SPONTANEOUS SPEECH COLLECTION FOR THE CSR CORPUS

Jared Bernstein and Denise Danielson

SRI International
Menlo Park, California 94025

1. ABSTRACT

As part of a pilot data collection for DARPA's Continuous Speech Recognition (CSR) speech corpus, SRI International experimented with the collection of spontaneous speech material. The bulk of the CSR pilot data was read versions of news articles from the *Wall Street Journal* (WSJ), and the spontaneous sentences were to be similar material, but spontaneously dictated. In the first pilot portion of the data collection, twelve subjects including nine journalists were located, and instructed in how to dictate using the data collection hardware and software at SRI. These talkers produced 1280 spontaneous sentences. In general, compared to read material, the spontaneous material took about two to three times more subject time to produce and about four times more experimenter time to produce, package, and ship. The paper provides details on the materials, subjects and procedures used in the study, and it describes the results in terms of speaker reaction and data production. The methods described are sufficient to collect fluent spontaneous recordings at a predictable rate. The spontaneous material differs in several characteristics from WSJ material; paragraphs and sentences tend to be longer, more word types are used, and by most measures, the material is more variable.

2. INTRODUCTION

The CSR (Continuous Speech Recognition) Corpus collection can be considered the successor to the Resource Management (RM) corpus[1], it focuses on the further development of speech recognition technology toward larger or open vocabularies, speaker and task independence, and is moving toward spontaneous speech. The default task in the pilot collection has been dictation of newspaper articles as if for the *Wall Street Journal* (WSJ). Thus, the largest part of the effort to collect a pilot version of the CSR corpus has been recording people reading selected short passages from the WSJ itself. The pilot CSR corpus was designed, however, such that a significant portion of the material was to be spontaneous and a subset of the speakers who read WSJ texts also were asked to dictate spontaneous articles in the WSJ style.

This paper describes the methods used in the collection of the spontaneous portion of the CSR corpus.

3. METHOD

User Interface. The speech data collection was performed with user interface software designed by Mike Phillips for collection of read speech [2]. MIT provided this software to SRI, where it was slightly modified for use in collecting spontaneous speech. The interface requires a *talk* button to be pushed and held down in order to record speech; another button must be explicitly pushed to accept the sentence. The most recent sentence is always available for playback.

Material Selection. Several issues seemed important in the selection of materials to be used in story generation.

1. It seemed appropriate to select material that would match the content of the WSJ, its vocabulary and topics.
2. To ensure that speech is truly spontaneous and not just read from source material, it is best to provide material that gives enough information, without giving it in a format that would encourage subjects just to read. Subjects need to come up with their own wording.
3. Subjects should be set up to maximize the likelihood of success. Any reasonable accommodation that produces appropriate spontaneous material is acceptable.

The materials provided to subjects changed over the course of the experiment. At first, subjects were provided with recent news articles or letters to the editor and were asked to prepare an outline of the material, put aside the original article, and then dictate from the outline or notes. In later sessions most subjects preferred to work from notes or press releases they had brought themselves and were, therefore, familiar with. Thus, subjects were encouraged to come to the session with topics and notes prepared.

Subjects. Twelve subjects generated spontaneous news stories. Four of the twelve generated two sets each of 80 spontaneous sentences. The other eight provided one set each of 80 sentences. The twelve included seven journalists

and three SRI employees. Three journalists were from the *Stanford Daily* and four from the *Peninsula Times-Tribune* (a local daily). The other two subjects were one journalist currently doing public relations work under contract and a former broadcast journalist.

Subject Recruiting and Selection. We chose to try to use journalists for this task. Not only did the particular task of news-style dictation lend itself to the use of journalists, but news writers seemed likely to be able to perform the task without undue effort.

Preference was given to individuals who had dictation experience. Given time constraints, we were unable to limit subjects to only those with dictation experience, and doing so might very well have also imposed an age constraint: most journalists who were in the field prior to the proliferation of PCs and word processors have dictated news stories; younger journalists have not.

We found subjects by first contacting a couple of local newspapers and speaking with the editor-in-chief or whoever they referred us to. After briefly describing the project and our needs, we asked for feedback about the level of interest that we could expect at what rate of pay. Journalists at major papers (where we would be more likely to find large numbers of speakers with dictation experience) typically wanted $35–$50/hr. At smaller papers we were able to find interest in the $20–$30/hr. range.

We were able to find enough people for the pilot study at a rate of $20/hr. Several of these speakers expressed an interest in coming back to do more dictation.

Potential subjects were first screened over the phone. After describing the project and the time commitment involved, we then asked the potential subject to "pick a topic of interest" — a story/column they are currently working on, or a current issue in the news — and dictate a brief story on that topic over the phone. We typically asked them to do this two or three times, to give us an idea of how easily they could come up with material.

Procedures. On arrival at the first recording session, the subject was asked to read a complete set of written instructions. The instructions are reproduced in the appendix. Next, subjects filled out a short information sheet about themselves, including a description of any prior dictation experience.

The data collection software was then demonstrated, and the subject was allowed to practice using the push-to-talk button. The practice session included 1–2 paragraphs of *Wall Street Journal* read text without any verbalized punctuation, and 1–2 paragraphs with verbalized punctuation. For 10 of the 12 subjects, this was the only exposure to the *Wall Street Journal* read text material prior to producing spontaneous data.

The first real data recorded from each subject consisted of 40 adaptation sentences that were read immediately following the practice session. Thus, by the time subjects were ready to start the spontaneous speech collection sessions, they were already fairly comfortable with the various controls available on the MIT collection software, including functions for reviewing, accepting, and rejecting utterances.

The remainder of the first recording session was devoted to spontaneous speech collection. Each set of 80 sentences comprised one ***session*** of 40 sentences with no verbalized punctuation (**NVP**) and one session of 40 sentences with verbalized punctuation (**VP**) [3]. All subjects generated the 40 sentences without punctuation first. The decision to order the collection this way was based on subject feedback regarding anticipated difficulty of the two tasks, and the experimenter's observations that subjects did in fact tend to have more difficulty with the verbalized punctuation condition.

Subjects were instructed to imagine that they were using a real speech-to-text dictation system to generate news-style articles as though intending to submit the articles for publication in a major newspaper. They were told that they could assume that the articles would be reviewed and edited before publication, so that they did not need to worry about making it perfect.

One goal of this project was to learn something about what people would expect to do naturally if they were using a speech-to-text dictation system. For this reason, the experimenter tried to control the process as little as possible. Subjects were allowed to use their own judgment as to whether or not a sentence was "good" and should be accepted.

The first two or three subjects were handed a variety of source materials and instructed to find topics of interest, jot down some notes, and then dictate from the notes. After some experience and feedback from these first few subjects, the experimenter began instructing subjects over the phone in advance to "come prepared." We briefly described the task as one in which the subject would be asked to make up and dictate short news-style articles. We asked that they have several topics in mind about which they could create brief stories.

Subjects were encouraged to use ideas from current stories they were working on and from recent articles they had done. It was suggested that they bring notes to work from as long as those notes were in "cryptic" form; they were specifically instructed not to bring in completed articles or notes in sentence form. Most subjects found this to be a much easier task than working from the SRI-supplied materials; however, the ability to control for content/vocabulary was lost. Most subjects thus did the majority of their "sto-

ries" from ideas they brought with them, and turned to the newspapers and other material provided by SRI only if they ran out of ideas before finishing the required 40 sentences.

Subjects returned for at least one additional recording session during which they completed reading the text portion of the collection, and also read back their own spontaneously produced sentences. The four subjects who generated a second set of 80 spontaneous sentences did so after having completed a significant amount of read text.

The schedule of a typical subject, then, was as follows:

Day 1:
- Read instructions
- Collection software demonstrated with NVP and VP WSJ material
- 40 adaptation sentences
- 40 spontaneous NVP
- 40 spontaneous VP

Day 2:
- Read WSJ
- Read spontaneous

4. RESULTS

The results of SRI's work with spontaneous speech are of three types: information about the cost in time or money to collect the material; information about the characteristics of the spontaneous material itself; and information about subject reactions to the procedure.

4.1. Production Cost

A principal concern about the collection of spontaneous speech is that the cost is high and variable. Because the pilot CSR data collection had pairs of collection sessions from the same speaker, one spontaneous session and another session during which a clean, written version of this spontaneous material was read, we have a good basis of comparison for the cost of spontaneous *vs.* read speech.

Figure 1 displays the distributions of recording session times in four collection conditions: spontaneous *vs.* read spontaneous, with verbalized punctuation (**VP**) *vs.* no verbalized punctuation (**NVP**). The recording session time is approximated by the difference in time between the completion of the first sentence and the last sentence in a 40-sentence session. This measure leaves out the variable preparation time that often occurs before the first spontaneous sentence in a session. This preparation time was typically about 5 minutes. There are 16 sessions in each of the 4 conditions.

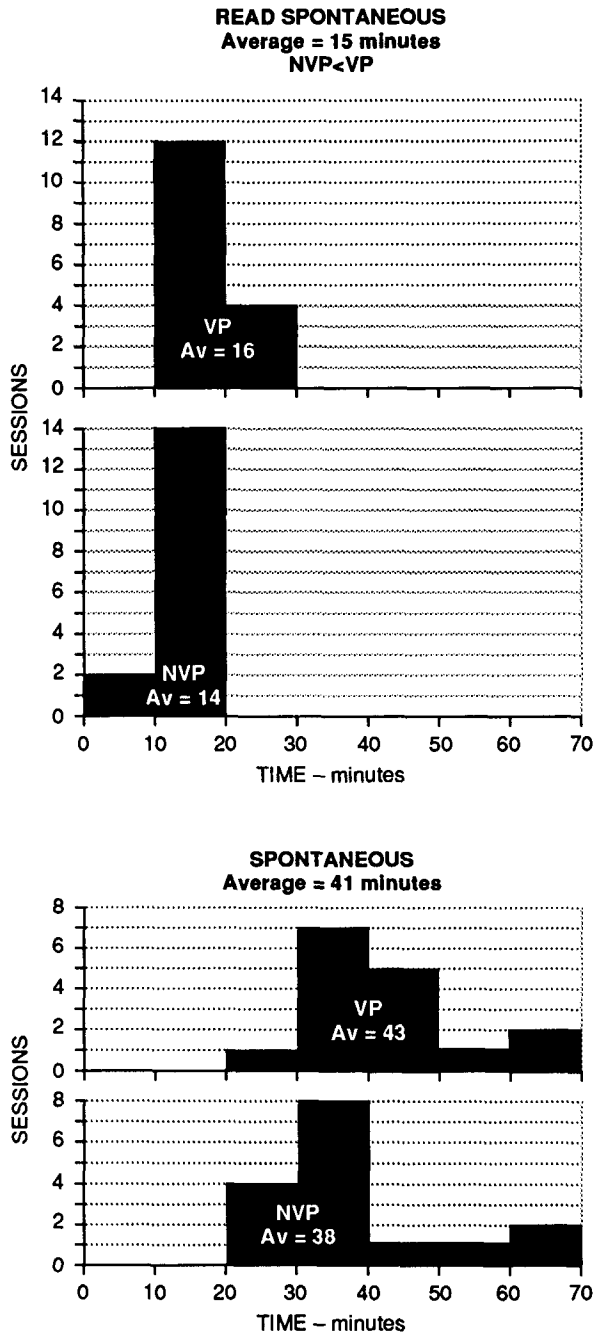

Figure 1. Elapsed time per recording session

Actual observed speaker time was about 5.5 hours for speaker-independent test conditions, which is slightly above our initial estimates. Data collection supervisor time for setup, microphone check, transcription, tape processing, scheduling, and other miscellaneous activities (including the elapsed data collection/speaker time) was approximately

12 hours for one such subject, as shown below.

Times for CSR Pilot Collection Tasks

Description	Minutes
Speaker time (5.5 hours):	
Initial instruction and practice	30
40 adaptation sentences	30
160 read sentences (5k and 20k)	100
80 read spontaneous sentences	50
80 spontaneous sentences	120
Monitor time (12 hours per subject):	
Collection time	330
Transcribe 240 read sentences to .dot	90
Transcribe 80 spontaneous sentences to .ptx .dot	160
Directory, shortpack, exabyte	90
Scheduling, miscellaneous	40

It may help to make a direct comparison of the time it takes to collect and transcribe (orthographically) a single set of 80 sentences, when one set is read and the other is spontaneous. The following table gives times required for collection itself (Subject & Experimenter), and the transcription times for generating prompt texts (.ptx) and detailed orthographic forms (.dot).

Collection and Transcription Time for One Set of 80 Sentences

Description	Read	Spontaneous
Collection (S&E)	100 min	240 min
xcribe to .ptx	—	120 min
xcribe to .dot	20 min	40 min
Total	2 hours	6.7 hours

There were several additional costs for the spontaneous speech. The speaker cost is higher because we needed to pay more to attract journalists. The spontaneous recordings also involve costs for preparing materials. These costs were minimal for this study, but for future efforts we expect to gather or create "fact sheets" and other prompt materials.

4.2. Characteristics

The material generated in the spontaneous sessions differed from real WSJ text. The differences occurred in several characteristics: content, vocabulary, paragraph size, speech rate. Furthermore, there were differences in both central tendency and in variability in most measures. The following table lists several obvious differences.

Description	WSJ-20	Spontaneous
Topics per session	11.2	6.5
Words	22,374	27,757
Sentences	1,293	1,280
Words/sentences	17.3	21.7
Unique words (types)	4,062	4,905
Average types/session	389	406
Range of types/session	337–426	281–513
Punctuation types	16	32

First, the number of different paragraphic topics that comprise a 40-sentence session was 11 in the WSJ material and about 6 or 7 in the spontaneous material. That is, spontaneous speakers like to keep going on a topic for six or seven sentences, whereas the WSJ cuts stories into paragraphs of about three or four sentences. Second, in a similar number of sentences, the spontaneous talkers used more words and more different words to construct longer sentences. Even at the session level, speakers used more different word types. Third, and most characteristically, the sessions were much more variable in the spontaneous condition. The WSJ materials are relatively uniform, and the spontaneous materials are more varied, both in the range of word types (shown in the table) and in sentence length and other measures not shown.

Figure 2, below, displays speech rate measured for materials recorded in four different conditions: read WSJ text *vs.* spontaneous text; and each with verbalized punctuation *vs.* with no verbalized punctuation. The speech rate for these materials is approximated by dividing the number of words in a sentence (including verbalized punctuation words) by the length of the file in time, without any endpointing or allowance for sentence internal silence. Thus, the measure is adequate for comparisons here, but cannot be taken as absolute or be compared to other figures.

As can be seen in Figure 2, the speech rates observed are slower for spontaneous speech than for read speech, and are slower for speech produced with verbalized punctuation in either form. Again, the spontaneous material is also more variable than the read material.

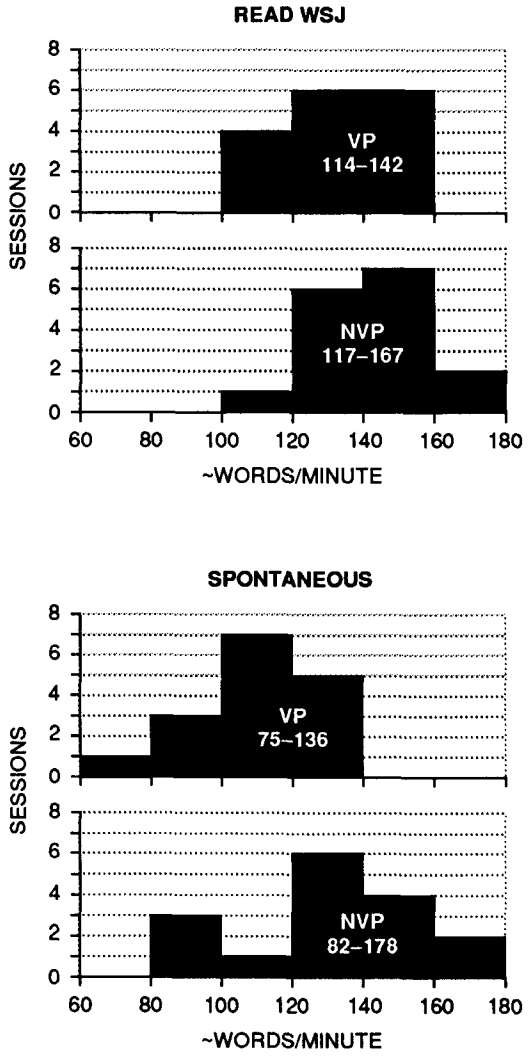

Figure 2. Speech rate in nominal words/minute

4.3. Subject Reaction

Materials. Subjects were most comfortable working from topics and materials that they brought with them. Most were not able, however, to be prepared on enough topics to do all of their collection this way. The next favored method was to use news releases or "fact sheets," as most journalists are accustomed to using these as sources.

An advantage of having subjects come with their own ideas and materials was that they tended to be more fluent and able to perform the task with greater ease when talking about topics with which they were familiar. In addition, they tended to produce longer and more complex sentences when covering familiar topics.

Verbalized Punctuation. The dictation with verbalized punctuation was perceived by subjects to be more difficult than dictation without punctuation. Subjects also reported that including all punctuation did not seem natural. Speakers did seem to become more comfortable with the task with practice, but their use of punctuation was inconsistent. Certain types of punctuation, such as quotation marks and commas used to offset items in a series, were seldom left out. Other punctuation marks, such as hyphens and dashes, were often omitted or used incorrectly.

The actual collected corpus does not really reflect the extent of these inconsistencies since either the speaker or the monitor would often catch the worst cases and the speaker would repeat the whole utterance.

Collection Paradigm. Speakers complained about having to speak one sentence at a time. They wanted to speak nonstop while the thoughts were there, and not have to wait for the machine.

Several subjects observed that a more natural way of doing dictation would be to speak in paragraphs, but with the capability to pause, i.e., stop recording while they think, then restart from the point where they left off. With this type of collection paradigm, some verbalized punctuation would be natural. In particular, it would seem natural to say "PERIOD" to indicate the end of one sentence before beginning the next.

5. CONCLUSIONS

Several results are evident:

1. The task can be done and with a fairly predictable rate of production. The total cost per sentence is about three or four times greater than similar read material.
2. The journalists do seem better at this task than others subjects of similar educational level.
3. Solicitations at local papers did not generate a large number of interested subjects.
4. Subjects with prior dictation experience do better at this task than those without such experience.
5. Subjects with more experience produced longer, more complex sentences.
6. Given the current editing tools, most subjects produce rather smooth and fluent spontaneous materials, primarily by rejecting whole utterances.
7. The spontaneous material is spoken slower and is generally much more variable than the read WSJ material.

Summary. Relatively fluent, spontaneously generated news stories can be collected at about four times the cost of read materials. Analysis of the materials is incomplete because

the collection was just finished and because the most important analysis will be done by the sites who use the data to run experiments.

Ongoing Research. SRI is currently collecting speech from an additional 8 test speakers. The current work includes experimentation with different ordering of collection sessions, and different materials and instructions for eliciting spontaneous speech. Because of the negative feedback regarding the verbalized punctuation condition, we are having some of the current subjects collect an extra set of 80 spontaneous sentences with no instructions regarding punctuation.

6. ACKNOWLEDGEMENT

SRI acknowledges support for this work from DARPA through NIST contract 50SBNB0C6211 D.O. 1040. The government has certain rights to this material. Opinions, findings, conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of any government agencies.

7. REFERENCES

1. Price, P., W. Fisher, J. Bernstein & D. Pallett 1988: "The DARPA 1000-word Resource Management Database for Speech Recognition," *IEEE Proc. ICASSP-88*, pp. 651-654.

2. Phillips, M., J. Glass, J. Polifroni, & V. Zue 1992: "Collection and Analyses of WSJ-CSR Data at MIT," in these Proceedings: *Fifth DARPA Workshop on Speech and Natural Language*, February 1992.

3. Paul, D., & J.Baker 1992: "The Design for the WSJ-CSR Corpus," in these Proceedings: *Fifth DARPA Workshop on Speech and Natural Language*, February 1992.

8. APPENDIX: SUBJECT INSTRUCTIONS

You will be asked to speak several different sets of sentences today. The first set consists of 40 individual sentences. These are all simple sentences which are fairly easy to read. Some of them do, however, use phrasing that may differ somewhat from the way you might normally say the same thing. It is important that you speak the sentences EXACTLY as they are written.

The remaining sentence sets fall into 3 different categories. The order in which these sets are collected may vary from one speaker to the next, but each speaker will contribute at least one set from each category. The 3 categories of sentence sets are described below:

READ TEXT: These sentences are taken from newspaper articles, and are presented 3 - 8 at a time in paragraph form. One sentence at time will be highlighted, indicating that the system is ready for you to read that sentence.

Type 1: **regular** — All punctuation marks are included in their normal fashion. The punctuation marks were left in to help you speak the sentences normally. Do NOT explicitly pronounce any of the punctuation marks.

Type 2: **verbalized punctuation** — For these sentences, each punctuation mark is written out and should be spoken as a regular word. For example ",COMMA" is pronounced "comma" and ".PERIOD" is pronounced "period."

SPONTANEOUS DICTATION: This will be sentences that you make up. You will be presented with a variety of material that can be used for ideas, and then asked to create news style "articles" as though planning to submit them to a newspaper for publication. For this task, we ask that you imagine that you are using a speech recognition, or "speech-to-text" system to create real articles. The "articles" that you create need not be complete — a total of 3 - 8 sentences per topic is fine — but they should consist of a group of sentences related to a single topic.

Type 1: **regular** — These sentences should NOT include spoken punctuation. Instead, we view these sentences as though the speech-to-text process is a "first-pass" effort, i.e., you can assume that appropriate punctuation would be inserted during later editing.

Type 2: **verbalized punctuation** — Again, these will be sentences that you make up, however in this set we ask that you explicitly say any punctuation marks that you would want to have appear in the article. Examples of some of the punctuation marks that we use (for the READ TEXT) are shown in the table below, but you may use whatever words you feel comfortable with.

Punctuation marks are spoken as:

, COMMA
. PERIOD
" DOUBLE-QUOTE
(LEFT-PAREN
) RIGHT-PAREN
- HYPHEN
-- DASH
: COLON
; SEMI-COLON

READ SPONTANEOUS:

For this set, you will be asked to read the sentences that you dictated previously. Both versions (with and without verbalized punctuation) will be presented, just like the standard READ TEXT.

NOTES: In the READ TEXT data there are a number of sentence fragments—the data was taken from a database of news articles and run through a screening process to eliminate most problem sentences, but in the attempt to automatically "clean up" the data other problems were sometimes created. A typical problem is that the algorithm used to break articles into distinct sentences was not terribly sophisticated. In some cases, this results in a false sentence break when there is an abbreviation followed by a word beginning with a capital letter.

Although they may sound strange, these sentences should be read as they are presented, with no attempt made to "correct" the mistakes. Another type of problem is that because the sentences were taken from a real newspaper, there are some strange words and uncommon proper names. Just do your best to come up with a reasonable pronunciation. Finally, all numbers have been written out to guide your pronunciation. The numbers and some of the sentences may be a little difficult to sight-read, especially at first. You may find that it helps to review the sentence silently before reading it.

Please try to read the sentences as accurately and naturally as possible. You may repeat a sentence as often as necessary before going on to the next sentence. From time to time, your monitor may also ask you to repeat a sentence, or to play it back so she/he can listen to it.

Have fun, and thank you for participating in this project!