USING UNTRANSCRIBED TRAINING DATA TO IMPROVE PERFORMANCE

George Zavaliagkos, Manhung Siu, Thomas Colthurst and Jayadev Billa

BBN Technologies
GTE Internetworking
Cambridge, MA 02138

ABSTRACT

This paper explores techniques for utilizing untranscribed training data pools to increase the available training data for automatic speech recognition systems. It has been well established that current speech recognition technology, especially in Large Vocabulary Conversational Speech Recognition (LVCSR), is largely language independent, and that the dominant factor with regards to performance on a certain language is the amount of available training data ([4]). The paper addresses this need for increased training data by presenting ways to use untranscribed acoustic data to increase the training data size and thus improve speech recognition.

1. INTRODUCTION

In the past few years, the Large Vocabulary Conversational Speech Recognition (LVCSR) community has attempted to address the problem of speech recognition on languages other than English. The data collected towards this goal resulted in a number of corpora in English, Spanish, Arabic, German, Mandarin and Japanese. These corpora are generically referred to as the Callhome corpora[1].

In a number of NIST-sponsored evaluations two facts became self evident. First, that Callhome recognition is a very hard task. Conversations consist of totally unconstrained conversational speech among familiar talkers with other compounding problems such as the use of foreign words, high out-of-vocabulary rates, overseas channel noise and simultaneous speech from more than one speaker. Second, that the technology currently used for English is very much portable to other languages. The dominant factor in determining the performance in a particular language being the amount of data available for training in that language.

Consider the problem of building a recognizer in a new language, where none or very little training data is available. From an operational point of view, it would be ideal to only have to transcribe a few hours of speech, build a recognizer, and use this recognizer to process large quantities of untranscribed training data (which is presumably much easier to collect). If the recognizer has the ability to identify areas where automatic transcription is sufficiently accurate, we could then feed this data into the training pool, thus enlarging the training set size with the attendant improvement in system performance.

In this paper we will address this possibility of using untranscribed data to improve system performance. In particular, we will present a paradigm to automatically transcribe data and then explore various issues such as tradeoffs between accuracy and size of transcribed data, efficiency of this paradigm as it relates to initial system performance and also how the nature of the transcribed data affects the new system performance.

The paper is organized as followed: Section 2 gives a overview of Callhome Corpora and current state-of-the-art performance across languages. Section 3 describes the paradigm for "using untranscribed data". Section 4 explores the various considerations and tradeoffs involved with this procedure as well as providing various simulation results. Section 5 discusses the results. Section 6 then presents a brief summary and final conclusions.

2. OVERVIEW OF CALLHOME CORPORA

Callhome corpora present a unique challenge for speech recognition research. The data consists of spontaneous speech between familiar parties with attendant dysfluencies posing challenges in itself coupled with small training data sets. A typical recognition task in Callhome evaluations is a (roughly) 5 minute conversation between two or more talkers. Most systems perform a two-pass recognition: a first pass that generates tentative hypotheses which are used to adapt the recognition model to each of the talkers and a second pass that recognizes using these adapted models. A typical evaluation test set contains 20 such conversations. Performance on the latest NIST evaluations across languages for the BBN Byblos system are given in Table 1. As we see in Table 1,

Language	Training speech	Available text	Word Error
English	150hrs	3M words	53.7%
Spanish	60hrs	0.8M words	57.4%
Arabic	18hrs	0.3M words	59.6%

Table 1: Callhome recognition training data and performance across 3 languages.

there are only small differences across languages, with the error rate ranging form 53% to 60%. The high error rate is attributed to the difficulty of the tests. For example, typical OOV rates for Callhome tests is 3-4%, and for approximately 1% of the test words even human transcribers failed to provide any transcription. Looking at the amount of training data available for each of the three languages[2] we see that the amount of training data available correlates reasonably well with the performance across languages.

Another fact that comes out from the Callhome evaluations is

[1] after the data collection protocol which involved offering callers free phone calls to their native country.

[2] We should note that not all 150hrs are of English Callhome are Callhome data. The data include 134 hours of Switchboard and 16 hours of Callhome training.

the portability of technology developed for English. For example, Table 2 shows the gain due to adaptation and Speaker Adaptive Training (SAT) [2] in various languages. It is remarkable that we get the same gain in terms of absolute reduction of the word error rate at three different operating points: approximately 5% absolute reduction in error rate.

	Word Error %		
	Switchboard	Spanish	Arabic
SI	32.3%	64.1%	66.7%
SI-adapted	28.2%	61.1%	62.6%
SAT-adapted	27.2%	59.3%	61.5%

Table 2: Gains due to adaptation and SAT for Switchboard and Foreign Callhome

3. A PARADIGM FOR UNSUPERVISED TRAINING

As described in the previous section, reasonable sized corpora are available for the few Callhome languages. However, when we want to port to a new, different language *quickly*, we can only expect small amounts of training data to be transcribed and available. We would like to explore whether we can use *untranscribed* data (presumably available in huge quantities) to enhance the performance of models built on minimal amounts of available training. In particular, we will assume that

- A text corpus, not necessarily in domain, is available.
- A couple of hours of speech is transcribed -preferably small amounts of data from many speakers.
- Much more untranscribed data is available.

The paradigm is as follows:

- Create an initial model from the available transcribed data.
- Decode all the untranscribed data using the initial model with available language modeling.
- Estimate a confidence score indicative of the decoder "confidence" in the correctness of the hypothesized word for each word.
- Under the assumption that the output confidences correlate to true performance, select a threshold on the confidence: words below this threshold are discarded, and the accuracy on the retained words is controlled by the value of this threshold.
- Add viable transcriptions to the training data set and retrain.

The selection procedure is best illustrated by an example. Assume that the decoder output was the sentence:

```
sentence-id   "example-utt"
hypothesis:   SIL w1   w2   w3   w4   SIL w5   w6   SIL
start frame:   0    0   21   42   57   63   69   81   101
confidence:        .15  .83  .91  .67        .9  .3
```

and that we decided that the confidence threshold was 0.8. This means that only words w2, w3 and w5 will be kept. For retraining, we will retain and add to the training the following two segments:

```
sentence-id:      "example-utt:part1"
reference:        w2 w3
speech segment:   start 210msec; end 570msec

sentence-id:      "example-utt:part2"
reference:        SILENCE w5
speech segment:   start 630msec; end 810msec
```

Note that our system does not output confidence for silence frames, so we make the assumption that silence frames are retained only if they are next to a word that is retained.

For language model training there is no need to split the sentence. Instead, we map all the low-confidence words to a "garbage" word token, such that the sentence would be added to the language model training as:

```
<garbage> w2 w3 <garbage> w5 <garbage>
```

The garbage word token will not obviously be part of the recognition lexicon.

4. SIMULATION EXPERIMENTS

We simulated the scenario described above twice, first using data from the Callhome Spanish Corpus and second using data form the (English) Switchboard ([1]) Corpus. The n-best frequency, together with language model counts, n-gram scores and acoustic scores were input to a Generalized Linear Model (GLM) trained to generate confidence estimates for each of the words [3].

4.1. Callhome Spanish

For Callhome Spanish, we used 3 hours of transcribed speech to train phonetically tied continuous density models (2000 Gaussians in total). The language model was trained on 42K words. We used the remaining 50 hours of speech in the corpus as the untranscribed training data, and we pruned the search such that recognition was run at 10xRT. The confidence threshold was selected such that the retained data had about 20% error (80% accuracy) retaining about 3hrs. Two test sets were used to evaluate the results: one for speakers that were included in the training, and one for all other speakers. For the in-train set we used some of the remaining speech from the training speakers (approximately 2 hours), and for the out-of-train set we used the same test as the one used in the Fall 1996 NIST evaluation. We refer to the two sets as TrainTest and Eval96. The results of our experiments are summarized in Table 3.

Training data (hrs)		% Word Error Rate	
true	retained	TrainTest	Eval96
3	-	68.9	76.0
3	3	67.3	75.7
6	-	65.9	75.4

Table 3: Callhome Spanish Simulation results

Table 4 presents the trade-offs between the percentage of the data that is retained and their error rate as a function of the prescribed threshold for the Callhome Spanish system. Ideally, we would like to select the portion of the data that has the best possible accuracy. However, as Table 4 indicates, for 87% accuracy we retail only 1% of the data. At the 1995 Fall LVCSR workshop, Dragon Systems presented an experiment where the training transcription where randomly corrupted. The baseline word

threshold	% words retained	%ccorrect in retained data
0.54	15%	59%
0.69	4%	75%
0.71	3%	80%
0.76	1%	87%

Table 4: Trade-offs between accuracy and amount of data retained for confidence thresholding

error rate (no corruption) for this experiment was 55%; results presented indicated that corrupting the data by 20% caused noticeable degradation. We therefore assume that a 20% error rate on the retained data is a minimum.

4.2. Switchboard

For Switchboard we trained two state-clustered tied mixture systems ([4]) with 32,000 and 64,000 Gaussians on 8hrs of speech, and built a language model on 2M words of Switchboard and 100M words of CNN. The decoder was run at 15xRT on 70hrs of speech, and unsupervised speaker adaptation was also performed. In all, 8 hours of data were retained with 8% corruption. The results are summarized in the Table 5 below: From the

Training data (hrs)		% Word Error Rate	
true	retained	32K Gaussians	64K Gaussians
8	-	38.6	37.7
8	8	38.1	37.3
16	-	37.6	36.3

Table 5: Switchboard Simulation results

results we see that there is a gain with addition of the retained data even with increase in model parameters. Also clear from the results is that the retained data help less than "true" data and this difference increases with increase in model size.

In a nutshell, the conclusion of these experiments stands as follows: we observe that performance improves by half as much as it would improve had we added a same amount of humanly transcribed data. The improvement is bigger for the matched speaker, channel and conversation topic condition. The gain for folding the automatically generated data would be bigger if we had increased the number of parameters of the training model. However these conclusions may change when the starting error rates of the system are not hopelessly high.

5. DISCUSSION

Although the result of our experiment is positive, one could take a negative point of view and argue that the improvement is miniscule. To see whether there is any practical implementation of our experiment, let us attempt to answer the following questions:

I. How efficient is the proposed paradigm as a function of starting word error rates? In other words, what if we were starting with a system whose baseline performance was 30% or less, rather than 70%?

II. Is the nature of the retained data an issue? The retained data by design is fragmented since we keep single words or word chunks not sentences.

First consider Question I: To obtain points in the curve for various operating points, we will use data from the Callhome English Spring 1997 test set and the Switchboard-II Spring 1997 test set. The error rate for these two sets with the BBN Byblos system stands at 53.7% and 35.1%. Together with the Spanish results, we have data points for retention based on pre-specified accuracy and confidence estimates for operational points that vary from 35% to 78%.

The results are summarized in Table 6. As we see, the trade-off shifts towards the automatic process. For example, for Switchboard-II we can retain 42% of the data at a 10% corruption, which may mean that we may just need closer to 10 times more untranscribed data to achieve the same effect as transcribed data.

Corpus	W.E.R	error in retained data	retention
SWBD-II	35.1%	10%	42.0%
CHome-Eng	53.7%	15%	17.5%
CHome-Span. 3hr training	68.9%	20%	3.0%

Table 6: Retention versus corruption of retained data for various corpora

Now consider Question II: By design the retained data is fragmented, we pick words or chunks of words rather than sentences. This results in both incorrect time boundaries for the words as well as incorrect insertion of a non-existent sentence boundaries for each retained fragment. To see the effect of this fragmentation we repeated the switchboard simulation experiment using the true segmentation for the retained 8hrs the results are summarized in Table 7. Clearly, fragmentation has a detrimental effect on the

Training data (hrs)		% Word Error Rate	
true	retained	32K Gaussians	64K Gaussians
8	-	38.6	37.7
8	8	38.1	37.3
8+8	-	37.9	37.0
16	-	37.6	36.3

Table 7: Switchboard Simulation results using truth for retained data.

end performance of the retrained system. There are several options to consider to minimize fragmentation: One could require a minimum word length for retained segments. The drawback of this approach would be the inevitable reduction in retention. Also instead of using portions of the new data one could conceive of ways to use all the data but weight it by their confidence scores. One should also consider the fact that as error rates decrease this problem becomes less important.

6. SUMMARY AND CONCLUSIONS

We hope that our experiments give some insight into human learning. We have shown that even an 80% error rate system can improve itself automatically, although requiring large quantities of data and with a slow pace in improvement. We have demonstrated that as the system gets better, the self-learning process also accelerates, in the sense that relatively more of the new data that is encountered can be used to improve the system. For high error rate systems, we have looked into the fragmentation of retained and presented techniques for addressing this problems and their drawbacks.

It should be mentioned that we have omitted two approaches that could improve the behavior of the unsupervised learning experiment. First, once the system improves by some measurable amount, one could conceivably iterate the process and thus increase the system performance. Second, confidence estimation methods have been researched only for the past few years, it is plausible that better confidence estimation algorithms will become available in the future, improving the efficacy of this paradigm.

7. REFERENCES

1. J.J. Godfrey et. al. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, San Francisco, March 1992. IEEE.
2. J. McDonough, T. Anastasakos, G. Zavaliagkos, and H. Gish. Speaker-adapted training on the switchboard corpus. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Munich, April 1997. IEEE.
3. M. Siu, H. Gish, and F. Richardson. Improved estimation,evaluation and application of confidence measures for speech recognition. In *Proceedings of EUROSPEECH-97*, Rhodes, September 1997.
4. G. Zavaliagkos, J. McDonough, D. Miller, A. El-Jaroudi, J. Billa, F. Richardson, K. Ma, M. Siu, and H. Gish. The bbn byblos 1997 large vocabulary conversational speech recognition system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Seattle, May 1998. IEEE.