UNSUPERVISED ACOUSTIC AND LANGUAGE MODEL TRAINING WITH SMALL AMOUNTS OF LABELLED DATA

Scott Novotney, Richard Schwartz and Jeff Ma

BBN Technologies, Cambridge MA 02138, USA

ABSTRACT

We measure the effects of a weak language model, estimated from as little as 100k words of text, on unsupervised acoustic model training and then explore the best method of using word confidences to estimate n-gram counts for unsupervised language model training. Even with 100k words of text and 10 hours of training data, unsupervised acoustic modeling is robust, with 50% of the gain recovered when compared to supervised training. For language model training, multiplying the word confidences together to get a weighted count produces the best reduction in WER by 2% over the baseline language model and 0.5% absolute over using unweighted transcripts. Oracle experiments show that a larger gain is possible, but better confidence estimation techniques are needed to identify correct n-grams.

Index Terms— Unsupervised Training, Word Confidence, Conversational Telephone Speech, Language Modeling

1. INTRODUCTION

State of the art performance in large vocabulary speech recognition (LVCSR) usually requires hundreds to thousands of hours of manually annotated speech and millions of words of text. But manual transcription is often too expensive or impractical. Even worse, many domains do not have millions of words of text required to build strong language models. However, we can rely upon the assumption that any domain which requires LVCSR technology will have hundreds to thousands of hours of audio available. Unsupervised acoustic or language model training builds initial models from small amounts of transcripts or text and decodes hundreds to thousands of hours of audio. We then train new models using these automatic transcripts. We hope to drastically reduce the labeling requirements for LVCSR in sparse domains.

Previous work on unsupervised acoustic model training (AM-UT) worked remarkably well with very little amounts of labeled training data. Lamel et al. [1] reported that training with only ten minutes of labeled data achieved a 33% relative reduction in WER when using 135 hours of unlabeled audio. They saw gains with both large (one billion) and small (one million) word *in-domain* LMs. Recent work by Ma and Schwartz [2] used a strong *out-of-domain* LM and one hour of labeled audio to decode 2000 hours of unlabeled audio. There was a dramatic gain from a starting WER of 51.2% to 27.0%. In this paper, we will extend these results and measure the impact on AM-UT using weaker out-of-domain language models. Inspired by the success with AM-UT, we will consider several methods for producing n-gram counts for unsupervised language model estimation (LM-UT). Then we will attempt to improve word confidences and directly model n-gram confidences and understand the limits of confidence estimation for language modeling with oracle experiments.

1.1. Corpus and System

All results are on English conversational telephone speech (CTS), primarily using the Fisher corpus [3]. We constructed a 2000 hour corpus out of 2300 hours of Fisher data, balancing for gender. Out of this larger set, we selected smaller labeled and unlabeled training sets, the smaller sets always being subsets of the larger. The terminology 1+2000 means one hour of supervised audio plus 1999 hours (2000 - 1) of unlabeled audio. All reported error rates are on the three hour Dev04 test set from the NIST Hub5 English evaluation. We also verified our results with the six hour Eval03 set. Our language modeling text consists of 1.1 billion words: 200M from broadcast news text and 900M words of 'conversational-like' text from the web [4]. We select subsets of these two sources to model acoustic and language model resource conditions.

BBN Technologies' speech recognition system, BYBLOS, is a multi-pass LVCSR system that uses state-clustered tied-mixture models [5]. To save time, we only used maximum likelihood estimation instead of discriminative training. Decoding requires three passes: a forward and backward pass to generate an n-best list, which is then rescored using a strong language model. These three steps are repeated after speaker adaptation using CMLLR.

1.2. Metric

We use the "WER Recovery" metric introduced by Ma and Schwartz [2] to gauge success of our unsupervised tech-

We would like to thank the JHU HLTCOE for funding this effort and our colleague Damianos Karakos from CLSP-JHU for his insightful advice.

niques. Since we have manual annotations for our "unlabeled" audio set, we can compare the WER with initial models (I), unsupervised models (U), and supervised models (S).

$$\text{WER Recovery} = \frac{\text{WER}_I - \text{WER}_U}{\text{WER}_I - \text{WER}_S} \quad (1)$$

WER Recovery measures what fraction of the gain from supervised training can be recovered by unsupervised training.

2. UNSUPERVISED ACOUSTIC MODEL TRAINING

In [2] Ma and Schwartz compared different unsupervised training strategies and found that the "use-all-data" strategy works better than any incremental bootstrapping procedure. Their approach was to first build acoustic and language models with the limited available manually labeled data. Using these poor initial models, they decoded the unlabeled audio and then estimated the confidence of each utterance to have a WER below a threshold [6]. Finally, they rejected low confidence utterances (25% - 50% depending on the quality of recognition) and trained new acoustic models. They saw a small additional gain after a second iteration of this process but not typically after a third. We use this strategy in all experiments reported in this paper.

2.1. Results with Weaker Language Models

Here, we study the impact of language model quality on unsupervised acoustic training. While we expect WER Recovery, from eqn. (1), with weaker language models to degrade, the hope is that unsupervised training performance will still be robust and give significant gains. We report results for a variety of acoustic conditions with three language models: 1.1 billion words of broadcast news and web data (1B), 1 million words of broadcast news (1M), and the 100k words from the initial 10 hours of transcripts (100K). Results with the 1B word LM come from Ma and Schwartz [2] and are directly comparable to our results.

Table 1 details results for various starting conditions. Of course, the baseline and supervised WER are higher with weaker LMs. But the WER Recovery numbers also show that the weaker the LM becomes, unsupervised training is able to recover less of the gain from supervised training. Starting with the 1B LM, WER Recovery degrades by 10% on average for the 1M word LM and by 20% for the 100k LM. As the amount of data increases, the unsupervised training recovers more of the supervised gain. This trend, first established by [2], continues to hold with weaker LMs. Our weakest condition is with only the ten hours (100k words) of manual transcripts (last two lines of Table 1). The baseline and supervised WER increase by 7% absolute and the WER Recovery with 2000 hours of audio decrease to 48% compared to the 1M LM. It is clear that a strong external knowledge source increases both absolute performance and WER Recovery.

LM	Audio	Base	Unsup	Sup	Recovery
1B BN	1+32	51.2	38.7	30.1	59%
	1+200	51.2	32.3	24.5	70%
	1+2000	51.2	26.8	21.0	80%
	10+200	36.4	29.0	24.5	62%
	10+2000	36.4	26.1	21.0	67%
1M BN	1+32	56.5	46.9	36.6	47%
	1+2000	56.5	35.3	26.6	71%
	10+200	41.8	37.5	30.7	38%*
100k	10+200	43.9	39.0	32.4	43%
	10+2000	43.9	36.7	28.8	48%

Table 1. **Unsupervised Training with Different Language Models** - These results are with various combinations of labeled/unlabeled audio (Column 2)and initial acoustic models. The baseline (Column 3) uses an acoustic model trained only on the labeled audio. Unsupervised training (Column 4) is after 2 iterations and we compare that to the supervised condition (Column 5) of using manual transcripts for all available audio and report WER Recovery (Column 6). The 1M LM degrades Recovery by 10% on average and the 100k LM by 20%. *The 10+200 condition with 1M LM is after only one iteration.

2.2. Direct Comparison of Acoustic Models

From Table 1, one could assume that the language model is the key to high WER Recovery on AM-UT. However one must differentiate between the LM used for training and the one used for decoding the test set. Since we performed the 10+200 training condition with all three LMs, we can separate the effect of these two different LM uses. Going from top to the bottom of Table 2, one sees the 1B LM improves AM-UT by 4% on average. However, its direct impact on decoding the test set is 7.5% as can be seen when going from left to right. The implication is that the quality of the LM influences absolute WER more than the quality of the unsupervised acoustic model.

Training LM	Decoding LM		
	100k	1M	1B
100k	39.3	37.7	32.5
1M	39.3	37.5	32.3
1B	35.0	33.0	29.0
Baseline 10hrs	43.9	41.8	36.4

Table 2. **Direct Comparison of Acoustic Models** WER on Dev04 with the effect of the LM separated out. LMs are used to decode the unlabeled audio for training and also to decode the test set to report WER. The 1B LM produces a 4% better AM, but decreases WEr by 7.5% from the 100k LM.

3. UNSUPERVISED LANGUAGE MODEL TRAINING

The fundamental mechanism for unsupervised training is that an external source corrects errors the initial model might make, thus increasing the quality of the automatic transcriptions and estimated acoustic models. AM-UT uses the lan-

guage model and the phonetic dictionary as this external source to directly compensate for errors made by the AM. In contrast, LM-UT uses the existing AM (and LM) to generate new strings of words that never occurred in the training text.

To find the best method of generating these n-gram counts, we measured the benefit of using *word confidences* to threshold or weight n-grams which appear in the ASR output after AM-UT. These come from a general linear model (GLM) trained to predict the probability of a word being correct given features from the recognizer [7].

We first trained an LM on all of the automatic transcripts without weighting. This is the simplest method to learn new n-grams and improve the initial LM. We used the unsupervised AM trained with the baseline 100k LM to decode 2000 hours and build an LM from the counts. This resulted in an improvement over the 100k LM from 36.7% to 35.2%. In a second experiment, we trained on all n-grams present in the *decoding lattices* using expected counts and we saw no significant improvement over the 100k LM. The fact that the large number of incorrect n-grams in the lattice overwhelms the correct n-grams is the probable cause of this degradation in WER. Extracting these correct n-grams from the lattices is too challenging and so we focus on improving one best estimates.

3.1. Combining Word Confidences

Since rejecting unlikely utterances for acoustic model training increased performance, we explored similar methods for LM-UT, this time thresholding on n-grams instead of utterances. Specifically, we rejected n-gram observations from the decoded 2000 hours with an estimated confidence below some threshold. Also, we weighted the n-gram counts by the estimated confidence. We used the product of the constituent word confidences to compute the confidence of n-grams, leading to a reduction in perplexity on our test set from 209 (for the baseline 100k LM) to 144[1]. This nicely matches our intuition that the probability of an n-gram being correct is the probability of the constituent words being correct.

Our first experiment used a 200 hour subset from the 2000 hour unlabeled set so that we could experiment with many different conditions, giving us insight into confidence thresholding. Table 3 shows WER on our test set using estimated n-gram counts. These counts for each LM were either selected randomly and without weighting, or thresholded and weighted by confidence. Going from the top to the bottom of Table 3, WER decreases, which suggests that confidence selection does not give any benefit. As we compare columns two and three, weighting n-grams by their confidences improves WER by 0.3% to 1.1% over the baseline 100k LM. However, as can be seen from the fourth column, this is only 16% of the supervised 6.7% gain for training on manual transcripts, leaving plenty of room for improvement.

[1]Other simple combination methods, such as averaging, were significantly worse.

	N-gram Selection Method		
% Selected	**Random**	**Weighted**	**Trans**
0	44.2	44.2	44.2
10	44.6	44.2	41.1
25	44.0	43.7	38.7
50	43.5	43.7	38.7
75	43.6	43.2	38.3
100	43.4	43.1	37.5

Table 3. **WER for Unsupervised LM Techniques** - Text chosen from 200 hours of one best hypotheses using 100k LM and unadapted 10hr AM. There is no benefit to selection, either randomly choosing n-grams or ranking and weighting by the product of word confidences. Confidence weighting improves over unweighted counts by from 0.8% to 1.1% absolute from the baseline. Yet only an additional 10% of manual transcripts is more powerful than counting all weighted n-grams.

Since the previous experiment showed that the best strategy is to select all data and weight counts by the confidence, we repeated this experiment with four larger acoustic conditions. Table 4 shows the results. The four acoustic models were trained with AM-UT using the baseline LMs. Since we have manual transcripts for the unsupervised set, we can select only those n-grams from the decoded transcripts which are actually correct, giving us an upper bound on n-gram selection, simulating perfect confidences. As with acoustic modeling, WER Recovery increases as we train on more unlabeled audio. LM-UT decreases WER by up to 2% with 2000 hours, but only recovers half the gain from oracle selection. Furthermore, oracle selection recovers at most 53% of the supervised gain when training on manual transcripts. While a 2% reduction in absolute WER is meaningful, the *recovery of LM-UT is much less than AM-UT.*

LM	Audio	Base	Unsup	Oracle	Sup	Rec
100k	10+200	39.3	38.5	37.3	33.9	15%
	10+2000	36.7	34.7	32.8	29.2	26%
1M	10+200	37.5	36.8	36.2	33.3	17%
	1+2000	35.2	33.3	32.0	29.2	32%

Table 4. **Unsupervised Language Modeling across Different Acoustic Conditions** - We report WER on Dev04. Two different LMs (column 1) were used to decode the unlabeled audio, using acoustic models after AM-UT (column 2). The baseline WERs (column 3) were reduced by 0.5%-2.0% after LM-UT (column 4). However, the oracle WERs (column 5) is still far off from the supervised WERs (column 6) when using manual transcripts. WER Recoveries (column 7) are much less than AM-UT. (50%-80%)

3.2. Oracle Selection

The previous oracle results in Table 4 correspond to perfect *selection* - correctly recognized n-grams were selected and all

others rejected. In Table 5, we applied a different strategy reflecting accurate confidence *estimation* where an n-gram has a weight equal to its accuracy. E.g., a trigram with only two words correct was assigned a weight of 0.67. We thresholded n-grams according to their true accuracy and then weighted the counts (column two). We also weighted the selected n-grams by the *estimated* confidence (column three). While confidences correctly discount unlikely n-grams when all data are accepted, they slightly hinder performance as the average n-gram accuracy increases. Our confidence system cannot accurately distinguish correct n-grams, so weighting provides the best compromise.

	Weighting Method	
Threshold	**Accuracy**	**Confidence**
Base	36.7	36.7
0	35.2	34.7
0.25	34.6	34.6
0.5	34.1	34.3
0.66	33.7	33.6
0.75	32.9	33.2
1	32.8	33.1

Table 5. **Results with Oracle Confidences** - WER on Dev04 using oracle confidence weighting. We extend the oracle confidence selection from Table 4 and select n-grams with accuracy less than one (Column 1). We also replace the true accuracy weight with the estimated confidence, but still use oracle selection (Column 2).

3.3. Improved Confidences

The very good performance of oracle selection, compared to fair confidence estimation, implies that better confidence estimates would result in significant gains. To better distinguish between correct and incorrect n-grams, we tried adding various features to our GLM for word confidence estimation. These included:

- The estimated confidences of the neighboring words.
- The product of many feature pairs (to capture non-linear effects that the GLM cannot model).
- The estimates of the true posteriors, where a heldout set was used to tune the estimation.

We used these new confidence features to re-estimate confidences for 2000 hours decoded with the 100k LM and unsupervised AM. This reduced perplexity on Dev04 from 127 to 113, but there was no meaningful reduction in WER. While word confidences may be independently modeled, word accuracy is not independent of the surrounding words. Instead of using the product of the constituent words to estimate an n-gram confidence, we directly modeled n-gram confidences. Using analogous features from the word confidence system, including the product of the word confidences, we built separate GLMs for each n-gram order. This reduced perplexity further to 112, but again with no WER reduction.

4. DISCUSSION

Unsupervised acoustic modeling is very robust even with a very weak LM. While recovery degrades by 10% moving from the 1B to 1M word LM and 20% on average to the 100k LM, recovery is still near 50%. LM-UT benefits from using weighted n-grams from ASR output. The absolute improvement from a weak acoustic model on ten hours of training to two thousand hours supervised is 15% while the comparable LMs only vary by 8%. Additionally, the relative Recovery is much less than acoustic modeling - 15%-30% versus 50-80%. We hypothesize that since n-gram modeling techniques only memorize the observed data, the LM is unable to aggregate observations and smooth out noise. We could increase the gain from LM-UT slightly by repeating acoustic model training and we intend to report our results in a subsequent publication.

5. REFERENCES

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

[2] Jeff Ma and Rich Schwartz, "Unsupervised versus supervised training of acoustic models," in *INTERSPEECH 2008*, Brisbane, Australia, 2008.

[3] Owen Kimball, Chai-Lin Kao, Teodoro Arvizo, John Makhoul, and Rukmini Iyer, "Quick transcription and automatic segmentation of the fisher conversational telephone speech corpus," in *RT04 Workshop*, 2004.

[4] Ivan Bulyko, Mari Ostendorf, and A. Stolcke, "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures," in *Proc. HLT-NAACL 2003*, Edmonton, Canada, 2003, pp. 7–9.

[5] Rohit Prasad et al., "The 2004 BBN/LIMSI 20xRT English conversational telephone speech recognition system," in *INTERSPEECH 2005*, Lisboa, Portugal, 2005, pp. 1645 – 1648.

[6] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *Proc. ICASSP 2006*, Toulouse, France, 2006, pp. 1057–1060.

[7] M Siu and H Gish, "Evaluation of word confidence for speech recognition systems," *Computer Speech and Language*, vol. 13, no. 4, pp. 299–318, 1999.