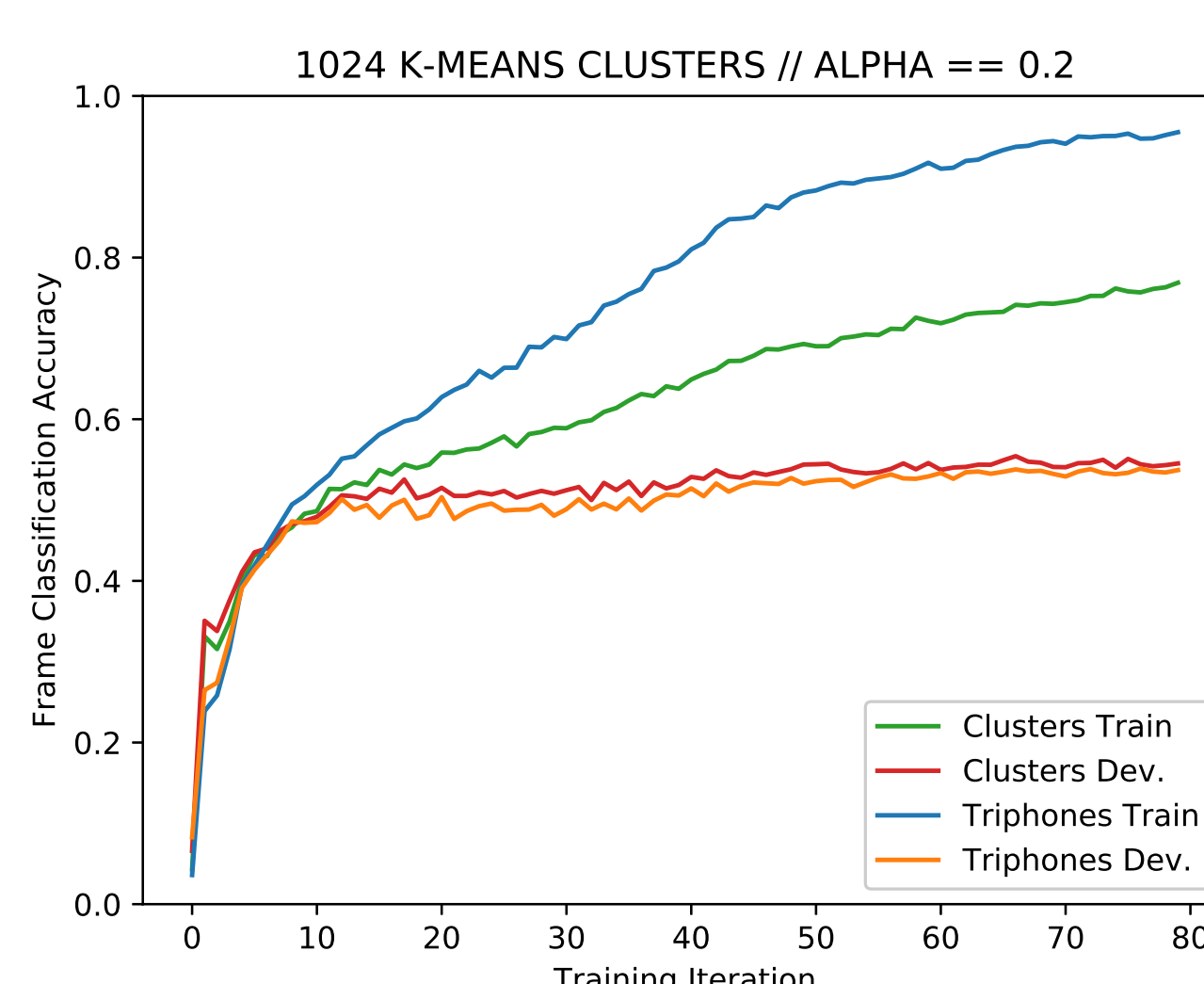


Figure 6: Model Accuracy During Training (Simple Loss)

7. Testing Setup

- ▶ k-folds cross-validation ($k == 5$)
 - ▷ 511 utterances for train
 - ▷ 100 utterances for test
- ▶ Decoded with 1-gram LM
 - ▷ Acoustic model more import

8. Results: Traditional Weighting Scheme

- ▶ Loss = $((1 - \alpha) * MAIN + \alpha * AUX)$
- ▶ WER better than Baseline in 4/9 experiments

Table 2: WER% for Traditional Weighting Scheme

	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
Single Task Baseline		57.55 ± 1.82	
+ 256 clusters	57.93 ± 1.63	57.04 ± 1.58	57.66 ± 1.24
+ 1024 clusters	57.69 ± 3.78	56.99 ± 3.08	57.60 ± 0.79
+ 4096 clusters	57.25 ± 2.87	58.07 ± 1.35	57.45 ± 0.32

9. Results: Simple Weighting Scheme

- ▶ Motivation: downweighting MAIN makes learning hard
- ▶ Loss = $(MAIN + \alpha * AUX)$
- ▶ WER better than Traditional Loss
- ▶ WER better than Baseline in 6/9 experiments

Table 3: WER% for Simple Weighting Scheme

	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
Single Task Baseline		57.55 ± 1.82	
+ 256 clusters	57.33 ± 2.49	58.02 ± 2.09	57.18 ± 0.56
+ 1024 clusters	57.74 ± 3.06	56.88 ± 1.33	57.13 ± 1.55
+ 4096 clusters	57.56 ± 2.53	57.49 ± 3.17	57.31 ± 1.31

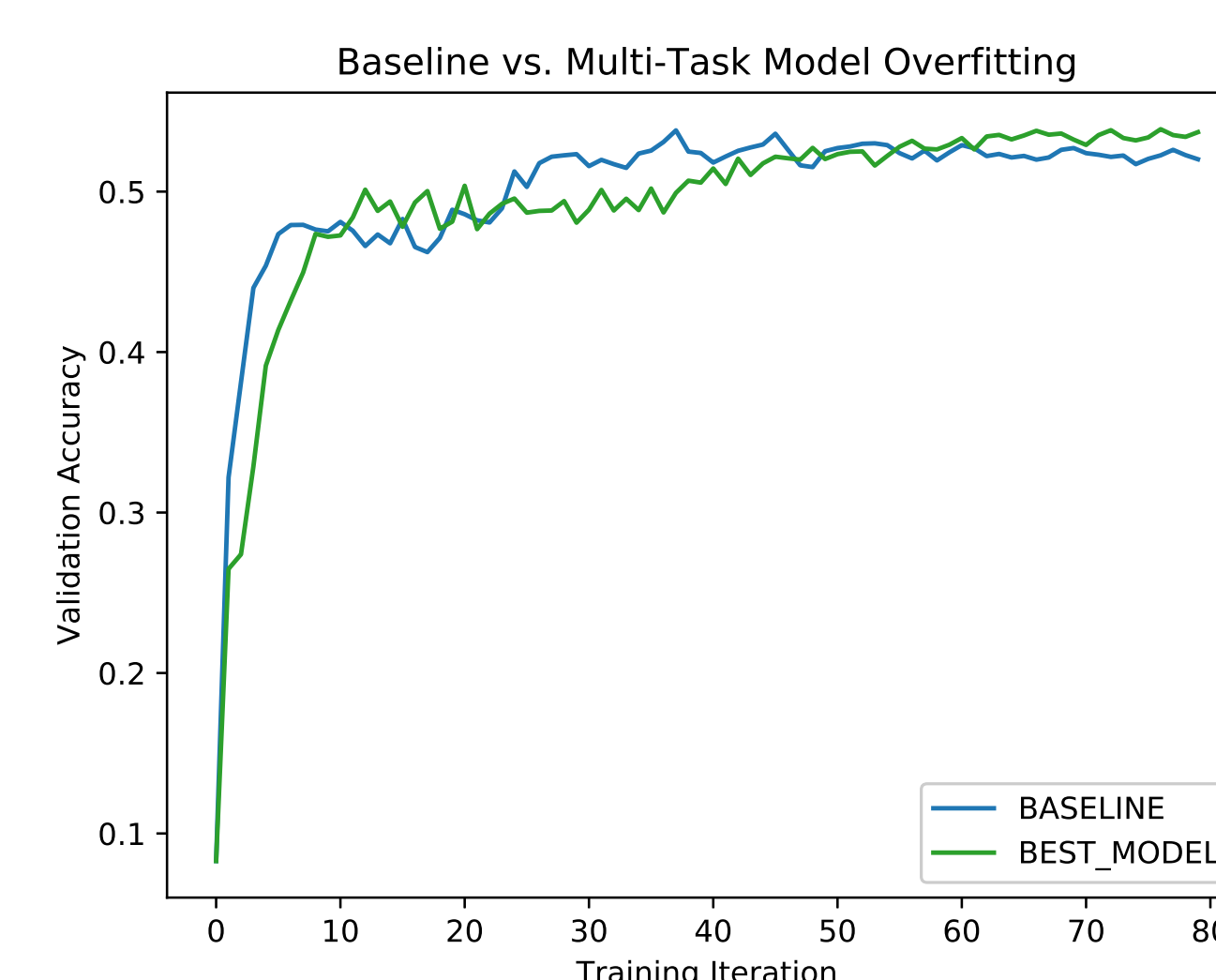


Figure 7: Model Accuracy on Dev. Data During Training

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 author(s) and do not necessarily reflect the views of the National Science Foundation.