Adaptation of DNN Acoustic Models using KL-divergence Regularization and Multi-Task Training

László Tóth, Gábor Gosztolya

Research Group on Artificial Intelligence

Hungarian Academy of Sciences

and the University of Szeged

Szeged, Hungary

MAGYARORSZÁG MEGÚJUL

A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Deep Neural Nets in Speech Recognition

- Hidden Markov Modeling has been the dominant speech recognition technology for about 30 years
- But DNN-based models now clearly outperform standard HMMs
  - Turning HMM/GMMs into HMM/DNNs is quite straightforward
    - The GMMs (estimating $p(X|s)$) are replaced by a DNN (estimating $P(s|X)$)
    - DNN-based posteriors → Bayes' rule → scaled likelihoods
    - This is called the "hybrid" HMM/DNN modelling method
- However, a lot of HMM/GMM refinements cannot be trivially transferred to HMM/DNNs
  - E.g.: context-dependent modeling or speaker adaptation

Context-dependent phone models

- Instead of modeling phones independently of their context („a”), we create models for each possible context („b-a+b”, „b-a+c”,…)
  - Standard for HMM/GMMs, and now for HMM/DNNs as well
    - (we use the same old, Gaussian-based technology for HMM/DNNs…)
- There are a lot of CD models→ few training examples per model
  - Solution: state tying – shared models for similar phones
  - Hierarchical state tying: the number of parameters can be tuned between the two extreme points (fully CI or fully CD models)
  - **We adjust the number of states to the amount of training data**
    - 3 hours of data → 1000 states
    - 300 hours of data → 5000-10000 states

Speaker adaptation

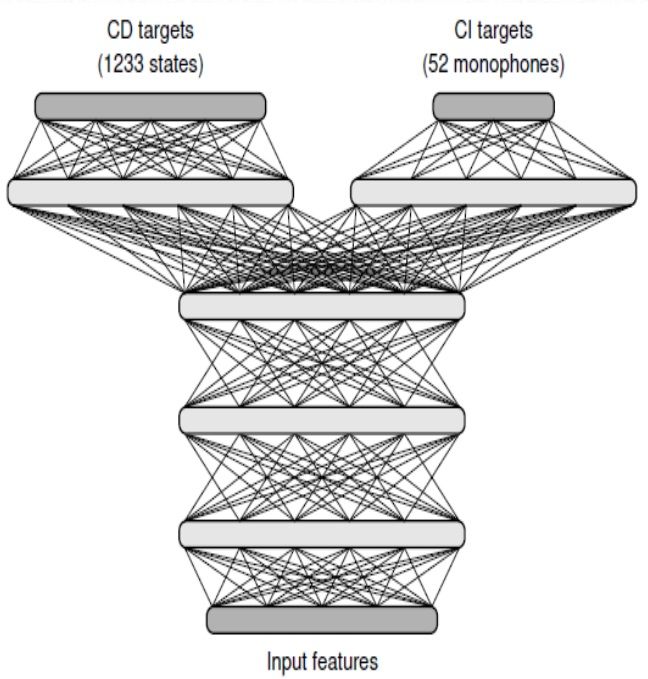
- Goal: to adapt the model to the voice of the actual speaker
  - Supervised: we know (have transcript) for what the speaker said
  - Unsupervised: the transcript is only estimated (by the recognizer)
- HMM/GMM: well established, GMM-specific methods
- HMM/DNN: active research topic, no widely accepted solution
  - Common: all methods train the net further on the adaptation data
- Goal: to use as small adaptation data as possible
- Problem: the adaptation data set is orders of magnitudes smaller than the train set
  - Danger of overfitting the adaptation set!

Adaptation with CD models

- CD models: number of states is adjusted to the train set size
  - Overfitting is almost sure
  - A lot of states will have zero examples in the adaptation set
- Some possible solutions
  - Restrict the number of parameters to be trained (e.g. one layer)
  - Allow only linear transformations (by adding a linear layer)
  - Estimate targets for the classes not seen in the adaptation set
  - Extend the target function with a regularization term
    - Yu et al.: penalizes when the output of the adapted model strays too far from the output of the unadapted model
  - **Here we propose to use multi-task training**

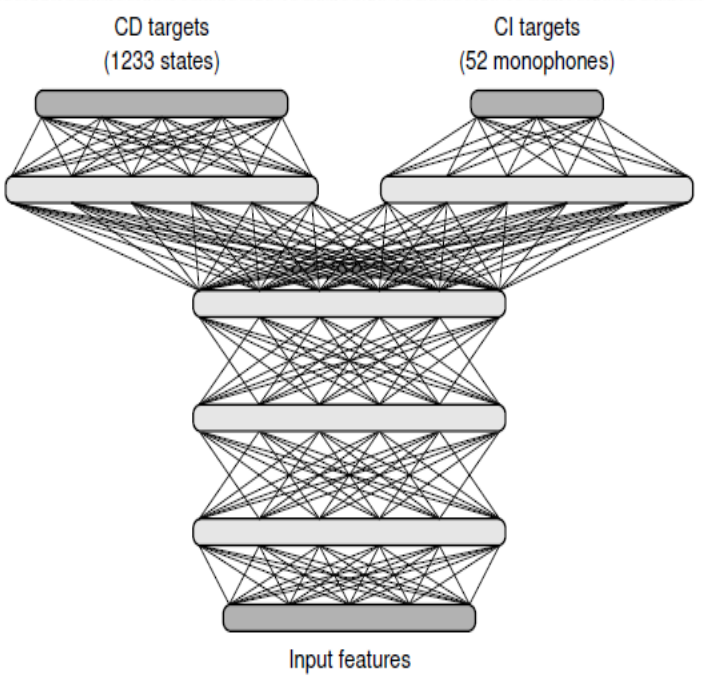
Multi-task training

- The network has to learn more tasks in parallel
  - Dedicated output layers for each task
  - The hidden layers are shared, so they have to learn all tasks
  - (We allowed 1-1 task-specific hidden layers, with a slight improvement)
  - During training, each batch of data is randomly assigned to one of the paths
  - Multi-task training is known to improve the generalization of the network
  - First use in ASR: Microsoft, 2013



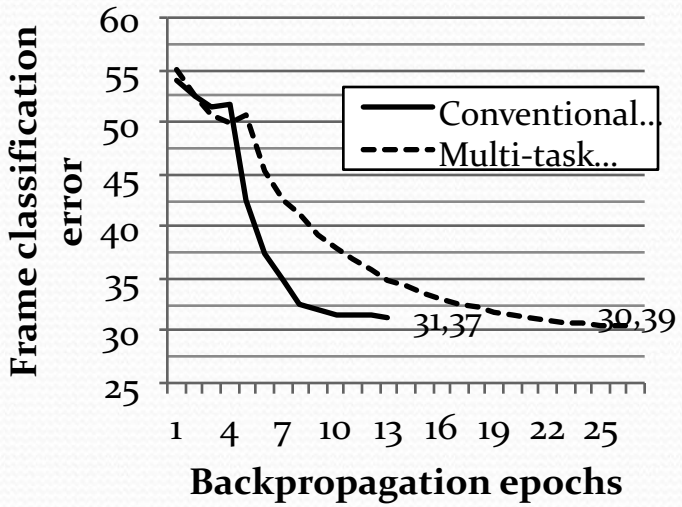
Multi-task training of CD models

- Task 1: CD states, Task 2: phones
  - We have both type of training labels for each training vector
  - During training, the error of both the CD and the CI targets gets minimized
    - CI targets have a regularization effect
  - During recognition we use only the CD output
  - ***During adaptation we train only the CI output***
    - alleviates the problem of missing labels



Results with multi-task training (no adaptation yet!)

- Data set: 28 hours of Hungarian broadcast news, 1233 states



Training method	FER %		WER %	
	Train set	Dev. set	Dev. set	Test set
Conventional	25.9%	31.4%	17.7%	17.0%
Multi-task	23.5%	30.4%	17.4%	16.5%

- During training: slower convergence but slightly better results
- Final WER: about 3% relative WER reduction

Adaptation experiments

- Our broadcast news corpus is not optimal for adaptation tests
  - The files are not annotated by speaker
  - However, there is no speaker change within a file
  - The duration of files ranges from 3 to 100 seconds
- We experimented only with unsupervised adaptation
  - First, the ASR recognizes the given file using the unadapted DNN
  - Then we perform adaptation training on the given file using the estimated transcript obtained in the previous step
  - Finally, we recognize the file again using the adapted DNN

Refinements to adaptation

- We found that multi-task training with CI units is not enough (the results had a huge scatter, suggesting overfitting)
  - We restricted adaptation to the uppermost shared hidden layer
  - We used the regularization method of Yu et al. (2013)
- KL-divergence based regularization
  - Penalizes when the output of the adapted model strays too far from the output of the unadapted model
  - Formalized using the KL-divergence of the two outputs

KL-divergence based regularization

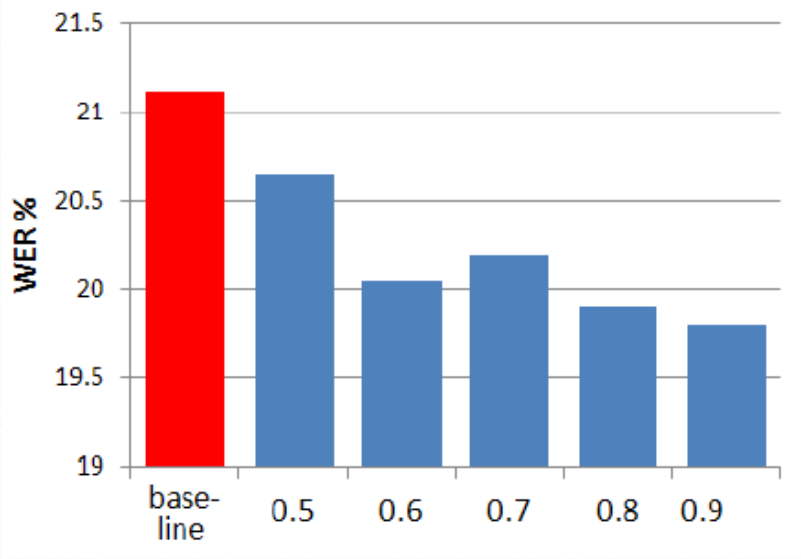
- After some derivation, KL-divergence regularization boils down to smoothing the hard training labels estimated by the recognizer with the output of the unadapted network

$$(1-\alpha)p(y|x) + \alpha p_{un}(y|x)$$

- $p_{un}(y|x)$: output of the unadapted network (0.2 0.2 0.1 **0.1** 0.2 0.2)
- $p(y|x)$: estimated "hard" training targets (0.0 0.0 0.0 **1.0** 0.0 0.0)
- α : linear interpolation weight (e.g. 0.5)
- (0.1 0.1 0.05 **0.55** 0.1 0.1)
- Larger α means we do trust less in the estimated (hard) targets and more in the unadapted (probabilistic) outputs

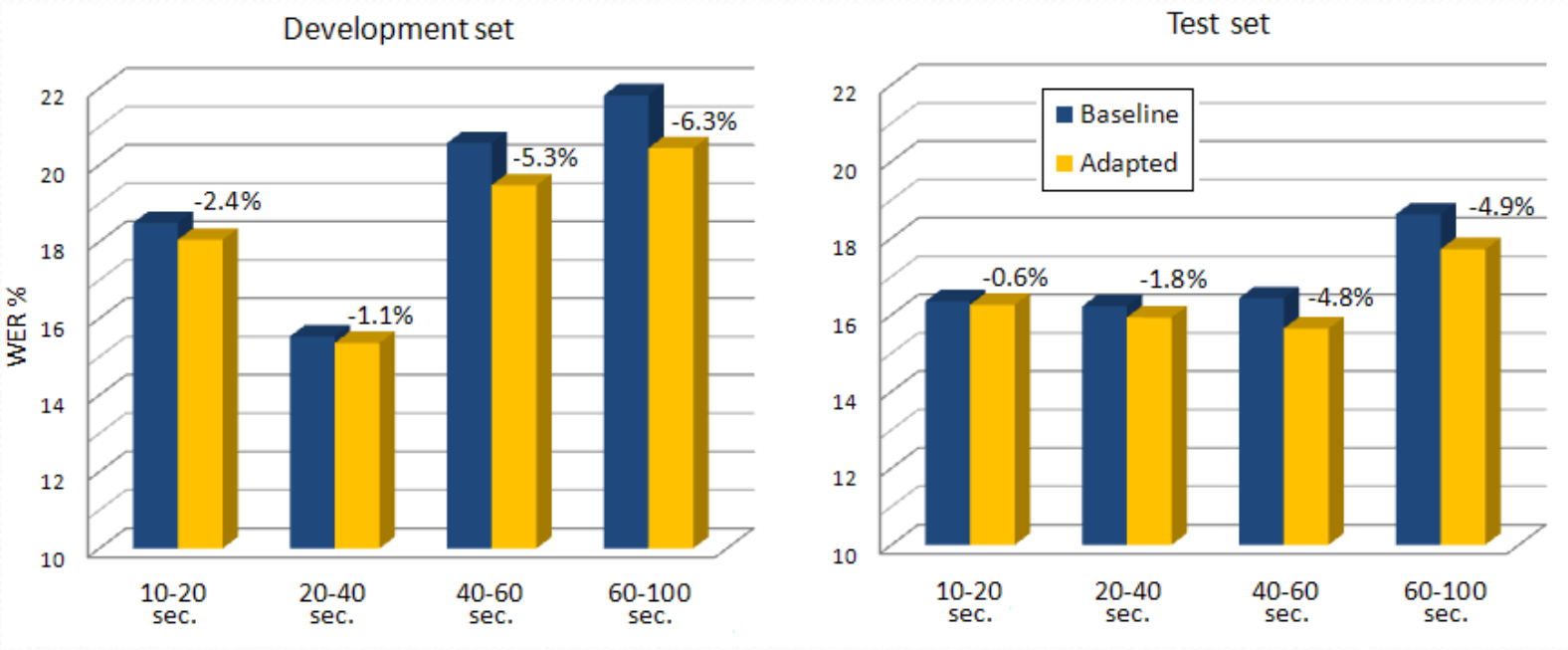
The effect of KL-regularization

- Demonstrated for 40-100 seconds of adaptation data, dev set



- Stable behavior requires strong regularization (α close to 1)
- Note: α=1 makes no sense, as the error becomes zero…

Efficiency of adaptation as a function of adaptation data length



- 10-40 sec of adaptation data seems to be insufficient
- After 40-100 of adaptation, the WER reduction is 5-6% relative

Summary

- Multi-task training of CD and CI units improves the results of recognition (with CD units)
- It yields a trivial way for adaptation using only the CI targets
- Combination with KL-divergence based regularization improves the adaptation results further
- With adaptation data of only 40-100 sec, we could achieve WER reduction of 5-6% relative
- We plan to evaluate the method with longer adaptation times and also with supervised adaptation

Thank you for your attention!

tothl@inf.u-szeged.hu