interspeech 2008 incorporating SST 08

Unsupervised versus Supervised Training of Acoustic Models

Jeff Ma and Richard Schwartz

BBN Technologies

ABSTRACT

In this paper we reports unsupervised training experiments we have conducted on large amounts of the English Fisher conversational telephone speech. A great amount of work has been reported on unsupervised training, but the major difference of this work is that we compared behaviors of unsupervised training with supervised training on exactly the same data. This comparison reveals surprising results. First, as the amount of training data increases, unsupervised training, even bootstrapped with a very limited amount (1 hour) of manual data, improves recognition performance faster than supervised training does, and it converges to supervised training. Second, bootstrapping unsupervised training with more manual data is not of significance if a large amount of un-transcribed data is available.

Index Terms— unsupervised training, supervised training, Fisher conversational telephone speech

10.21437/Interspeech.2008-122

1. INTRODUCTION

A great amount of work has been reported on unsupervised training [1,2,3,4,5,6]. However, no work has been reported to rigorously compare behaviors of unsupervised training with supervised training (with manual transcriptions), especially on a large amount of data, such as thousands of hours.

[5, 8] reports unsupervised training experiments on 153 hours of broadcast news (BN) data. They considered a variety of conditions, such as bootstrapping the training with as little as 10 minutes of manual data and using a small language model (LM) trained with 1.8 million words. Even with the 10 minute data and the small LM, the unsupervised training was able to dramatically decrease the word error rate (WER) from 65% to 37% with the 153 hours data. They tried to compare this performance with supervised training that was conducted on 126 hours of manual data, which had no overlap with the 153 hours. They didn't try more data to see how the training scales up.

[2, 3] reports the first unsupervised training experiments on very large amounts of English and Arabic broadcast news data. But the experiments began with large amounts (~140 hours) of well-transcribed data for both languages. The unsupervised training produced substantial gains on 1,900 hours of audio data, even after the minimum mutual information (MMI) training. Since the 1,900-hour BN English data has closed captions available, they compared the unsupervised training with the lightly supervised training [7, 8, 9], which was designed to take advantage of the availability of closed captions. They found that the unsupervised training produced only slightly worse performance (13.7% v.s. 13.2%).

In this paper we focus on comparisons of unsupervised and supervised training. In addition, we also believe it is important to verify that given a limited amount of transcribed data, the techniques for unsupervised training continue to scale up with larger amounts of speech, and we also want to confirm that the techniques work for less formal styles of speech (other than BN data). Therefore, we extended the work of [5] in two dimensions. Instead of BN, the domain is conversational telephone speech. And we extended the training up to 2,000 hours of speech to confirm that the trend continues (as far as we can follow it today). Finally, we compared the results for unsupervised training with supervised training using the exact same speech.

This paper is organized as follows. Section 2 describes data and performance measures; Section 3 addresses unsupervised training strategies; Section 4 presents results and observations we obtained on large amounts of audio data; Section 5 summarizes this work.

2. DATA, PERFORMANCE MEASURES, AND SUPERVISED TRAINING

2.1. Data

For the reasons described above, we chose the English "Fisher" conversational telephone speech corpus released by LDC to experiment with, because it consists of a large amount (~2,300 hours) of telephone conversational speech and all of the data was transcribed (in a fast way [10]) with a quite reasonable accuracy.

We explored two scenarios in which we were assumed to have 1 hour and 10 hours of manual data, respectively. We performed unsupervised training experiments with three different amounts of audio data, 32, 200, and 2,000 hours. The 1, 10, 32, 200, and 2,000 hour data sets were extracted from the 2,300 hours, and the smaller sets were always subsets of the larger ones. Thus, when we say unsupervised training on the 32, 200, and 2,000 hours later, we actually use 31, 199 and 1,999 hours of un-transcribed audio,

Accepted after peer review of full paper

respectively, if the training is bootstrapped with the 1-hour transcribed data. To set up these data sets, we randomly selected speakers and extracted the first 3 minute speech of each speaker. We tried to balance genders while selecting speakers.

The development set we used here was the Hub5 English Dev04 test set, which includes 3 hours of speech. To validate results, we used the Hub5 English Eval03 evaluation set, which includes 6 hours of data and is completely independent of the Dev04 set.

For language model training, in addition to manual transcriptions (in-domain data), we also used about 1.1 billion (1B) extra words (out-of-domain data), among which roughly 200 million are from BN texts and 900 million from texts that had been pulled from the web by University of Washington [11]. For the scenario in which we had only the 1-hour manual data, we trained a LM with the 1B words plus the 1-hour manual transcriptions (about ~10K words). We denote this LM as "LM1". For the scenario in which we had the 10-hour manual data, we train a LM with the 1B words plus the 10-hour manual transcriptions (~100K words). We denote this LM as "LM10".

2.2. Performance measures

To compare unsupervised with supervised training, we consider two performance. The first measure is WER recovery against supervised training. It is the absolute WER reduction that unsupervised training produces divided by the absolute reduction that supervised training produces. We assume that the upper bound for unsupervised training would be the performance of the supervised training on the same data, which has a WER recovery of 100%. The second measure of performance is a WER ratio. It is the ratio of the WER unsupervised training produces to the WER supervised training produces.

2.3. Supervised training

For the purpose of the performance measures, we trained models on the 1, 10, 32, 200 and 2,000 hours (or with their manual transcriptions (or supervised training), respectively. To save time, we didn't train discriminative models, like MMI. We only trained Maximum-Likelihood (ML) models with an improved speaker adaptation [12]. In all decoding, we used a 70K dictionary and ran two passes, without (1st pass) and with (2nd pass) speaker adaptation. The adaptation normally yields about 15% relative gain. All WERs given in this paper are from the 2nd decoding pass.

Error! Reference source not found. shows the WERs of these models with the two language models, LM1 and LM10, on the Dev04 and Eval03 test sets. As expected, the WER decreases rapidly as we increase the amount of training data. But, after 200 hours the decrease becomes slower. On the 200 and 2,000 hour data, the "LM10" gives slightly better performance (0.3-0.4% absolute). It is due to the use of 9 more hours (=10-1) of manual transcripts in its training.

Table 1: Performance of models trained with 1,10, 32, 200 and 2,000 hours of manual transcriptions and different LMs

Data (hrs)	WERs with LM1		WERs with LM10	
	Dev04	Eval03	Dev04	Eval03
1	51.2	54.5	-	-
10	-	-	36.3	39.9
32	30.1	34.3	-	-
200	24.5	28.1	24.2	27.7
2,000	21.0	24.3	20.7	24.0

3. TRAINING STRATEGIES

Before conducting unsupervised training on the large amounts of data, we explored different training strategies on the 32-hour audio data since the turnaround time was shorter.

There are different ways to add new data to unsupervised training. In [5] the amount of new data added was close to be doubled at each time. We first compared a similar data-doubling strategy with a "use-all-data" strategy. The doubling strategy is to double the amount of data at each iteration of unsupervised training, We used the model from the first hour of manual data to recognize a second hour of audio data. Then a model was estimated using these 2 hours of data. The resulting model was then used to transcribe 2 more hours of new data plus the old 1-hour data. This process was repeated until we had used the full 32 hours of audio data. The "use-all-data" strategy is to use all available data in each iteration. That is, we used the model from the first hour of manual speech to recognize all the 31 hours of audio and then used these transcriptions to estimate a new model, and then we used the resulting model to recognize the speech again for more iterations.

Table 2: compares the results for the two strategies. In the table, the label "1+1+2+4+8+16" signifies that the data was doubled gradually and the label "1+31" indicates that we used all the 31 hour audio data at each iteration. As we can see, the data-doubling performs slightly better than the 1st iteration but worse than the 2nd iteration of the "use-all-data" strategy. The doubling strategy actually decoded 57 hours (=31+26) of data, which is close to the 62 hours that the "use-all-data" strategy decoded with 2 iterations. So the "use-all-data" method is better (1.7% absolute).

In [3] we developed an explicit method to estimate confidence for hypotheses, in which the accuracies are directly involved, and we found the data selection based on the estimated confidence performed better than the use of all data. We used the same data selection on the 32-hour data. The last two rows in Table 2: show the results, where

"1+24" and "1+25" means 24 and 25 hours were selected from the 31 hour data at the 1st and 2nd iterations, respectively. In this paper we use "n+m" to represent "n" hours of initial manual data plus "m" hours of selected data. As can be seen, there is 1.2% absolute gain for discarding about 25% of the automatically transcribed audio data at the 2nd iteration. For simplicity, we call the "use-all-data" strategy plus the data-selection a "use-all-selected" strategy. We used this strategy in all experiments reported later.

Table 2: Comparison of unsupervised training strategies

Training data	WERs on Dev04
1+1+2+4+8+16	41.6
1+31 (1st iteration)	41.9
1+31 (2nd iteration)	39.9
1+24 (1st iteration)	41.1
1+25 (2nd iteration)	38.7

The supervised training with the 1-hour and 32-hour manual transcriptions yielded WERs 51.2% and 30.1% (shown in **Error! Reference source not found.**), respectively. So, for the "use-all-selected" strategy (at the 2nd iteration), the two performance measures, the WER recovery and the WER ratio, are 59.2% = (51.2-38.7)/(51.2-30.1) and 1.28 =38.7/30.1, respectively.

4. EXPERIMENTS ON LARGE AMOUNTS OF DATA

After choosing the training strategy, we greatly increased the amount of un-transcribed audio. On the 32-hour data the best unsupervised training produced about 60% of the gain that the supervised training produced. The question is whether we can scale up this benefit as we increase the amount of audio data. So we conducted experiments with the 200 hour and the 2,000 hour audio data. We explored the two scenarios, in which the training was bootstrapped with the 1-hour and 10-hour manual data, respectively.

Table 3: shows results of the unsupervised training that was bootstrapped with the 1-hour manual data on the 200 and 2,000 hour data. The first column lists the amounts of audio data. The column "Iter" indicates the iteration of the unsupervised training conducted on the corresponding amounts of audio, and the "-" sign in this column denotes the supervised results on the same audio data (shown again here for convenience).

On the 2,000 hour audio, besides the "use-all-selected" training strategy, we also tried an incremental strategy (the row labeled "2000+inc"). That is, we used the 2nd iteration model trained on the 200 hour data to transcribe the remaining 1,800 hours (=2,000–200 hours) in the 1st iteration, and then we used the newly trained model to transcribe all the 2,000 hours at the 2nd iteration. For the "use-all-selected" strategy we ran the 3rd iteration on the 2,000 hour data, because the WER reduction was large (4.6% absolute) from the 1st to the 2nd iterations. For the incremental strategy, there was no need to run an extra iteration because the gain is small (1.1%) from the 1st to the 2nd iteration. The "use-all-selected" strategy produces slightly better performance (0.2% absolute) than the incremental strategy. It verifies what we observed on the 32-hour data. But the incremental strategy saves time since it conducts one less iteration on the 2,000 hour data.

Table 3: Performance of unsupervised training on the 200 and 2,000 hour data, bootstrapped with the 1-hour data

Audio (hrs)	Iter.	Train data (hrs)	WERs	
			Dev04	Eval03
1	-	1	51.2	54.5
200	1st	1+150	35.7	39.6
	2nd	1+155	32.2	36.0
	-	200	24.5	28.1
2000	1st	1+1365	32.5	36.4
	2nd	1+1615	27.9	32.1
	3rd	1+1645	26.8	30.6
	-	2000	21.0	24.3
2000+inc	1st	1+1637	28.1	32.0
	2nd	1+1702	27.0	30.8

Table 4: shows the results on the 200 and 2,000 hour data when the training was bootstrapped with the 10-hour manual data. We didn't think the 0.2% difference affected any conclusions we would draw, so we used the incremental strategy on the 2,000 hour data to save time.

Table 4: Performance of unsupervised training on the 200 and 2,000 hour data, bootstrapped with the 10-hour data

Audio (hrs)	Iter.	Train data (hrs)	WERs	
			Dev04	Eval03
10	-	10	36.3	39.9
200	1st	10+147	29.7	34.0
	2nd	10+166	28.8	32.7
	-	200	24.2	27.7
2000+inc	1st	10+1238	26.1	30.3
	2nd	10+1772	25.5	29.2
	-	2000	20.7	24.0

From these results, first we see that as the amount of audio data increases, the WER from the unsupervised training continues to decrease, even faster than it does from the supervised training. Second, after the 2,000 hour data was added, the unsupervised training in the two scenarios yielded very close performance (27% v.s. 25.5%), although the initial performance was significantly better with the 10-hour manual data (15%=51.2%-36.3%). So manually transcribing more data is not crucial for unsupervised training to achieve decent performance, if a large amount of data is available.

In Table 5:, we show the WER recovery and the WER ratio measures computed at the last iterations of the unsupervised training shown in Table 4:.

Table 5: WER recoveries and WER Ratios of the unsupervised training on the 200 and 2,000 hour data, bootstrapped with the 1 and 10 hour manual data

Manu data	audio (hrs)	Dev04		Eval03	
		WER recovery	WER ratio	WER recovery	WER ratio
1 hour	32	59.2%	1.28	57.9%	1.25
	200	71.1%	1.32	70.1%	1.30
	2,000	80.8%	1.28	79.1%	1.26
	2,000+inc	80.1%	1.29	78.5%	1.27
10 hour	200	62.0%	1.18	59.0%	1.18
	2,000+inc	69.2%	1.21	67.3%	1.22

We see that the WER recovery of the unsupervised training increases as more audio data is added. In the scenario with the 1-hour manual data, the WER recovery measured on the Dev04 test set is 59.2% on the 32-hour audio data and increases to 71.1% on the 200 hour data and then to about 80% with the 2,000 hour data. In the scenario with the 10 hour manual data, the WER recovery is 62.0% on the 200 hour data and then increases to 69.2% on the 2,000 hour data. The recoveries on the Eval03 are about the same as on the Dev04 set. The unsupervised training bootstrapped using less with data, although starting with a worse point, decreases WER faster. The WER ratio may be more easily understood and also better-behaved. It appears to remain constant as the amount of data increases. It is roughly 1.3 for the scenario starting with the 1 hour manual data and 1.2 for the scenario with the 10 hour data on both test sets. Naturally, as the amount of initial transcribed data increases, this ratio will approach 1. The fact that this ratio keeps roughly constant for a given amount of initial training means that as the WER of the supervised training decreases with more training data used, the absolute difference between unsupervised and supervised training will decrease. So, as we get more audio data, there is less motivation for manually transcribing the audio data.

5. SUMMARY AND FUTURE WORK

We have shown that as the amount of un-transcribed audio data increases, the unsupervised training continues to decrease the WER faster than the supervised training does, and the supervised training bootstrapped with less manual data (1 hour) deceases the WER faster than bootstrapped with more data (10 hours). We have compared the unsupervised training with supervised training by using two measures. The WER recovery of unsupervised training keeps increasing as more data is added. In our experiments it was as high as 80% when the initial training was 1 hour and the total amount of audio was 2,000 hours. The WER Ratio seems to depend only on the initial amount of manual data and remain constant as more data is added. This characteristic implies that there is less motivation to manually transcribe more data if a great amount of audio data is available. In our experiments, it remained roughly at 1.3 and 1.2 when the initial data was 1 and 10 hours, respectively.

It has generally been assumed that a good LM is the key to success with unsupervised training. The assumed mechanism for improvement is that the LM causes the system to recognize some of the training words that would not be recognized correctly with only the acoustic model. The LMs we used here were trained with 1 billion words, which can be considered a powerful one. To investigate effect of weaker LMs on the unsupervised training, we have trained a LM with only 1 million BN words. Although the absolute WER is higher, our initial experiments with this weak LM have been showing similar trends as those reported in this paper.

6. REFERENCES

[1] G. Zavaliagkos and T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," in DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, USA, Feb. 1998, pp. 301-305

[2] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised Training on Large Amounts of Broadcast News Data," in IEEE Int. Conf. On Acoustics, Speech, and Signal Processing, Toulouse, France, May 2006, Vol. 3, pp. 1057-1060.

[3] J. Ma and S. Matsoukas, "Unsupervised training on A large amount of Arabic Broadcast news Data", in IEEE Int. Conf. On Acoustics, Speech, and Signal Processing, 2007.

[4] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: recent experiments", Proc. Eurospeech 99, pp 2725-2728, September 1999.

[5] L. Lamel, J. L. Gauvain, G. Adda, "Unsupervised acoustic model training", Proc. ICASSP 2002, pp 877-880, 2002.

[6] C. Gollan, S. Hahn, R. Schlüter, and H. Ney. "An improved method for unsupervised training of LVCSR systems", In Interspeech 2007, pages 2101-2104, Antwerp, Belgium, 2007.

[7] L. Nugyen, B. Xiang, "Light supervision in acoustic training", Proc. ICASSP 2004, pp 185-188, 2004.

[8] L. Lamel, J.L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training", Computer Speech and Language, 16(1):115-229, 2002

[9] H.Y. Chan, P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training", ICASSP2004, pp 737-740, 2004.

[10] O. Kimball, C. Kao, R. Iyer, T. Arvizo and J. Makhoul, "Using quick transcriptions to improve conversational speech recognition", in Proc. of Int. Conf. Spoken Language Processing, Jeju, Korea, Sept. 2004.

[11] I. Bulyko and M. 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, PP. 7-9.

[12] S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections," *ASRU 2003*.