Spontaneous Speech Effects In Large Vocabulary Speech Recognition Applications

John Butzberger, Hy Murveit, Elizabeth Shriberg, Patti Price

SRI International
Speech Research and Technology Program
Menlo Park, CA 94025

ABSTRACT

We describe three analyses on the effects of spontaneous speech on continuous speech recognition performance. We have found that: (1) spontaneous speech effects significantly degrade recognition performance, (2) *fluent* spontaneous speech yields word accuracies equivalent to read speech, and (3) using spontaneous speech training data can significantly improve performance for recognizing spontaneous speech. We conclude that word accuracy can be improved by explicitly modeling spontaneous effects in the recognizer, and by using as much spontaneous speech training data as possible. Inclusion of read speech training data, even within the task domain, does not significantly improve performance.

1. INTRODUCTION

Recognition of spontaneous speech is an important feature of database-query spoken-language systems (SLS). However, most speech recognition research has focussed on acoustic and language modeling developed for recognition of read speech [1]. Read speech has been used extensively in the past for both training and testing speech recognition systems because it is significantly less expensive to collect than spontaneous speech, and because the lexical and syntactic content of the data can be controlled.

The multi-site data collection effort [3] has provided a challenging corpus for research and development in the Airline Travel Information System (ATIS) domain. We have observed a significant increase in word error rate compared to the previous task domain, the read-speech naval Resource Management (RM) task [2,6]. Word error rates for RM systems have typically been in the 5% range, whereas ATIS word error rates have exceeded 10% [4], for comparable perplexities.

The speaking style typically exhibited in the RM domain had a very consistent rate and articulation, within and across sentences, and across speakers. There were no disfluencies, such as word fragments, hesitations, or self-edits, since utterances containing these effects were removed from the corpus. The utterances tended to be short and direct (3.3 seconds long, on average). No pause fillers (uh, um), false starts, repairs, or excessively long pauses occurred. The speakers were able to concentrate on speech production, rather than query formation or problem solving. Furthermore, the training and testing texts were generated using a fixed vocabulary, and with the same, known language model, which quite adequately represented the source and target languages.

The speaking style typically exhibited in the ATIS domain differs from that in the RM domain all of the above aspects. The speaking rate is highly inconsistent, both within utterances, across utterances within a session, and across sessions and speakers. The articulation is highly variable, with stressed forms of function words and reduced forms of content words typically not observed in read speech. The sentence lengths vary widely, and are typically longer than RM sentences (7.5 seconds long, on average). Some words in ATIS sentences may not exist in the recognizer's lexicon, and an appropriate language model must be developed.

Most importantly, however, ATIS speech contains spontaneous effects and disfluencies: filled pauses, stressed or lengthened function words, false-starts and self-edits, word fragments, breaths, long pauses, and extraneous noises such as paper rustling and beeps. Data collected using systems containing automatic speech recognition and natural language components contain frequent occurrences of hyperarticulated words, elicited by the subjects in an attempt to overcome recognition or understanding errors [5]. Additionally, the data have been collected in normal office conditions (rather than in a soundproof booth), and recording quality and conditions vary across sites [3].

2. ERROR ANALYSIS

We begin by analyzing the errors that occurred in the February 1991 evaluation set of 148 Class-A sentences, for which our recognition word error rate exceeded 18%. These sentences are examined because they are believed to be a particularly difficult sampling of ATIS speech.

Phonetic alignments were automatically generated corresponding to both the reference and recognized word strings, and we listened to each utterance was listened to very carefully. The acoustic and language model scores were compared, and a subjective judgment was made as to the likely source of the error (the acoustic model, the language model, the articulation quality of the segment, or other effects such as breaths, out-of-vocabulary words, or extraneous noise).

We found that 30% of the errors (Table 1) could be attributed to poor articulation or poorly modeled articulation (usually reductions, emphatic stress, or speaking rate variations), 20% were due to out-of-vocabulary words or poor bigram probabilities, 20% were due to unmodeled pause-fillers (uh, um, breaths), and the remaining portion unexplainable, but probably due to inadequate acoustic-phonetic modeling.

We see that 70% of the errors are due to effects observed in the ATIS domain, but not in the RM domain. If these errors were removed, we would approach an error rate typically seen in a comparable RM system (with a perplexity 60 wordpair grammar).

Corpus	Cause for Error	Portion
ATIS only	Poor Articulation	30%
	Vocabulary and Grammar	20%
	Pause Fillers	20%
ATIS and RM	Other	30%

Table 1: Summary of error sources for the Class-A Feb91 ATIS evaluation set (148 sentences).

3. READ VS. SPONTANEOUS SPEECH

To determine the impact of spontaneous versus read speaking styles on recognition performance given a fixed training condition, a recognition experiment with two test sets was constructed. The first set contained spontaneous speech utterances; the second set contained read versions of those same utterances, given later by the same subjects.

The training data consisted of RM, TIMIT, and pilot-corpus ATIS utterances (with the read-spontaneous and spontaneous test data held out). This left rather little ATIS-specific data for training, almost none of it spontaneous. The recognition was run without a grammar (perplexity 1025) to remove any corrective effects of the grammar, so that only the acoustic effect of the spontaneous speech could be evaluated. The spontaneous test sentences were categorized as either fluent or disfluent based on the existence of special markings in their corresponding SRO* files.

We found that the primary difference in error rates between the read and spontaneous test sets was due directly to disfluencies in the spontaneous speech (Table 2). *Non-disfluent spontaneous speech had the same error rate as read speech.* The disfluencies include pause-fillers, word fragments, overly lengthened or overly stressed function words, self-edits, mispronunciations, and overly long pauses. This list of disfluency types is derived from the special markings used in the SRO transcriptions. The observation that non-disfluent spontaneous speech error rate approaches read speech error rate is consistent with the fact that the test speech much more closely resembles the training data. The training data was fluently and consistently articulated, just as was the non-disfluent spontaneous speech.

Characteristic	Num Sents	Word Error
Read	241	33%
Spontaneous	241	43%
Spontaneous - Disfluent	97	56%
Spontaneous - Fluent	144	32%

Table 2: Error rate versus speaking style. Read speech and fluent spontaneous speech have equivalent error rates.

The breakdown of error rate versus disfluency type (Table 3) shows that a significant portion of the errors were due to filled pauses, long pauses, lengthenings, and stress. Sentences with these disfluencies had twice the word error rate of fluent speech. The filled pause errors happened because there were no models for breath/uh/um events in this particular recognizer's lexicon. The stress and lengthening errors happened (most likely) because of the lack of sufficient observations of these events in the training data, and because of the lack of explicit models for these effects. The long pauses usually caused insertions within the pause regions neighboring the phrase-initial and phrase-final words.

From these observations, we conclude that more training data containing these effects would improve the match between the acoustic models and the spontaneous test speech, and therefore would improve the recognition performance. Furthermore, these effects should be explicitly modeled in the recognizer's lexicon, once sufficient training data is obtained. However, this process depends on the reliability of the SRO labeling across sites, which tends to be subjective and inconsistent.

*The SRO transcription contains a detailed description of all the acoustic events occurring in a utterance.

Disfluency Type	Num Sents	Disfluency Causes Error
Self-Edit	7	71%
Filled Pause	24	92%
Long Pause	17	94%
Lengthening	36	81%
Stress	22	59%
Mispronunciation	2	100%
Fragment	5	100%

Table 3: Number of sentences afflicted with each disfluency type, and the percentage of occurrences where the disfluency causes an error.

4. TRAINING DATA VARIATIONS

Further evidence for the importance of modeling spontaneous phenomena is found by manipulating the content of the training data sets that are used for acoustic-phonetic modeling. In this experiment, we compare spontaneous speech recognition performance given different combinations of read, spontaneous, ATIS, and non-ATIS training subsets.

The training subsets (Table 4) consist of the standard RM and TIMIT training data, and read and spontaneous subdivisions of all the ATIS and MADCOW data available as of October 1, 1991. The "Breaths" corpus refers to an internally collected database of inhalations and exhalations, used to train a breath model, which is allowed to occur optionally between words during recognition. Much of the ATIS-read data was also collected internally at SRI.

Corpus	Size
ATIS-Read	7,932
ATIS-Spontaneous	6,745
TIMIT	4,200
Resource Management	3,990
Breaths	800

Table 4: Training data subsets, which are combined in various ways to determine the impact of read and spontaneous training data on recognition of spontaneous speech.

Recognition was conducted using a development test-set of 447 spontaneous MADCOW utterances [3], with a perplexity 20 bigram grammar trained on all the available spontaneous speech transcriptions (roughly 10,000 sentences). All of the experiments outlined below use discrete-distribution HMMs, and every training set combination includes the 800 breath utterances.

Using all the available ATIS and MADCOW data yielded a system with a word error rate of 9.6% (Table 5). Using only spontaneous ATIS speech reduced performance by only 6%, to 10.2% word error. Training with a roughly equivalent quantity of read ATIS speech increased the error rate significantly, by 58% to 15.2%. This suggests that having training data which is consistent in speaking mode with the test data can significantly improve performance. However, the effect of lexical and phonetic coverage in the training sets might be a factor in causing this performance difference. This issue is discussed in Section 5.

Training Set	Size	Error
ATIS-Read	8,732	15.2%
ATIS-Spontaneous	7,545	10.2%
ATIS-All	15,477	9.6%

Table 5: Training set variations for ATIS-only systems. This table indicates that having speaking-mode-consistent data is a major contributor to performance improvement.

We also look at the impact of using non-ATIS read speech for additional training data (Table 6). Using successively more training data gives the expected result, an improvement in performance. However, when using all the available data (RM, TIMIT, ATIS and MADCOW), the performance matches that of the system trained exclusively on ATIS and MADCOW data. Furthermore, the performance of the system trained using all the available read speech (16,922 sentences) performed much worse than the system trained only on spontaneous speech (7,545 sentences).

Training Set	Size	Error
TIMIT	5,000	26.9%
TIMIT + RM	8,990	20.5%
TIMIT + RM + ATIS-Read	16,922	14.6%
TIMIT + RM + ATIS-All	23,667	9.6%

Table 6: Training set variations using non-ATIS data. The error rates is reduced when ATIS-read data is added, and is reduced further when ATIS-spontaneous data is added.

We can conclude from these experiments that having speaking-mode-consistent training data is more important than simply having a large quantity of training data. However, we cannot be certain that the phonetic content of the ATIS-spontaneous training set better matches the development set than the ATIS-read training set. This issue is addressed in the next section.

We compared the errors of two different recognizers used on the same test set of spontaneous speech. Both recognizers were trained on a comparable number of utterances, but one recognizer was trained on read speech only (TIMIT+RM+ATIS-Read), and the other on spontaneous speech only (ATIS-Spontaneous). We found that substitutions of one function word for another form a significant portion of the errors in both test sets, and in roughly the same proportions. However, there were significantly fewer substitutions of content words for other content words for the recognizer trained on spontaneous speech compared to the recognizer trained on read speech.

Similarly, the recognizer trained on spontaneous speech manifested significantly fewer errors in substitution of a pause filler for a function word. "Homophone" errors, which can lead to understanding errors, formed a significant portion of the errors in the recognizer trained on read speech, although almost none of these appeared for the recognizer trained on spontaneous speech. We believe that this is because many words that can be homophonous in read speech ("for"-"four" and "to"-"two", for example) are no longer homophones in spontaneous speech ("fer"-"four" and "tuh"-"two").

5. Phonetic Coverage Analysis

One potential reason for the dramatic performance variations could be that the phonetic content of the development test-set is better covered by the ATIS-Spontaneous subset than the ATIS-Read subset. In this section, we attempt to disprove that theory, giving further strength to the argument that speaking-mode consistency is the primary factor affecting performance.

We reason that the more detailed (more context-dependent) acoustic-phonetic models there are available for testing, the more adequate the training data has been in representing this dimension (the better the phonetic coverage). Therefore, for this analysis, we determine the average context level (or detail) of HMM states that each frame of test data visits during recognition. This is computed by assigning an integer-valued number to each model type (increasing as context level increases), then computing the percentage of all frames of data visiting states corresponding to a particular level of context.

The series of context-dependent model types used in the DECIPHER system is listed in Table 7. A model with a particular context level will be generated by the DECIPHER trainer if there is sufficient data to train that model.

Model Type	Context Level
Monophone	1
Left-general biphone	2
Right-general biphone	2
Left biphone	3
Right biphone	3
General triphone	4
Left-general triphone	5
Right-general triphone	5
Triphone	6
Word-specific	7

Table 7: Assignments of an integer-valued context level to each context-dependent model type. Models with increasing detail are assigned higher context level values.

The expectation is that the higher the average context level encountered during recognition, the better the performance. This trend is indeed captured in Table 8, where the system with the least task-specific training data (TIMIT) had the least average context level (and the lowest performance), and the system with the most training data (TIMIT+RM+ATIS-All) had the highest average context level (and the highest performance).

The important point to note is that the average context level of the best-trained read speech system (TIMIT+RM+ATIS-Read) was roughly equal to that of the best spontaneous-only system (ATIS-Spontaneous), but the error rate was significantly higher (14.6% versus 10.2%, respectively). This suggests that although models of equivalent detail are being used for recognition, the performance difference is due to the spontaneous speaking-mode of the training set, which is consistent with the speaking-mode of the test set.

Training Sets	Error Rate	Context Level
TIMIT+RM+ATIS-All	9.6%	6.31
ATIS-All	9.6%	6.26
ATIS-Spontaneous	10.2%	6.03
TIMIT+RM+ATIS-Read	14.6%	6.14
ATIS-Read	15.2%	5.96
TIMIT+RM	20.5%	5.06
TIMIT	26.9%	4.56

Table 8: Context level versus word error. This table indicates that despite similar model detail (context level), the spontaneous-trained system significantly outperforms the best read-trained system.

6. CONCLUSION

These studies have convinced us of the importance of using as much spontaneous speech material as possible in training our system. Furthermore, we have found that spontaneous speech effects can significantly degrade recognition performance, although *fluent* spontaneous speech yields word accuracies equivalent to read speech.

Word accuracy can be improved by using as much spontaneous speech training data as possible, and by explicitly modeling such effects in the recognizer's lexicon (such as optional interword breath and pause-filler models). Inclusion of read speech training data did not significantly improve performance, given that the phonetic coverage of the training sets were roughly equivalent.

Acknowledgments

We gratefully acknowledge support for this work from DARPA through the Office of Naval Research contract N00014-90-C-0085. The government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the government funding agencies.

REFERENCES

1. Price, P., W. Fisher, J. Bernstein, and D. Pallet, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *Proc. ICASSP*, 1988.

2. Pallet, D., J. Fiscus, and J. Garofolo, "DARPA Resource Management Benchmark Test Results June 1990", *Proc. DARPA Speech and Natural Language Workshop*, R. Stern (ed.), Morgan Kaufmann, 1990.

3. MADCOW, "Multi-Site Data Collection for a Spoken Language System," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

4. Murveit, H., J. Butzberger, and M. Weintraub, "Performance of SRI's DECIPHER Speech Recognition Systems on DARPA's ATIS Task," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

5. Shriberg, E., E. Wade, and P. Price, "Human-Machine Problem Solving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

6. Murveit, H., J. Butzberger, and M. Weintraub, "Speech Recognition in SRI's Resource Management and ATIS Systems," *Proc. DARPA Speech and Natural Language Workshop*, P. Price (ed.), Morgan Kaufmann, 1991.