INTERSPEECH 2009 BRIGHTON

Analysis of Low-Resource Acoustic Model Self-Training

Scott Novotney[12], *Richard Schwartz*[1]

[1] BBN Technologies, Cambridge, MA, USA
[2] JHU Human Language Technologies Center of Excellence, Baltimore, MD

snovotne@bbn.com, schwartz@bbb.com

Abstract

Previous work on self-training of acoustic models using unlabeled data reported significant reductions in WER assuming a large phonetic dictionary was available. We now assume only those words from ten hours of speech are initially available. Subsequently, we are then given a large vocabulary and then quantify the value of repeating self-training with this larger dictionary. This experiment is used to analyze the effects of self-training on categories of words. We report the following findings: (i) Although the small 5k vocabulary raises WER by 2% absolute, self-training is equally effective as using a large 75k vocabulary. (ii) Adding all 75k words to the decoding vocabulary after self-training reduces the WER degradation to only 0.8% absolute. (iii) Self-training most benefits those words in the unlabeled audio but not transcribed by a wide margin.

Index Terms: speech recognition, self-training, English

10.21437/Interspeech.2009-86

1. Introduction

State of the art Large Vocabulary Conversational Speech Recognition (LVCSR) requires large amounts of transcribed audio to train acoustic and language models with high accuracy. Self-trained (or unsupervised, lightly-supervised, semi-supervised – the community isn't quite sure) acoustic modeling tries to lessen these resource requirements and minimize the need for audio transcription or collection of text. The general technique uses a small amount of labeled audio to build an initial (usually poor) model which is then used to generate automatic transcripts on a large amount of unlabeled audio. These automatic transcripts are then used to build a new, much larger (and presumably) much better model.

The initial experiment with acoustic model (AM) self-training established the technique we use in this paper on the Spanish Callhome corpus. (Zavaliagkos, 1998) A model built on three hours of transcriptions was used to decode twenty-five hours of unlabeled audio, with an average word error rate (WER) of 70%. Utterances with a WER below 20% were selected by estimating confidences; however, this resulted in only forty-five minutes of audio – due to the small unlabeled set. The authors instead used manual transcriptions to estimate selection from a 1000 hour corpus, giving a 1.6% reduction in WER over the initial three hours.

Later experiments conducted AM self-training with very small amounts of labeled data and showed that a low WER was unecessary. Lamel et al. (2002) demonstrated that ten minutes of transcribed broadcast audio are sufficient to achieve a 33% relative reduction in WER when using 135 hours of unlabeled audio and a strong broadcast news language model (LM). They further extended this condition to a weak one million word in-domain dictionary.

Recent work by Ma and Schwartz (2008) with conversational English used one hour of labeled audio, 2000 hours of unlabeled audio and a large non-conversational LM. Self-training reduced the starting WER of 51% to 27%. This is an 80% recovery of the WER reduction possible had the 2000 hours been transcribed.

We differ from Ma and Schwartz in that we assume only ten hours of manual transcriptions are available from the domain to derive an AM, LM and pronunciation dictionary. This low-resource starting point reflects certain LVCSR conditions which do not have outside text corpora available, such as colloquial Arabic. Next, we assume a larger dictionary is given to us – such as a set of target keywords – and then wish to recognize these words.

Clearly, retraining with these new words will provide the greatest benefit. However, repeating self-training is not a trivial task. Our experiments with 2000 hours of speech required two weeks per iteration on a state-of-the-art compute cluster. Looking to the future, iterative self-training on ten thousand hours or more will require months of turnaround time. To justify this cost, we measure the gain of retraining with new words versus only including them at decoding time. We then analyze individual word performance based upon this training and decoding partition.

2. Corpus and System Description

2.1. Corpus

We report results using one resource condition from Ma and Schwartz (2008). 200 hours were selected from the English Fisher conversational telephone speech corpus (Kimball, 2004) as the 'unlabeled' set with a ten hour labeled subset used to build the initial acoustic model and language model. We use the three hour Dev04 test set from the NIST Hub5 English evaluation.

In addition to the 100k word in-domain LM (from the initial ten hours), we also refer to two out-of-domain language models. The first consists of 200M words from broadcast news text and 900M words of 'conversational-like' text from the web (Bulyko, 2003). The second consists of 1M words of broadcast news text. All three LMs were estimated with Kneser-Ney smoothing.

2.2. System

We used a multi-pass state-of-the-art LVCSR system with state-clustered Gaussian tied-mixture models at the triphone and quinphone level depending on the decoding step. The audio features are transformed with vocal track length normalization, cepstral mean subtraction and dimensionality reduction using HLDA (Prasad, 2005).

To save time, we only used maximum likelihood estimation instead of discriminative training. We were not overly concerned as the gain of self-training on two thousand hours easily outperforms discriminative training on one or ten hours of manually transcribed speech.

Decoding requires three passes: a forward and backward pass using triphone models and a trigram LM to generate an n-best list, which is then rescored using quinphone acoustic

models and a fourgram language model. These three steps are repeated after speaker adaptation using constrained maximum likelihood regression.

This setup is unchanged throughout our self-training experiments. First, an initial model is trained on small amounts of labeled data. We next estimate word and utterance-level confidences on a development set. Confidences estimate the probability that either a word is correct or an utterance has a WER below a target threshold usually set to one minus the initial WER.

After decoding the unlabeled audio, we select estimate confidences and then select those utterances with confidence greater than 50%. We train a new model (with a larger number of parameters due to automatic heuristics which grow the size of the clusters and pdfs for the system) using the initial manual data and the automatic transcripts. The gain for self-training quickly converges after two iterations, with a third sometimes providing a small benefit.

2.3. Metric

We use the WER Recovery metric introduced by Ma and Schwartz (2008) to gauge success of self-training. Since we have manual transcriptions for the 'unlabeled' audio, we can compare the WER with the initial model (I), self-trained model (U) and supervised model (S).

$$WER\ Recovery = \frac{WER_I - WER_U}{WER_I - WER_S}$$

WER Recovery measures what fraction of the gain from supervised training is recovered by self-training. A recovery of 100% means self-training with automatic labels is as effective as supervised training with manual transcriptions. A negative recovery means that the self-trained model is worse than the initial model, but we have yet to see such results.

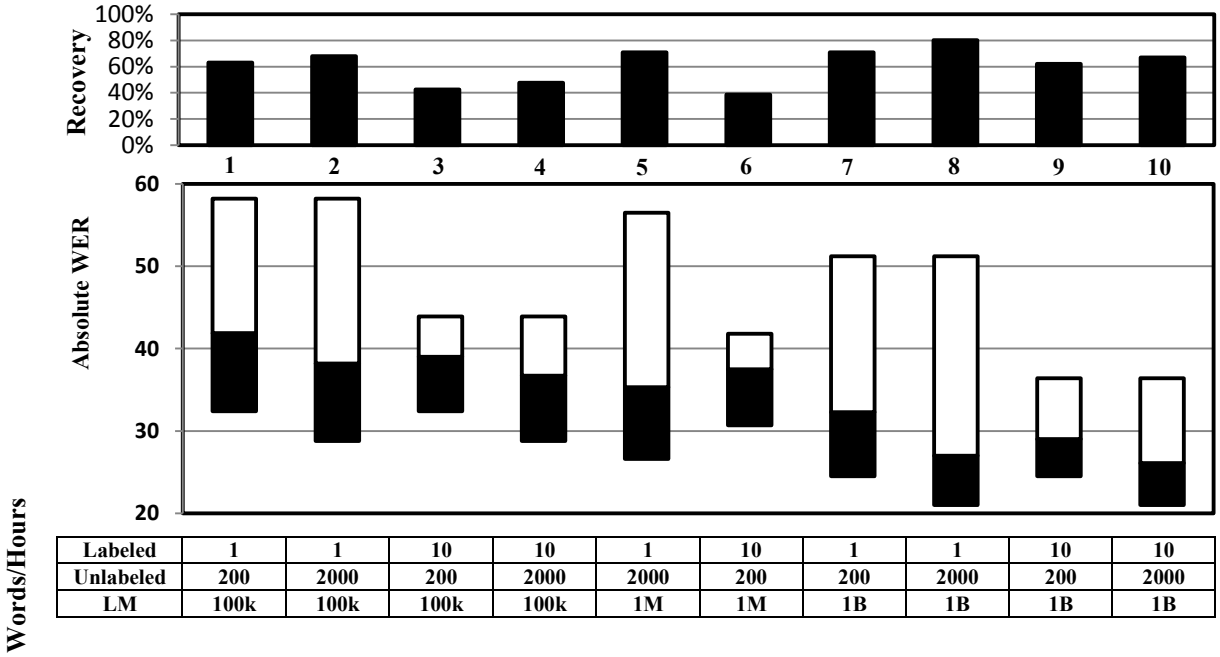
3. Visualizing Previous Results

To show major trends, we introduce a new way to visualize self-training experiments. Each WER Recovery number requires three models: an initial, self-trained and upper-bound fully supervised model. We vary the amount of labele and unlabeled audio as well as the amount of text used for a language model.

Additionally, we either self-train the acoustic or language model. While some scenarios may seem artificial, they help demonstrate self-training trends. These results come from our own recent experiments and the previous work of (Ma and Schwartz, 2008) and are shown in Figure 1.

First, notice that self-training gives substantial reductions in WER for all tested conditions. However, it is most beneficial when the amount of initial labeled data is small (one hour versus ten hours). Additionally, a stronger language model improves performance twice: it reduces the starting WER and also improves the quality of the self-trained acoustic model (increases recovery). Compare the parallel experiments of the three LMs: 100k (bars 2 and 3), 1M (bars 5 and 6), and 1B (bars 8 and 9).

Finally, consider an operating scenario with the available resources of: 1B words of language modeling text, one hour of labeled audio and 200 hours of unlabeled audio (bar 7). There are two options: either collect ten times as much audio (bar 8) or transcribe ten times as much (bar 9). The absolute reduction (24%) for collecting audio is *more* than transcribing (22%). WER Recovery is higher as well (the upper half of Figure 1).Once the 2000 hours are collected, the added benefit of ten hours of transcriptions (bar 10) is minimal – WER is only 0.9% lower than starting with one hour of transcription. These trends continue with the 100k LM (bars 1-4).



Labeled	1	1	10	10	1	10	1	1	10	10
Unlabeled	200	2000	200	2000	2000	200	200	2000	200	2000
LM	100k	100k	100k	100k	1M	1M	1B	1B	1B	1B

Figure 1 – *Prior Experiments in Acoustic Self-Modeling.* Each column is a separate experiment consisting of an initial labeled audio set, unlabeled audio set and a fixed LM detailed in the bottom of the chart. There are three WER values displayed in the middle section. The top of each bar is the WER of the initial AM (one or ten hours). The middle split marks the WER after self-training using the unlabeled audio (200 or 2000 hours). Finally, the bottom of each bar is the lower bound performance had the unlabeled data been transcribed. WER Recovery (the top section) is then the fraction of the entire bar covered by white.

4. Small Vocabularies

We repeat self-training using the 5k vocabulary present in ten hours of manual transcripts instead of a full 75k word dictionary. As seen in Figure 2, this small number of words is sufficient for a low out of vocabulary (OOV) rate. We use 200 hours of 'unlabeled' audio and an LM built from the ten hours – since we assume no additional resources (transcripts or text) are available.

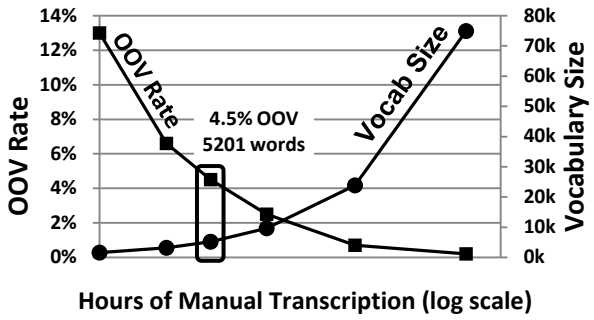


Figure 2 - *Vocabulary growth as a function of manual transcripts*. Calculated on the 3hr test set, 5,201 words types are sufficient to cover 95.5% of the test tokens.

Reducing the vocabulary from 75k to 5k words increases WER by 2% absolute for all three acoustic models despite the OOV rate rising from 0.18% to 4.5%. However, this degradation has almost no impact on self-training: WER Recovery is 41% with the 5k vocabulary versus 42% for the full 75k.

Self-training requires two dictionaries: a *training* dictionary used to decode the unlabeled audio and a final *testing* dictionary to report WER on the test set. When the missing 70k words are added to the *test* vocabulary after self-training of the 5k AM, WER improves to within 0.8% of using all 75k words during training (the gray bar in Figure 3). These additional words have no training samples (manual *or* automatic) and have only back off probabilities in the LM.

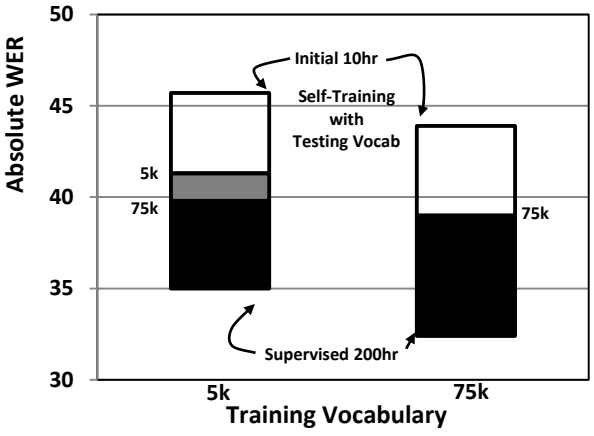


Figure 3 – *Impact of Small Training Dictionaries*. The semantics of these bars are the same as in Figure 1. The right-hand bar is the same as bar 3 of Figure 1, which uses a training and testing dictionary of 75k words. The AM trained with the 5k vocabulary was tested with both the 5k and the 75k vocabulary – the gray bar is the gain for including 70k more words.

5. Detailed Analysis

The experiment in the previous section analyzed the importance of including words in the phonetic dictionary by measuring WER, which is a metric of the *average* performance. But this obscures the behavior of individual words and is heavily influenced by a few most frequent words. We now dig beneath WER and analyze performance of specific words by category.

5.1. Classifying Words by Training Category

A self-training experiment (Section 4) partitions words in the test set into three categories:

- **Labeled** – In the 10hrs of manual transcripts.
- **Unlabeled** – Only in the 200hrs of audio.
- **OOT** – words in neither (Out Of Training).

The 5,201 words in the labeled ten hours dominate the test set tokens, as seen in column 3 of Table 1. From these figures alone, we conclude that words which do not appear in the ten hours of transcripts have negligible impact on word error rate – they just don't occur frequently enough to matter. However, the benefit of self-training on these words should be measured if individual word performance matters (for example, searching for particular rare words).

Vocabulary Set	Type Counts in Test Set	Token Counts in Test Set
Labeled	1,659	34,268
Unlabeled	807	1,520
OOT	147	174

Table 1 – *Frequencies of words in 3hr Test Set.* The types and tokens in the test set are partitioned based upon their appearance in the training data. Words are either in the ten hours of manual transcriptions (*Labeled*), in the unlabeled audio, but not in the transcripts (*Unlabeled*) or not in the training data (*OOT)*. We can compute this distribution since our unlabeled data actually have manual transcriptions.

5.2. Analysis by Training Category

Using the partitions from the previous subsection, we re-analyzed the vocabulary experiments from Section 4: ten hours of labeled audio, 200 hours unlabeled audio and a 100k LM (from the ten hours). Since we didn't count insertions against a word category, we measured word accuracy instead of error rate. Figure 4 details the results.

The first experiment from the previous section used the 5k dictionary for self training and the full 75k dictionary to decode the test set. All three categories of words benefit from self-training. However, *Labeled* words have a much higher initial accuracy and also higher recovery than the other two sets and have both acoustic and language model training examples (since they appear in the initial ten hours). Compare the size of the black recovery bars in Figure 4. The *Unlabeled* and *OOT* improve only through parameter sharing.

The second experiment included all words in the 75k dictionary during self-training and there is a very large benefit for doing so, although absolute performance barely changes. The *Unlabeled* and *OOT* words improve (in terms of Recovery) by four times as much - 16% to 63% and 6% to 28% respectively - from the first experiment (compare the size of the white recovery bars in Figure 4). The absolute

improvement in accuracy - 23% to 32% and 15% to 18% respectively - also significantly improves (the size of the gray bars over the black bars in Figure 4). Since the *Labeled* words were already present in the training dictionary, there is no additional gain in this experiment. Most importantly, *Unlabeled* words now have a much higher recovery than the *Labeled* words in training.

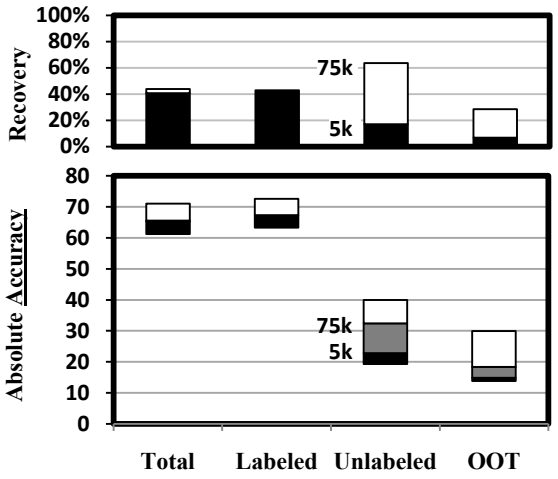


Figure 4 – *Word Accuracies by Training Category*. Instead of WER, we analyzed accuracy, so higher numbers are better. Along with the overall performance (*Total*), columns 1-3 correspond to each of the three categories of words described in Section 5.1. Results start at the bottom of each bar, with initial performance of the 10hr acoustic model. The top line is the upper bound when using the fully supervised 200hr AM. In between these two lines is the performance of self-training when using either the 5k or 75k vocabularies during training. All results used the 75k dictionary during testing. Recovery is shown with 5k results in black and the gain for 75k in white.

6. Conclusions and Future Work

By visualizing previous self-training experiments on the English Fisher corpus, we demonstrated the appropriate scenario for self-training: a small amount of initial transcribed data and a very large amount of unlabeled audio. Even though a strong language model improves WER Recovery, a weak language model is sufficient to improve the acoustic model. Finally, self-training squashes the difference in WER of initial acoustic models. Even though a ten hour supervised acoustic model has a more than fifteen percent lower absolute WER than one hour, the difference after self-training on two thousand hours is less than one percent.

Only a few hours of manual transcriptions are required to generate a vocabulary with low OOV rates (4.5% from ten hours of manual transcriptions) on Fisher English. This smaller vocabulary of five thousand words, while raising absolute WER by 2%, negligibly impacts self-training (measured by WER Recovery). However, the accuracy of individual word types greatly depends on whether they are present during self-training. They can still be recognized and improved on once added to the decoding word models, even though the component triphones may not have been present in training and they only appear in the LM with backoff unigram weights.

Novel words benefit from self-training the most (from 15% to 60% WER Recovery) once they are included in the training vocabulary. Although they may not appear in the manual training, they still improve if they appear in the unlabeled audio (which is assumed to be likely in thousands of hours of speech). The implication is that those words which are not in the initial training (not well trained) but in the unlabeled audio (can improve) benefit the most from self-training.

Earlier work (Ma and Schwartz, 2008) demonstrated that self-training works despite high error rates of greater than 50%. Our recent work demonstrates a more striking case for the *Unlabeled* category of words – those in the unlabeled audio, but not in training. Self-training improves absolute performance from 19% accuracy (over 80% error) to 32% accuracy. Restated, recognizing only one in five words correctly is sufficient to recover 60% of the gain as having all the instances transcribed.

We are extending our self-training work to both Spanish and Arabic to demonstrate success in other languages. One could also apply self-training to domain adaptation, such as from modern standard Arabic to colloquial Arabic. Instead of a poorly trained model (from one hour) we are now trying to improve a well-trained, but *biased* model.

7. References

[1] George Zavaliagkos and Thomas Colthurst, "A comparison of the data requirements of automatic speech recognition systems and human listeners", Proc. EUROSPEECH, 2582-2584, 2003

[2] Ivan Bulyko, Mari Ostendorf and Andreas Stolcke, 2003. *Getting more mileage from web text sources*, in Proc. Of HLT/NAACL, 2003. PP. 7-9.

[3] Owen Kimball, Chai-Lin Kao, Rukmini Iyer, Tresi Arvizo, and John Makhoul. 2004. *Using quick transcriptions to improve conversational speech recognition,* in Proc. Of International Conference on Spoken Language Processing, Jeju, Korea.

[4] Lori Lamel, Jean luc Gauvain, and Gilles Adda, *Lightly supervised and unsupervised acoustic model training*, 2002. *Computer Speech and Language*, vol. 16, no. 1, PP. 115-129.

[5] Jeff Ma and Rich Schwartz. 2008. *Unsupervised vs. Supervised Training of Acoustic Models*, in INTERSPEECH 2008, Brisbane, Australia.

[6] Rohit Prasad et al., *The 2004 BBN/LIMSI 20xRT English conversational telephone speech recognition system*, in INTERSPEECH 2005, Lisbon, Portugal. PP 1645-1648.