REDUCING WORD ERROR RATE ON CONVERSATIONAL SPEECH FROM THE SWITCHBOARD CORPUS

P. Jeanrenaud *E. Eide* *U. Chaudhari* *J. McDonough* *K. Ng* *M. Siu* *H. Gish*

BBN Systems and Technologies
70 Fawcett Street 15/1c
Cambridge MA 02138 USA

ABSTRACT

Speech recognition of conversational speech is a difficult task. The performance levels on the Switchboard corpus had been in the vicinity of 70% word error rate. In this paper, we describe the results of applying a variety of modifications to our speech recognition system and we show their impact on improving the performance on conversational speech. These modifications include the use of more complex models, trigram language models, and cross-word triphone models. We also show the effect of using additional acoustic training on the recognition performance. Finally, we present an approach to dealing with the abundance of short words, and examine how the variable speaking rate found in conversational speech impacts on the performance. Currently, the level of performance is at the vicinity of 50% error, a significant improvement over recent levels.

1. INTRODUCTION

We have previously presented approaches on word spotting and topic identification tasks using the Switchboard corpus [2, 3]. Both approaches made use of large vocabulary speech recognition systems. The word error rate at which these systems operated was in the vicinity of 70%. Such poor performance could be partially explained by the difficulty of the task: conversational speech is typically unstructured, and includes frequent pauses, pause fillers, non-speech events, repeats, restarts, and other artifacts. In this paper, we address the issue of recognizing this conversational speech and describe how we are able to significantly reduce the word error rate. In the discussions, we try to provide some insights on the various problems that arise from recognizing conversational speech, as compared to read or prompted speech, and try to quantify their impact on the word error rate. An in depth analysis of understanding and improving recognition performance can be found in [1].

The approach we have taken is to first consider continuous speech recognition (CSR) techniques that have been shown to improve the performance on read speech such as the Wall Street Journal (WSJ) task. These techniques include the use of more complex models, as well as additional training for both acoustic and language models. We also investigated the use of read speech to train the acoustic models. In combination with these techniques, we started to look specifically at ways of reducing the word error rate on spontaneous speech. In particular, we present a word-pairing approach that reduces the effect of frequent short words and examine the effect of the speaking rate on the word error rate.

The paper is organized as follows. In Section 2, we describe the Switchboard corpus and compare it with a corpus of read speech. Next, in Section 3, we describe the experimental paradigm. Then we present a variety of experimental results in Section 4. Finally, we conclude with a discussion of the various results in Section 5.

2. THE CORPUS

Switchboard is a large corpus of conversations recorded over the telephone, in which the two speakers are asked to discuss one of 70 different topics such as pets, crime, or air pollution. There is a total of 2300 conversations (or 4600 sides) with each conversation averaging five minutes in duration.

Segmentation. Because the conversations are typically five minutes, we had to break up the conversation into smaller sentence-sized units for processing. Each side from a Switchboard conversation was segmented automatically at turns and pauses using the time markings that originally came with the data to create these "sentences." Because the time markings are not completely accurate, we estimate that approximately half of the sentences contained small errors at the boundaries (word missing, inserted, or cut-off). Another result of the segmentation algorithm is that a large fraction of these sentences are really partial sentences that are probably harder to recognize than complete sentences, at least from a language modeling point of view.

Training and Test. The entire corpus was partitioned into a training and testing set. The test set was drawn from ten different topics, with an even balance of male and female speakers. All other sides that included these test speakers were taken out from the training. Finally, the test set was cleaned by correcting the transcriptions and adjusting the boundaries to include some silence on both ends. Statistics for the two data sets are described in Table 1.

	# sides	# spkrs	hours	# sent.	# words
TRN	3097	382	143.2	200 K	2 M
TST	20	20	0.5	524	6724

Table 1. Description of training and test set used: Total number of different sides, number of unique speakers, total duration in hours, total number of sentences, and total number of words.

Comparison with Read Speech. In order to characterize the differences between conversational speech and read speech, we compare some statistics computed from the Switchboard data to those computed from the Wall Street Journal (WSJ0) data. These statistics, presented in Table 2, are: Percent coverage with 200 words, percentage of function words, and average length of the words (in phonemes). The main observation is that function words are much more frequent in the conversational data, which translates into a higher coverage with 200 words, and shorter words on average since function words are typically short.

0-7803-2431-5/95 $4.00

	% cov. with 200 words	function words (%)	avg length words
SWB	75	60	3.0
WSJ0	60	38	4.1

Table 2. Weighted coverage with the 200 most frequent words, percentage of function words (from a list of 245 words) in the test sets, and average word length for Switchboard and WSJ0.

3. BASELINE SYSTEM

We used for these experiments BBN's Continuous Speech Recognition System [6], Byblos, which is a semi-continuous, tied mixture system. In this system, context-dependent phonemes are modeled with 5-state, left-to-right HMMs. A set of 52 phonemes is used, including 7 special symbols dedicated to non-speech events such as laughter or breathing noise. A forced alignment pass was used to reject the sentences with errorful transcriptions. The features used are 45 mel-cepstra per 10ms frame, including first and second derivatives. During the training process, context independent units are first trained from flat estimates; the resulting context independent models are then used to bootstrap the training of the context dependent units.

The decoder is a multi-pass system where the last pass uses a lattice word graph that can be expanded to include trigram weights and cross-word models. The final output is a word-dependent N-best list of hypotheses that can be rescored using additional knowledge sources.

3.1. Lexicon and Language Model

In order to pick the lexicon, we measured the unweighted and weighted coverage using the 2k, 5k, 10k and 22k most common words in the training set. Based on the numbers reported in Table 3, we decided that the 5k lexicon seemed like a reasonable trade-off between coverage and system complexity. We did, however, "close" the lexicon by adding the words missing from the test to facilitate the interpretation of the results.

	2k	5k	10k	22k
unweighted cov.	52.8%	75.2%	87.4%	94.6%
weighted cov.	93.0%	96.7%	98.3%	99.3%

Table 3. Weighted and unweighted coverage using the 2k, 5k, 10k and 22k most common words in training.

For language modeling training, we considered two sets of transcriptions. In the first set, only the 40,000 sentences from the same 10 topics as the test set were used. The second set included the sentences from all 70 topics for a total of 200,000 sentences. We measured the 2-gram and 3-gram perplexity on the test set using each training set and compared it with the perplexity of the WSJ0 task. The results are presented in Table 4. We notice first that even though the additional training transcriptions do not come from the same topics as the test, the 70-topic language model results in lower perplexities than the 10-topic grammar. Another observation is that the 3-gram perplexity is worse than the 2-gram perplexity with the 10-topic set, and marginally better with the 70-topic set. This seems to indicate that we are probably experiencing a lack of training, even with 2 million words. But the bottom line is that the perplexity numbers are fairly high compared to the WSJ task perplexity, thus indicating that the transcription task on Switchboard is a hard problem from a language modeling point of view.

Grammar	Training (# sent)	Training (# words)	2-gram perp.	3-gram perp.
10 topic	40 k	0.4 M	136	138
70 topic	200 k	2 M	128	122
WSJ0	2 M	35 M	110	77

Table 4. Number of sentences, number of words, 2-gram and 3-gram perplexity for the 10-topic and 70-topic training data, compared to WSJ0 5k.

4. EXPERIMENTAL RESULTS

We describe in this section experimental results on the test set. First, we concentrate on improving the acoustic modeling by using more complex models, additional acoustic training, and read speech data from a different corpus. Next, we try to address, more specifically, the conversational nature of the speech. We examine the problem of frequently occurring short words and the impact of the speaking rate on the performance.

4.1. Improving the Acoustic Model

We describe in this section how we are able to reduce the word error rate by using more complex acoustic models and increasing the amount of acoustic training. For all these experiments, we used the 70-topic language model. As a reference, the original large vocabulary speech recognition system [2, 3] used 3-state, gender-dependent models for each context-dependent phoneme, and the input feature vector consisted of cepstra and their first derivatives. The word error rate for this system was in the vicinity of 70%.

Using our baseline system described above, with 4.5 hours of acoustic training, first (D) and second difference (DD) cepstra, and 5-state HMM models, we achieve a word error rate of 67.5%. As we include cross-word triphone models and a trigram language model, the word error rate is further reduced to 61.9%. This indicates that word accuracy improves consistently as the complexity of the models increases. We then increase the amount of acoustic training to include, first, 32.2 hours and then the full 143.2 hours. At this point, the word error rate reaches 53.1%. In Figure 1, we show the change in word error rate as the amount of acoustic training data is increased. We notice that below 30 hours, the decrease in word error rate is around 2% for each doubling of the data. Above 30 hours, the decrease levels off slightly to only 1% for each doubling of the data. Finally, our best run (50.5%) is achieved by reducing the pruning levels and using improved crossword models. All these results are summarized in Table 5 and Figure 1.

systems	acoustic training (hours)	word error rate (%)
5-state, D+DD, 2-gram	4.5	67.5
5-state, D+DD, xword, 3-gram	4.5	61.9
5-state, D+DD, xword, 3-gram	32.2	55.1
5-state, D+DD, xword, 3-gram	143.2	53.1
5-state, D+DD, xword, 3-gram, imp.	143.2	50.5

Table 5. Word error rate and amount of acoustic training used for systems with different model complexity.

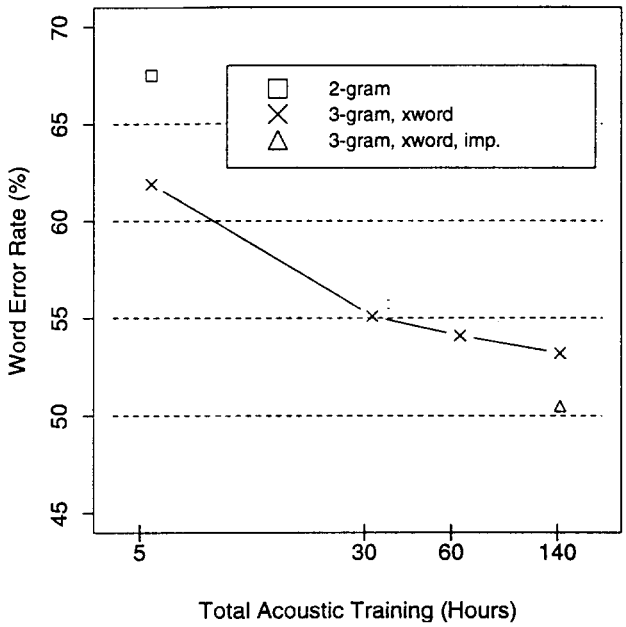

Figure 1. Word Error Rate as a function of the amount of acoustic training (hours) for 3 systems: 2-gram, non crossword models (□); 3-gram, crossword models (×); 3-gram, crossword models with improved search and modeling (△).

4.1.1. *Using Read Data*

One major concern was the possibility that the Switchboard acoustic data was not suitable for training due to its poor quality (see Section 2.). We believed that this was particularly an issue since we train our phoneme models from flat initial estimates. This motivated us to examine the use of an alternate corpus to train the acoustic models. We chose Macrophone [5] which is a large corpus of read and prompted speech recorded over the telephone. Each speaker was asked to record sentences from 15 different types. For our experiment, we used sentences from the TIMIT, WSJ and ATIS types, which amounted to a total of 27 hours of speech. For the first experiment, we trained the acoustic models using only the 27 hours of Macrophone data. The word error rate measured on the standard test set, with the standard language model, was 66.7%. This is significantly worse than 61.9%, the word error rate achieved with only 4.5 hours of Switchboard training. Further analysis shows that the triphone coverage on the test set using the Macrophone training is actually worse that the triphone coverage using the 4.5 hours of Switchboard. This is not surprising given that the Macrophone data comes from an entirely different domain than the test data. Compounded to this is the fact that the speaking styles are different. We hoped that these deficiencies would be compensated by the quality of the data. This, however, does not seem to be the case. This observation is confirmed in our second experiment where the Macrophone training was used only during the initial training phase, to generate the context-independent units from the flat estimates. In this case, the context-dependent units were trained using the 4.5 hours from Switchboard. The word error rate was 62.2%, almost unchanged from the error rate achieved using only the Switchboard data. These results are summarized in Table 6.

description	init	training (hours)	word error rate (%)
SWB	SW	4.5	61.9
MAC	MAC	27	66.7
MAC,SWB	MAC	4.5	62.2

Table 6. Word error rate using only Switchboard training, only Macrophone training, and Switchboard training bootstrapped from Macrophone.

4.2. Modeling Conversational Speech

Even though we were able to significantly reduce the word error rate with additional training and more complex models, the performance achieved on the Switchboard test set is still out of line with standard results on read speech. This difference cannot be completely explained by the high perplexity of the task or the quality of the data (recordings over the phone); we still need to model the conversational nature of the data better. Based on the analysis presented in [1], we identified two major areas to address. The first is the frequent use of short non-content words that are by nature more difficult to recognize, and the second is the speaking rate.

4.2.1. *Short Words*

We observed in Section 2 that one major difference between conversational (Switchboard) and read (WSJ) speech is the average length of the words. When we try to characterize the complexity of a task, we usually report perplexity, which is measured at the word level. The perplexity P is measured as:

$$P = 2^{-\frac{1}{N}\sum \log_2(p)}$$

where N is the total number of words in the test set, and p is the probability, estimated from training, of the observed transition. The assumption is that the average word length is comparable across the different domains. However, a more meaningful quantity to measure is a phoneme-based perplexity. This quantity can be expressed as:

$$P_p = 2^{-\frac{1}{N\mu}\sum \log_2(p)} = P^{\frac{1}{\mu}}$$

where μ is the average word length, measured in phonemes, in the test set. Observe that this quantity is normalized by the average word length and is thus more relevant in reflecting the actual difficulty of the task since our basic units are phonemes. On Switchboard and WSJ, the phoneme-level perplexities are 5.0 and 2.9 respectively. We can translate these differences into a word-level perplexity: a phoneme-level perplexity of 5.0 on WSJ translates into a word perplexity of around $5.0^{4.1} = 730$, which is very high. Fortunately, the concept of a word is arbitrary and we may want to consider common strings of words such as "YOU KNOW" and "A LOT OF" as a single word. We have implemented a scheme where pairs of short words which occur frequently define a single word that will be added to the original lexicon. In addition to reducing the effective perplexity of the task, this approach has other potentially beneficial effects in that it allows for variable n-grams around these words; it also allows for better acoustic modeling by artificially allowing crossword triphones to be included in the passes prior to the lattice pass.

In our experiment, we only considered the merging of pairs of words that have four phonemes or less, and we included in our lexicon only pairs that occur more than 100 times. This resulted in about 4000 new words. Using this augmented lexicon, a new

n-gram grammar was built and we measured a word error rate of 48.7% on the standard test set. This result is summarized in Table 7.

description	normal words	compound words	word error rate (%)
Baseline	5k	0	50.5
Compound-word	5k	4k	48.7

Table 7. Vocabulary size and word error rate for systems with and without compound words.

4.2.2. *Speaking Rate*

We present here the analysis of our results according to speaking rate (SR). At this point, we limit ourselves to "number of vowels per second" as the measure for speaking rate and we assume that this quantity is more or less constant over a sentence. We measured the speaking rate for each sentence in the test set and we partitioned the data into three groups of equal size (slow, medium, and fast) based on the SR value. We present in Table 8 the thresholds, word error rate, together with other statistics, for the full test set and the three groups. The medium group has, by far, the lowest error rate: 46.2%; the slow group does worse: 48.2% errors; and the fast group has the highest word error rate: 55.9%. The error rate discrepancy between the medium group and that of the two extreme sets shows that speaking rate is a significant performance indicator. Considering the table in more detail, we see two trends. Going from slow to fast, the number of fillers per sentence decreases while the number of words per sentence increases. The slow data have the most filler words and also the highest perplexity. In addition, the sentences are shorter on average, than those classified as medium or fast, which could result from segmenting sentences at long pauses. These observations are indicative of what may be a global change in grammatical structure and word usage resulting from speakers' indecision or hesitation. Further refinements in the language model may be required to adequately capture these effects. The fast data set has few fillers and higher perplexity compared to the medium data set. One hypothesis for the significantly higher error rate on the fast data is that there may be stronger coarticulation effects.

	SR (vow/s)	3-gram perp.	#fill	err	#sents	#wrds
base.	N/A	122	200	50.5	524	6724
slow	≤ 4	138	79	48.2	175	1683
med.	4 to 5	110	76	46.2	175	2370
fast	≥ 5	123	45	55.9	174	2671

Table 8. Speaking rate (vowels/second), 3-gram perplexity, number of filler words, word error rate, number of sentences, and number of words as a function of speaking rate group.

So far, we have mentioned effects that are sentence or even conversation based. But, it is possible that local effects, that is those occurring at the level of a few words, are important as well. This is especially true with the fast speech. Duration modeling might be one way to capture the local variation. However, we believe this must be done at the word level rather than at the phoneme level as preliminary results indicate that there are commonly occurring words with a large durational variance.

5. DISCUSSION OF THE RESULTS

In trying to reduce the word error rate on conversational speech, we presented a range of experimental results using the Switchboard corpus and the Macrophone corpus. We showed that significant gains could be achieved by applying standard techniques that work effectively on read speech such as WSJ. By combining more complex models (trigram, crossword acoustic models) with increased acoustic (143.2 hours of data) and language modeling training, the word error rate was reduced from the 70% to the 50%. An additional gain was obtained by reducing the effect of the short words by using a word pairing approach. It is important to note that the overall reduction in word error was not dominated by one or two single factors but was the result of a combination of all the various changes made to the original system.

Although the overall improvement is significant, the final error rate is still very high. We believe a great deal still needs to be done to address the real nature of conversational data. We plan to investigate acoustic models that can handle the large variability due to the difference in speaking rates and the coarticulation of words. At the same time, we are considering ways of improving the language modeling to deal with pause fillers, and repeats and restarts.

ACKNOWLEDGMENT

The authors would like to thank Richard Schwartz and Francis Kubala for their useful comments and insights.

REFERENCES

[1] E. Eide, "Understanding and Improving Speech Recognition Performance through the Use of Diagnostic Tools." *Proc. IEEE ICASSP*, 1995

[2] P. Jeanrenaud, M. Siu, K. Ng, R. Rohlicek, H. Gish, "Phonetic-based Word Spotter: Various Configurations and Application to Event Spotting." *Proc. ESCA Eurospeech*, 1993, vol. II, pp. 1057-1060

[3] J. McDonough, K. Ng, P. Jeanrenaud, M. Siu, H. Gish, R. Rohlicek, "Approaches to Topic Identification on the Switchboard Corpus." *Proc. IEEE ICASSP*, 1994, vol. I, pp. 385-388.

[4] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research Development" *Proc. IEEE ICASSP*, 1992, pp. I-517-520

[5] J. Bernstein, K. Taussig, and J. Godfrey, "Macrophone: an American English Telephone Speech Corpus for the Polyphone Project" *Proc. IEEE ICASSP*, 1994, pp. I-81-84

[6] F. Kubala, A. Anastasakos, J. Makhoul, L. Nguyen, R. Schwartz, G. Zavaliagkos, "Comparative Experiments on Large Vocabulary Speech Recognition." *Proc. IEEE ICASSP*, 1994, vol. I, pp. 561-564.