Improving DNNs for Limited-Resource Acoustic Modeling

Problem
- DNNs become the state-of-the-art for LVCSR tasks, outperforming the traditional GMM-HMM.
- However, DNNs encounter special challenges when the amount of training data is highly limited (e.g., < 10 hours of transcribed speech) (new language, domain, ...)
- A comparison we observe on the BABEL corpus

Solution #1
Dropout
- Proposed as a strategy to curb overfitting in DNN training [1]
- It brings gains to LVCSR performance with highly limited training data [2].

Maxout Networks
- Applied to LVCSR by two ASRU’13 papers[3,4]
- Promising improvement shown under limited-resource conditions [3]
- A generalized version of max-pooling and its effectiveness for both ASR and KWS [5]

Solution #2
Lang-Universal Feature Extractors (LUFEs)
- Multilingual DNNs trained over rich-resource languages act as LUFEs [6]
- How to train them? Hidden layers are shared; features and classification layers are specific

Cross-Language Hybrid Systems
- On a new target language, features are feed into the LUFEs; hybrid systems are built over the deep feature representations
- Maxout networks as sparse LUFEs [3]
- Deep Convolutional Networks as LUFEs [7]

Potential Reasons
1. Large parameter space $\rightarrow$ overfitting
2. Unreliable class labels from weak GMM-HMM
3. Class priors are estimated unreliably

Solution #3, 4, 5, … …

Questions?
- Does the class inbalance in training data matter?
- How can we take advantage of vast amount of untranscribed speech?

We look forward to

Your Ideas