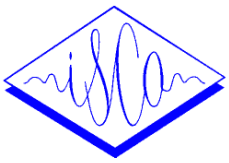

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Collection and Analyses of WSJ-CSR Corpus at MIT[1]

Michael Phillips, James Glass, Joseph Polifroni, and Victor Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

ABSTRACT

Recently, the DARPA community in the United States started a new data collection initiative in the Wall Street Journal (WSJ) domain to support research and development of very large vocabulary continuous speech recognition (CSR) systems. Since August 1991, our group has actively participated in the development of the WSJ-CSR corpus. The purpose of this paper is to document our involvement in this process, from recording and transcription to analyses and distribution. We will also present the results of an experiment investigating the preprocessing of the prompt text.

10.21437/ICSLP.1992-279

INTRODUCTION

One of the key ingredients that has contributed to the steady improvement in speech recognition technology in recent years is the availability of large speech corpora [1, 3, 7, 8]. With the help of these corpora, researchers have been able to develop recognition systems and obtain reliable estimates of system parameters. Perhaps just as important, these corpora, together with standardized performance evaluation procedures and metrics, have encouraged objective comparison of different systems, leading to better understanding and cross fertilization of research ideas [4].

The various speech corpora that the DARPA community has collected serve a wide range of purposes. The TIMIT corpus was designed with acoustic-phonetic research in mind. The Resource Management corpus addresses the needs for developing recognition systems with moderate vocabulary (1,000 words) and perplexity (60, with a word-pair language model). The VOYAGER and ATIS corpora contain spontaneously generated speech, and are useful for spoken language system development. All the presently available corpora have moderate vocabulary sizes and perplexities, and thus cannot adequately support research and development of very large vocabulary continuous speech recognition (CSR) systems in American English[2]. As a result, the DARPA community recently initiated an effort towards the construction of a new corpus to meet these needs.

The domain chosen by the community was the Wall Street Journal (WSJ), and the text prompts were selected from the CD-ROM distributed by ACL/DCI [5]. While the ultimate goal is to collect around 300 hours of speech from more than 100 speakers, it was thought that we should collect a pilot corpus of approximately 40 hours, to satisfy near term needs and to debug the text preprocessing and data collection processes. Since August 1991, our group has been one of three that has actively participated in the collection of the WSJ-CSR pilot corpus[3]. The purpose of this paper is to document our involvement in this process, present some analyses of the resulting data, and describe an experiment investigating the preparation of the prompt text.

DATA COLLECTION

All the MIT data were collected in an office environment, where the ambient noise level was approximately 50dB on the A scale of a sound-level meter. All utterances were recorded simultaneously using two microphones. A Sennheiser HMD-410 noise-cancelling microphone was always used for one of the channels. For the other channel, we rotated among the sessions three microphones: a Crown PCC-160 phase coherent cardioid desk-top microphone, a Crown PZM-6FS boundary desk-top microphone, and a Sony ECM-50PS electret condenser lavaliere microphone. The data were collected using a Sun SPARCstation-II, augmented with an Ariel DSP S-32C board and ProPort-656 audio interface unit for data capture. The sampling rate was 16 kHz, and the signal was lowpass filtered at 7.2 kHz. The input gain was held constant, for all subjects, at a setting that maximized the S/N ratio without clipping. Rather than transferring each collected sentence immediately to a remote file server for storage, and thus increasing the amount of delay between sentences, we stored the speech data temporarily on a 200 MByte local disk.

The prompt text, i.e., the text used to elicit speech material from the subjects, was preprocessed at Lincoln Lab to remove reading ambiguities inherent in written text [5]. Approximately half of the prompt text contained verbalized punctuation. The prompt text was displayed one paragraph at a time in the hope that this would encourage the subjects to produce sentence-level prosodic phenomena. The sentence to be recorded was highlighted in yellow, and the highlighting automatically moved forward to the next sentence once the previous sentence had been accepted. Four buttons (icons that can be activated with the mouse) were available for the subject to record, playback, accept, or unaccept an utterance. A push-and-hold mechanism was used for recording. We developed this user interface environment in

[1]This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

[2]A large corpus of spoken French has recently been collected by French researchers[2]. The BREF corpus contains over 200 hours of speech, collected from over 100 subjects.

[3]The other two participants are SRI and Texas Instruments.

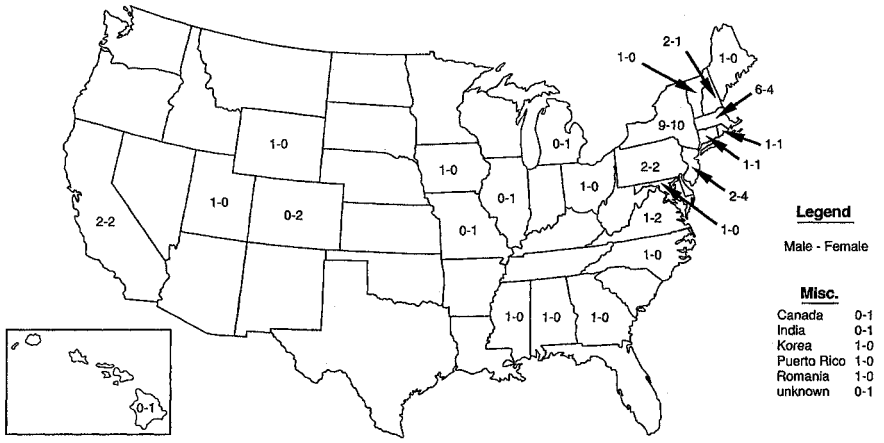

Figure 1: Geographical distributions of the subjects.

the hope that it would enable subjects to record the data with minimum supervision. Our experience with pilot data collection indicates that this is indeed the case. In fact, this software and hardware environment has also been adopted by one of the two remaining sites collecting WSJ-CSR data.

Subjects were recruited from the MIT community and vicinity via e-mail and posters. They were separated into three categories depending on how their data would be used for system development and evaluation: speaker-independent (SI), speaker-adaptive (SA), and speaker-dependent (SD). An attempt was made to balance the speakers by sex, dialect, and age, particularly for the latter two groups, since the total number of speakers in these groups is relatively small.

Data were collected in sessions of approximately 100 utterances (about 40 minutes per session). Each new subject was asked to read a set of instructions introducing them to the task. After that, the experimenter helped the subjects practice using the mouse for recording. The entire introduction took about 5 minutes. The subjects were then asked to read the designated set of 40 speaker adaptation sentences provided by Dragon Systems. The experimenter monitored the recording of the adaptation sentences, and asked the subject to repeat a sentence if a mistake was made. All subsequent recordings were made without supervision. Approximately half of the prompt texts for each subject contained verbalized punctuations. Subjects belonging to the SA and SD categories returned for multiple sessions. However, the introduction and the reading of the adaptation sentences took place only during the first session.

Once the data were recorded, they were authenticated using an interactive environment in which an experimenter could listen to an utterance, visually examine the waveform to detect truncation, and edit the orthographic transcription when necessary. Finally, the speech data and the corresponding orthographic transcriptions were written onto CD-ROM-compatible WORM disks for distribution.

We started the collection of WSJ-CSR data in early October, 1991, and completed the pilot collection by year end. Figure 1 shows the geographical distribution of all the subjects that we have recorded thus far. The age of the subjects ranges from 17 to 52, with an average of 27.1 and a standard deviation of 6.6. A breakdown of the amount of data collected in each of the three categories is shown in Table 1.

DATA ANALYSES

Since the WSJ-CSR speech corpus differs in many dimensions from the other corpora collected thus far in the DARPA community, we thought it would be useful to compute some of its vital statistics. In this section, we will describe some of the analyses that we have performed thus far.

All the analyses were based on the training set data, including the SI, SA, and SD categories[4]. The results are summarized in Table 2. In addition to computing various measures for the entire set, we have also analyzed the adaptation sentences, and those with and without verbalized punctuation.

Table 2 shows that the MIT training set contains nearly 15,000 sentences, and the number of sentences with and without verbalized punctuation are about equal. These sentences contain over 250,000 words, resulting in an average of about 17 words per sentence. The sentence length ranges from one word to 31 words with a standard deviation of 6.6 words. These sentences are considerably longer than any of the data that we have collected in other domains [1, 6, 8]. The adaptation sentences are generally shorter than the WSJ sentences although some speakers found them difficult to pronounce, and needed to be corrected repeatedly. On average, verbalizing the punctuations adds an extra 3 words to each sentence.

To compute the duration we first passed each sentence through an automatic begin-and-end detector to remove any extraneous silences. Altogether, the MIT training set contains almost 100,000 seconds of speech material, or about 27 hours. The average duration of the sentences is 6.5 seconds. The corresponding speaking rate is 156 words per minute, which is 30% higher than that for the spontaneous speech that we have collected [6]. This discrepancy is presumably due to the inherent difference in the way speech is elicited.

[4]We have excluded data from the development and test sets because of our desire to keep them uncontaminated for future system development and evaluation.

Category	Training Set		Development Set		Test Set	
	# sentences	# speaker	# sentences	# speaker	# sentences	# speaker
SI	6867 (6720)	49 (48)	747 (1600)	4 (8)	808 (1600)	4 (8)
SA	3206 (3840)	5 (6)	755 (960)	5 (6)	805 (960)	5 (6)
SD	4879 (4880)	2 (2)	295 (320)	2 (2)	324 (320)	2 (2)

Table 1: Statistics on the amount of data collected, expressed in terms of the number of sentences and the number of speakers, for each category and each data sets. The numbers in parentheses are the goals for the entire pilot effort.

Measurements	Adaptation	without VP	with VP	**Total**
# Sentences	2240	6410	6302	**14952**
# Words	29232	105533	120051	**254816**
Ave. # Words/Sentence	13.1	16.1	19.0	**17.0**
Duration (s)	11404	39053	47579	**98037**
Ave. Sentence Duration (s)	5.1	6.1	7.5	**6.5**
Ave. # Words/Min.	153.8	162.1	151.4	**155.9**
# Words Read with Errors	28	337	332	**697**

Table 2: Statistics of various measures for the adaptation sentences and sentences with and without verbalized punctuation.

In collecting the WSJ-CSR data, we hoped to provide an interface that was easy for the subjects to use, so that costly on-line monitoring was unnecessary. However, this potential cost reduction may be offset by the cost of authentication if the subjects produce too many errors. The sentences containing errors have the added disadvantage of not being well matched to the language model, which is constructed from the prompt text. To gain some insight into the magnitude of this problem, we tabulated the discrepancies between the final orthographic transcription and the corresponding prompt text. The result, summarized in the last row of Table 2, show that 697, or 0.27% of the words were read with error (including substitutions, insertions, and deletions). Note that, while the number of words read with errors for the adaptation sentences were one-tenth of that for the WSJ sentences, the *percentage* of errors for the adaptation sentences was only about one-third of that for the WSJ sentences. Recall that the adaptation sentences were read with an experimenter monitoring the process and instructing the subject to repeat when an error was detected. Thus, while monitoring the data collection process can reduce the errors by a factor of three, the magnitude of the problem is relatively small. Many of these errors are due to the speaker expanding abbreviations ("R. I." becomse "Rhode Island", for example). Since this would not occur in the verbalized punctuation text (the prompt would be "R .period I .period"), it is likely that these expanded abbreviations accounted for the slightly higher error rate in non-verbalized punctuation portions.

In the final analysis, the entire MIT training set, containing 27 hours of usable speech, was collected in approximately 125 40-minute sessions. Thus three hours of subject time is required to collect one hour of speech. Adding the overhead of recruiting and scheduling subjects, authentication, and other related administrative matters, we estimate that 6–8 hours of time is needed for one hour of speech.

TEXT PREPROCESSING EXPERIMENT

The prompt text used for the pilot collection was preprocessed by Lincoln Lab [5]. The rationale for this preprocessing step was at least two-fold. First, by converting numbers and abbreviations to a standard format, one removed any ambiguity concerning how these items should be read. Second, forcing the subjects to read the text in some pre-determined format would result in speech data that was consistent with the language model, which was derived from a considerably larger quantity of text data. However, some researchers felt that this preprocessing step might unnecessarily restrict the ways these items could be pronounced. Thus the collected data may not accurately reflect realistic situations in which a user is asked to dictate.

In order to gain some understanding of the effect of this preprocessing step, we recently conducted a small experiment. We first selected 100 sentences in the training set containing one or more items that were preprocessing candidates. Table 3 contains examples of selected sentences. These sentences were presented to the subjects, unprocessed, for recording. Following the recording, each utterance was carefully orthographically transcribed. This transcription was then compared with the processed prompt text used during the pilot data collection to determine if any discrepancies existed. We recruited 6 male and 6 female subjects, obtaining 12 readings for each of the 100 sentences. Half the subjects had served previously for the pilot collection effort.

Back then the distribution was $2.10 annually.
For the 1987 first 9 months, it had a $2.4 M net loss.
A W-4 form can be revived whenever necessary.

Table 3: Examples of candidate sentences for text preprocessing.

The results of the experiment can be analyzed in several ways. Figure 2 shows a histogram of the number of distinct renditions produced by the 12 subjects for the 100 sentences. There is considerable variation in the production of these sentences. The average number of distinct renditions is 3.9, with a standard deviation of 2.4. The figure shows that only 12 of the 100 sentences resulted in readings that agreed unanimously with the processed prompt text. Approximately half of the sentence tokens (601 out of 1,200) were identical to the corresponding prompt text.[5]

Figure 2 shows that, for almost 90% of the sentences used in this experiment, the subjects produced at least one rendition

[5]Although the data set size is small, we observed only small differences due to prior experience with the WSJ data collection. Experienced subjects agreed with the processed prompt text 315 times, whereas new subjects agreed only 286 times.

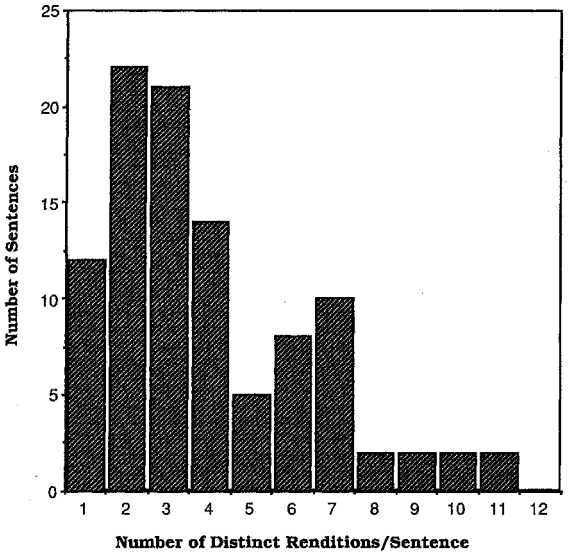

Figure 2: A histogram of the number of distinct renditions produced by the 12 subjects for the 100 sentences.

that differed in some way from the processed prompt text. But is this prompt text the preferred way of producing the sentences by our subjects? To answer this question, we computed the rank of the processed prompt text for each sentence which showed that the processed prompt text corresponds to (or is at least tied with) the most frequently produced rendition in over 60% of our sentences. Over 90% of the time, it was within the top three.

A closer examination of the 100 sentences showed that there were 171 locations where there was a discrepancy between the processed prompt text and at least one of the 12 recorded orthographies. 49 of these seemed to be reading errors and consisted of a single word deletion, insertion, or substitution, and were typically produced by only one of the 12 speakers. An additional 14 discrepancies were due the addition of verbalized sentence punctuation (the subjects were not asked to verbalize punctuation). Of the 1296 substrings associated with the remaining 108 locations, 635 (or 49%) of them matched the processed prompt text exactly. Almost all of the discrepancies (96.6%) can be classified into three categories: numbers, abbreviations, and dates. Discrepancies in 81 locations were due to the different ways numbers can be pronounced (e.g., "two hundred *and* thirty four" *vs.* "two hundred thirty four," "two point thirty four" *vs.* "two point three four"). Discrepancies in an additional 20 locations were due to the different ways abbreviations (e.g.,"Corp" or "E.S.T") can be said. Finally, there were 7 locations with discrepancies in the pronunciation of dates (e.g."ten" for "tenth").

The results of our investigation indicated to us that although there is a large variation in the way the subjects have spoken these unprocessed sentences, the types of variation is fairly limited. In addition, the magnitude of the these variations would be smaller in the overall corpus since we only presented unprocessed sentences that seemed to have ambiguous realizations. Nevertheless, we are still faced with the question of whether or not to preprocess the data. Before we can answer this question definitively, it is important that we conduct further study on a larger sample of sentences using a larger number of subjects. In the end, the decision of whether to preprocess the text will have to be determined by the community who will be the consumers of the resulting data, after considering the objectives of the research program and the trade-offs between a more reliable language model and more realistic speech data.

SUMMARY

This paper describes our involvement in the collection of the WSJ-CSR pilot corpus. By paying close attention to developing a computer interface that is easy to use, we were able to collect over 33 hours of speech from 64 subjects over a relatively short period. By using in-house equipment to produce CD-ROM-compatible WORM disks, we were able to distribute the data to interested researchers rapidly. Our analyses of the collected data show that the WSJ-CSR corpus differs significantly from other corpora in the research community. We expect that it will have long-lasting impacts on speech recognition research within the DARPA community and around the world.

Our preliminary text preprocessing experiment suggests that the current preprocessing scheme may not be adequate in capturing the ways people would naturally speak the sentences. Clearly, more extensive experiments must be performed. Whether one should preprocess the text at all is a decision that the DARPA community must decide collectively.

ACKNOWLEDGEMENTS

The collection of the WSJ-CSR data received help from many members of our group. In particular, Christie Clark Winterton was responsible for recruiting, scheduling, and assisting the subjects. She also authenticated a large fraction of the orthographic transcriptions of the collected data.

REFERENCES

[1] Lamel, L. F., R. H.Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop*, 100–109, February, 1986.

[2] Lamel, L.F., Gauvain, J.L., and Eskenazi, M., "BREF, a Large Vocabulary Spoken Corpus for French," *Proc. Eurospeech-91*, 505–508, September, 1991

[3] MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus," to appear in *Proc. DARPA Speech and Natural Language Workshop*, February, 1992.

[4] Pallett, D., "Benchmark Tests for DARPA Resource Management Database Performance Evaluations," *Proc. ICASSP-89*, 536–539, May, 1989.

[5] Paul, D. and Baker, J., "The Design for the Wall Street Journal-Based CSR Corpus," to appear in *Proc. DARPA Speech and Natural Language Workshop*, February, 1992.

[6] Polifroni, J. Seneff, S., and Zue, V., "Collection of Spontaneous Speech for the ATIS Domain and Comparative Analyses of Data Collected at MIT and TI," *Proc. DARPA Speech and Natural Language Workshop*, 360–365, February 1991.

[7] Price, P., Fisher, W., Bernstein, J., Pallett, D., "The DARPA 1000-Word Resource Management Database," *Proc. ICASSP-88*, 651–654, April, 1988.

[8] Zue, V., Daly, N., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., Seneff, S. and Soclof, M. "The Collection and Preliminary Analysis of a Spontaneous Speech Database," *Proc. DARPA Speech and Natural Language Workshop*, 126–134, October 1989.