**UNSUPERVISED ACOUSTIC AND LANGUAGE MODEL TRAINING WITH SMALL AMOUNTS OF LABELED DATA**

*Scott Novotney, Richard Schwartz, Jeff Ma*

BBN Technologies, Inc., 10 Moulton Street, Cambridge, MA 02138, USA
JHU HLT COE, 810 Wyman Park Drive, Baltimore, MD 21218, USA

{snovotne, schwartz, jma}@bbn.com

---

**Introduction**

State of the art LVCSR requires hundreds to thousands of (expensive) manual transcriptions.

**Can unlabeled audio reduce this cost?**

(Lamel 02) Self-Training on English Broadcast News (TDT-2)
- 10 min. of transcripts, 135 hours of unlabeled audio
- 1 billion word in-domain LM
- 33% relative reduction in WER

(Ma 08) Self-Training on English CTS (Fisher corpus)
- 1hr of transcripts, 2000 hours of unlabeled audio
- 1 billion word out-of-domain LM
- 47% relative reduction in WER

**How does a weakened LM (external knowledge) impact acoustic model self-training?**
- Use weaker language models of 1M and 100k words.
- Measure the impact of LM on the quality of the AM.

**Will self-training improve language modeling?**
- Estimate n-gram confidences using word confidences.
- Reject or de-weight unlikely n-grams using confidences.
- Directly model n-gram confidences.

---

**Self Training Method**

1. Build initial acoustic and language models from available manual data.
- Acoustic Model
- Language Model
- Unlabeled Audio

2. Recognize untranscribed audio data with initial model.
- Automatic Transcriptions

3. Use word confidences to select or weight observations.
- Filtered Transcriptions

4. Train new models on adapted transcripts and iterate. (twice usually enough)

**WER Recovery Metric**

We gauge success as the relative reduction in WER recovered by self-training compared to models trained with manual transcripts of the same ‘unlabeled’ audio.

Initial WER – Self-Trained WER
- 100% recovery means that the self-trained models performed identically to the supervised models.

---

**Analyzing Self-Training**

We measure WER Recovery as a function of:
- Labeled audio (1 or 10 hours)
- Unlabeled audio (200 or 2000 hours)
- Language modeling text (100k in-domain, 1M out of domain, 1B out of domain)
- Acoustic model or language model training

---

**Trends**

- Self-training is most effective with small amounts of transcribed audio.
- Larger improvement for collecting ten times as much audio as transcribing.
- Language model self-training less effective and smaller impact on WER reduction.

---

**Language Model Impact**

AM self-training requires an LM for training and final decoding of the test set. We separate these effects by varying the LM resources for the same audio condition.

- 10hrs manual transcripts
- 200hrs unlabeled audio
- 100k LM from 10hrs
- 1B LM is 100k LM + BN/Web

A stronger LM improves WER twice:
- Total WER reduction is 10.3% (diagonal)
- Improves WER during decoding (~6.5% horizontal)
- Improves quality of the AM during training (~4% vertical)

**Selection for LM Self-Training**

Unsupervised n-gram counts are the product of individual word confidences (which range from 0 … 1).
- Estimated with a general linear model on dev set.
- 32 features from ASR system such as AM and LM score.
- As usual, the token count is then sum of the type counts.

We can use these counts to reject n-gram types by ranking and thresholding before normal smoothing (Witten-Bell).

---

**Conclusions**

**AM Self-Training**
- Self training provides greatest benefit with small initial models and large amounts of audio.
- Best case: 80% recovery (24% out of 30% absolute WER reduction) with 1+2000hr AM and 1B word LM.
- Worst case: 42% recovery (5% out of 12% absolute WER reduction) with 10+200hr AM and 100k LM.

**LM Self-Training**
- Harder task:
  - No parameter sharing: model memorizes the data.
  - No external knowledge source to correct mistakes.
- Language model self training works, but provides much smaller benefit compared to acoustic models.
- Using oracle confidence increases recovery to only 50%.
- Improved confidences are not the answer.