CSR Corpus Development

George R. Doddington

SRI International
Menlo Park, CA

ABSTRACT

The CSR (Connected Speech Recognition) corpus represents a new DARPA speech recognition technology development initiative to advance the state of the art in CSR. This corpus essentially supersedes the now old Resource Management (RM) corpus that has fueled DARPA speech recognition technology development for the past 5 years. The new CSR corpus supports research on major new problems including unlimited vocabulary, natural grammar, and spontaneous speech. This paper presents an overview of the CSR corpus, reviews the definition and development of the "CSR pilot corpus", and examines the dynamic challenge of extending the CSR corpus to meet future needs.

OVERVIEW

Common speech corpus development and evaluation received major emphasis from the very beginning of the DARPA speech recognition program. At that time, a set of common corpora were defined to serve the needs of the research community. This resulted in the development of the TIMIT speech corpus, which was collected from a large number of subjects and intended to support basic research in acoustic-phonetic recognition technology. The Resource Management (RM) corpus, collected from fewer subjects but representing an application of interest to DARPA, provided the greatest focus of interest in technology throughout the research community. In the course of R&D using these two corpora, the first serious research and advances toward speaker-independent speech recognition were achieved.

Although the RM corpus served its intended purpose well, technology advances came to make its limitations painfully obvious. The language was artificial and limited, the speech was read and therefore unnatural, and the corpus completely avoided the central issue of understanding the meaning of the spoken utterances. In response to these limitations and to rapid advances in the performance of speech recognition technology on this RM task, a new research initiative was formed by combining speech recognition and natural language understanding tasks in a spoken language system (SLS) program.

The SLS program took shape with the definition of the Airline Travel Information System (ATIS) task, a database query task which supports research in both speech recognition and natural language. The ATIS corpus (corpora) is currently being collected to provide the experimental data for developing SLS technology. This ATIS corpus exhibits several desirable features regarding the speech recognition problem that were found lacking in the RM corpus. These features are namely the use of spontaneous goal-directed speech and the consequent use of a natural grammar and an open unrestricted vocabulary.

Although the ATIS corpus provides the kind of speech data desired by the speech recognition research community and required to address important problems in the application of speech recognition to real tasks, there is one unfortunate shortcoming of this corpus. This is that the cost and effort of collecting the data is too great to support the massive data requirements for advances in speech recognition technology. Some way of improving the efficiency and productivity of data collection was needed in order to support further advances in speech recognition technology. This need was the primary motivation for the creation of the CSR research initiative and its related CSR corpus.

The CSR research initiative, along with the CSR corpus development effort, was created in order to provide better support for advances in the state of the art in large vocabulary CSR. The primary focus in the CSR initiative has been on the design and development of a CSR speech corpus which is required to fuel the research and through which the research might be productively directed. Primary objectives of the CSR corpus have been to increase the realism of the speech data and at the same time to maximize the efficiency of collecting that data. Efficiency has been viewed as of paramount importance because it is generally believed that significant advances in speech recognition technology will require more comprehensive models of speech and correspondingly more massive quantities of speech data with which to train them.

Janet Baker was the principal champion and designer of the CSR corpus, working as the chair of a CSR corpus design committee. This committee dealt with a large and diverse set of research interests and corpus needs, which made the

task of designing a satisfactory corpus extremely difficult. For example, the desire to collect spontaneous speech was in direct opposition to the need to make corpus development efficient (because spontaneous speech requires a generally painstaking and expensive transcription task, whereas read speech can be transcribed far more efficiently and even largely automatically). [1]

Major Corpus Design Decisions

- Read speech versus spontaneous speech: On the issue of spontaneous speech, it was decided that the majority of the corpus (and in particular the majority of the training data) should be read speech, for economic reasons, whereas the majority of the test data (which comprises a small fraction of the total data) should be spontaneous speech. The reason for these decisions is that it was felt that large amounts of read speech would provide greater training benefits than smaller amounts of spontaneous speech, while using spontaneous speech for testing would better validate the technology for a relatively small increase in cost.

- Prompting text: Probably the most significant decision regarding the CSR corpus was the decision to work initially with the Wall Street Journal (WSJ). This decision was influenced by the richness of the WSJ language and by the existence of a preexisting and very large (50 million word) corpus of WSJ text (as part of the ACL-DCI effort). All of the read speech data is currently being collected using prompts derived from the WSJ. The spontaneous speech data is being collected using a news reporting dictation paradigm that simulates the WSJ dictation scenario.[2]

- Verbalized punctuation: In dictation, which is the nominal target application for the CSR technology development effort, dictation users typically say punctuation such as "comma" and "period" so as to aid in the proper punctuation of the dictated document. Therefore, in order to improve the verisimilitude of the CSR corpus, a strong opinion was voiced that such verbalized punctuation (VP) be included in the prompting text. Opposed to this view was the opinion that such predetermined VP may not represent realistic VP, may limit research on automatic punctuation, may restrict the task and perplexity, may unduly burden the corpus with VP words, and may present a difficult and artificial reading task to users. As a result, a compromise position was taken in which half of the corpus was collected in VP mode and half in non-VP mode.

- Speaker-independence: The CSR corpus, although directed primarily toward speaker-independent recognition, also supports research into speaker dependent recognition. Approximately half of the pilot corpus is dedicated to speaker-dependent work.

- Microphone independence: The primary microphone is the traditional Sennheiser model HMD-414. In addition, all data were collected also with a secondary microphone. Previously, this second microphone was a single far-field pick-up microphone, such as the desktop Crown model PZM-6FS. The CSR pilot corpus represents a departure from this practice and a first attempt at true microphone-independent recognition (in much the same spirit as speaker-independent recognition) by using one of many different microphones for the alternate (secondary) speech channel.

- Transcription: For the CSR pilot corpus, the original source text was preprocessed to produce a string of words that represented as well as practical the string of words that would result from reading the source text. This word string was then presented to the subject as the prompting text. This approach provided a very efficient transcription mechanism, because the prompting text could automatically be used as the transcription (except when the subject made errors in reading). Also, the language model, although perhaps a bit unnatural to the extent that the prompt string doesn't represent the statistics of the true language model, can be more easily and comprehensively estimated by preprocessing large volumes of text rather than by transcribing relatively small amounts of speech data.

The CSR Corpus Coordinating Committee

The charter of the CSR Corpus Coordinating Committee (CCCC) is to coordinate CSR corpus development and to resolve issues which arise in CSR corpus development and evaluation. There are currently 12 members of the CCCC, namely:

Janet Baker, Dragon
Jordan Cohen, IDA
George Doddington (chairman)

1. The design of the CSR pilot corpus is described in detail in the paper by D. Paul and J. Baker in this workshop's proceedings entitled "The Design for the Wall Street Journal-based CSR Corpus".
2. The spontaneous speech data collection effort is described in detail in the paper by J. Bernstein and D. Danielson in this workshop's proceedings entitled "Spontaneous Speech Collection for the CSR Corpus.

Francis Kubala, BBN
Dave Pallett, NIST
Doug Paul, Lincoln Labs
Mike Phillips, MIT
Michael Picheny, IBM
Raja Rajasekaran, TI
Xuedong Huang, CMU
Mitch Weintraub, SRI
Chin Lee, AT&T

This committee was formed at the SLS coordinating committee meeting in October 1991. Since that time the committee has met ten times, mostly via teleconference. CCCC activities have included:

- Definition of procedures for microphone gain adjustment and calibration.
- Definition of procedures for transcribing the speech data.
- Monitoring progress in speech data collection and transcription.
- Definition of the data distribution schedule and format.
- Definition of procedures for evaluation of vocabulary/speaker adaptive systems.
- Definition of procedures for scoring.
- Definition of recommended baseline performance evaluations.

The CSR pilot corpus

One of the primary motivations for creating the CSR task and corpus was to provide a sufficiently large corpus of data to properly support advances in speech recognition technology. This implies a very large effort, with many hundreds of hours of speech data being collected. Given the massive effort required, and appreciating the untried nature of many of the corpus parameters, it was decided that a pilot corpus should be collected first to determine the correctness of the many corpus design decisions and to allow modifications of these as necessary.

The CSR pilot corpus is described in a companion paper in these proceedings entitled "The Design for the Wall Street Journal-based CSR Corpus" by D. Paul and J. Baker. This corpus provides for the development and evaluation of both speaker-independent (SI) and speaker-dependent (SD) recognition. It uses the now-standard DARPA corpus approach of providing a three-part corpus: speech data for training the speech recognition system ("TRAINING"), speech data for developing and optimizing the recognition decision criteria ("DEVELOPMENT TEST"), and speech data for performing the formal performance evaluation ("EVALUATION TEST").

The CSR February 1992 dry run evaluation

The recommended baseline performance evaluations were defined by selection of training data set(s), testing data set(s), recognition conditions (vocabulary and language model), and scoring conditions. In the course of discussion on these issues it became clear that consensus was not possible on definition of a single set of evaluation conditions. This was in addition to the distinct differences between speaker-dependent (SD) and speaker-independent (SI) evaluation data and conditions. Some committee members felt that there should be no constraint on training material, to allow as much freedom as possible to improve performance through training data. Others believed strongly that calibration of performance improvement was paramount and therefore all sites should be required to use a single baseline set of training data. In the end, the committee was able only to identify a number of different training and test conditions as "recommended" alternatives for a baseline evaluation.

For training the recommended SI training corpus comprised 7240 utterances from 84 speakers. The recommended SD training corpus comprised the 600 training sentences for each of the 12 SD speakers. For the large-data speaker-dependent (LSD) training condition, the recommended SD training corpus comprised the 2400 training sentences for each of the 3 LSD speakers.

For testing there were a total of 1200 SI test utterances and 1120 SD test utterances. These data comprised, similarly and separately for SI and SD recognition, approximately 400 sentences constrained to a 5000-word vocabulary, 400 sentences unconstrained by vocabulary, 200 sentences of spontaneous dictation, and these 200 sentences as read later from a prompting text.

The vocabulary and language models used for the above-defined test sets were either unspecified (for the spontaneous and read versions of the spontaneous dictation), or were the 5000-word vocabulary and bigram grammar as supplied by Doug Paul from an analysis of the preprocessed WSJ corpus. (Actually, two different sets of bigram model probabilities were used, one modeling verbalized punctuation and one modeling nonverbalized punctuation. These two were used appropriately for the verbalized and nonverbalized punctuation portions of the test sets, respectively.)

Given the rather massive computational challenge of training and testing in such a new recognition domain, with larger vocabulary and greater amount of test data, not all of the test material was processed by all of the sites performing evaluation. Also, because of the variety of training and evaluation conditions, few results were produced that could be compared across sites. Two test sets, however, were evaluated on by more than a single site: Two sites produced results on the SD 5000-word VP test set (Dragon and Lincoln), and three sites produced results on the SI 5000-word

VP test set (CMU, Lincoln, and SRI). These results are given in a companion paper on "CSR Pilot Corpus Performance Evaluation" by David Pallett.

Future CSR corpus effort and issues

Several issues have been identified that bear on the CSR corpus and on potential changes in the design of the corpus:

- Verbalized punctuation. There is a significant argument to discontinue verbalized punctuation, for several reasons: It doubles the number of language models and test sets and thus the number of evaluation conditions. It is artificial in the sense that it is statistically unlike normal dictation, it is more difficult for many subjects to read, and it seems superfluous to the development of the underlying speech recognition technology.

- Preprocessed prompting text. There is argument to prompt the user with the natural unpreprocessed text from the WSJ rather than with the preprocessed word strings as produced by the text preprocessor. The reason is that the word strings do not represent the actual statistics of natural speech (see the companion paper by Phillips et. al entitled "Collection and Analyses of WSJ-CSR Data at MIT").

- Spontaneous speech. There is argument that the current paradigm for collecting spontaneous speech is not adequately refined to represent those aspects of spontaneous speech that are important in actual usage, and that spontaneous speech should remain in an experimental and developmental mode during the next CSR corpus phase.

- Adaptation. Speaker adaptation and adaptation to the acoustical environment has emerged as a major interest. It is clear that adaptive systems must be accommodated in the next phase of the CSR corpus.

- CSR corpus development effort. It is acknowledged that the CSR corpus development effort is a key activity in the support and direction of CSR research, and that this effort therefore requires program continuity and should not be treated as an occasional production demand that can be easily started and stopped.

These issues are currently under debate in the CCCC, and the next installment of the CSR corpus, to be called the CSR corpus, phase two, will no doubt reflect a continued distillation of opinion on these issues.