The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text

Christopher Cieri, David Miller, Kevin Walker

University of Pennsylvania, Linguistic Data Consortium, Philadelphia, PA, USA
{ccieri,damiller,walkerk}@ldc.upenn.edu

Abstract

This paper describes, within the context of the DARPA EARS program, the design and implementation of the Fisher protocol for collecting conversational telephone speech which has yielded more than 16,000 English conversations. It also discusses the Quick Transcription specification that allowed 2000 hours of Fisher audio to be transcribed in less than one year. Fisher data is already in use within the DARPA EARS programs and will be published via the Linguistic Data Consortium for general use beginning in 2004.

Introduction

Although progress in automatic speech recognition (ASR) has been rapid in many areas, conversational telephone speech has proven a consistent challenge for developers of (ASR) technology developers. Low bandwidth, noise, echo, distortion, differences in handset and telephone network, variation among speakers and the constant evolution of vocabulary have conspired to thwart attempts to build robust, high accuracy, continuous, large vocabulary systems. Chief among the problems addressed by the DARPA EARS program is the accuracy of conversational ASR systems. Chief among the needs of the program is a greater volume of transcribed telephone speech than has ever been available previously. This paper will describe a unique corpus of conversational telephone speech and the collection protocol under which it was created.

Data Needs within DARPA EARS

The DARPA EARS (Effective, Affordable, Reusable Speech-to-Text) program addresses the need for systems that generate high accuracy, readable transcripts (EARS 2004). Using the *common task* research management paradigm, EARS focuses attention on two data types: broadcast news (BN) and conversational telephone speech (CTS) across three languages: English, Chinese and Arabic. The program coordinates data collection, research, system development and technology evaluation in order to meet aggressive performance criteria. For example, EARS is strictly required to show annual progress adequate to culminate in real-time CTS systems that produce a five-fold improvement in accuracy. In contrast to some other common task programs, EARS go/no-go criteria encourage sites to cooperate in order to meet performance goals rather than compete against each other.

EARS sites include BBNT, Cambridge University, Columbia University, IBM, ICSI, IDIAP, LIMSI, Lincoln Laboratories, LDC, Microsoft Research, NIST, SRI, University of Pittsburgh and University of Washington who attack research problems individually and in multiple, sometimes overlapping teams. Research areas are designated: Novel Approaches, Speech-to-Text (STT) and MetaData Extraction. (MDE). The goal of the MDE area is to produce transcripts that indicate who said what when, identify disfluent speech and tag discourse structure so that its output can be either formatted for reading or processed by downstream systems.

The combination of challenging end goals with annual milestones and multi-disciplinary approaches drives the program forward aggressively. A critical ingredient in EARS' recipe for ultimate success is data. EARS sites require raw broadcast news and telephone conversations in English, Chinese and Arabic with transcripts and annotations to support metadata annotation in volumes never before available. The paragraphs that follow describe just a subset of EARS data activities specifically those dedicated to collecting and transcribing English conversational telephone speech using the Fisher protocol and Quick Transcription specification developed for this purpose.

Fisher Compared to Previous Telephone Collection Protocols

The Fisher telephone conversation collection protocol was created at the Linguistic Data Consortium (LDC) to address a critical need of developers trying to build robust ASR systems. Previous collection protocols, such as CALLFRIEND and Switchboard-II and the resulting corpora have been adapted for ASR research but were in fact developed for language and speaker identification respectively. Although the CALLHOME protocol and corpora were developed to support ASR technology they feature small numbers of speakers making telephone calls of relatively long duration with narrow vocabulary across the collection. CALLHOME conversations are challengingly natural and intimate. Under the Fisher protocol, a very large number of participants each make a few calls of short duration speaking to other participants, whom they typically do not know, about assigned topics. This maximizes inter-speaker variation and vocabulary breadth although it also increases formality.

Previous protocols such as CALLHOME, CALLFRIEND and Switchboard relied upon participant activity to drive the collection. Fisher is unique in being platform driven rather than participant driven. Participants who wish to initiate a call may do so; however the collection platform initiates the majority of calls. Participants need only answer their phones at the times they specified when registering for the study.

To encourage a broad range of vocabulary, Fisher participants are asked to speak on an assigned topic which is selected at random from a list, which changes every 24 hours and which is assigned to all subjects paired on that day. Some topics were inherited or refined from previous Switchboard studies while others are new.

Another important goal of the Fisher collection was to provide a representative distribution of subjects

across a variety of demographic categories including: gender, age, dialect region and English language fluency. Subjects also indicated their level of education, occupation and competence in languages other than English.

Although the targets for the Fisher collection have evolved with the DARPA EARS program that sponsored it, the most ambitious target expressed was to produce 2000 hours of conversational speech data from calls which individually lasted no more than ten minutes and from which eight minutes were generally expected to be useful. In this context "useful" means containing conversation on the assigned topic – or some other topic negotiated between the participants – and excludes greetings and leave takings.

The 2003 Fisher English Collection

Fisher data collection began in earnest in December 2002 and continued for nearly one year. The study began with the assumption that subjects would make from one to three calls each. However, in order to facilitate the selection of evaluation test sets that included subjects not seen elsewhere in the study, there was a period of nearly four months in which the Fisher robot operator only placed calls to unique subjects. Once the initial evaluation sets were developed, we removed this limitation and the rate of collection better than tripled from an average of 15 calls per day to an average of 54 calls per day.

In order to recruit the volume and diversity of subjects desired for Fisher it was necessary to automate recruitment and registration. The study was announced in newsgroups and Google banners related to linguistics, speech technologies and job opportunities. LDC placed print advertisements in the largest markets of each of the major dialect regions. E-mail lists devoted to finding money making opportunities for their members eventually discovered and publicized the study independently of LDC.

Subjects registered primarily via the Internet by completing an electronic form though a small percentage also called LDC's toll-free number. During registration subjects learned the terms of the study – in particular that their conversations would be collected for purposes of education, research and technology development, that their identities would be kept confidential and that they would be compensated per full-length call.

As mentioned above, Fisher's robot operator assigned a new topic each day from an initial list of forty. Toward the end of the first phase, LDC introduced a second set of 60 topics that were used throughout November and December of 2003. Thus each of the first set of topics was used 8 or 9 times. Each of the second set appeared once.

The main challenge in Fisher 2003 collection, other than maintaining a balance of male and female subjects was recruiting enough subjects to keep up with the collection platform. Because Fisher subjects typically complete only one to three calls, the platform accepted and then retired subjects as fast as the recruiters could register them. Even though registration was almost completely automatic, and recruiter teams could easily process up to 500 registrations per day, the number of actual registrations was consistently well lower than what the platform could handle.

Fisher Outcomes

During its first cycle of operation, the Fisher protocol has produced more data at faster rates than had been collected at LDC in the previous 12 years combined. In just over eleven months, LDC collected 16,454 calls averaging ten-minutes in duration and totaling 2742 hours of audio or 37% more than the most aggressive target requested by sites and sponsors.

Fisher sought to collect CTS data from representative sample of the U.S. population. The first parameter such studies generally attempt to balance is gender. Although one may target a perfect 50-50 gender distribution among subjects, previous studies have shown that in the United States females join such studies more frequently and participate more fully that males. Previous studies have struggled to attain a 60/40 female to male ration. In Fisher, gender is better balanced with females making just 53% of all calls.

In order to model the speaker population fully, age variation is also important. Many previous studies of conversational telephone speech have exploited college student populations. In Fisher, 38% of subjects are aged 16-29 while 45% are aged 30-49 and 17% are over 50.

In contrast with previous Switchboard collections which were regionally based, Fisher subjects represent a variety pronunciations including U.S. regional pronunciations, non-U.S. varieties of English and foreign-accented English. Using the major dialect boundaries drawn by William Labov's Phonological Atlas of North America (2003), the Fisher project recruited subjects primarily from the four major United States dialect regions: North, Midland, South and West. However the study also admitted Speakers from Canada, non-native speakers of English and speakers of other national varieties of English outside the US. Figure 1 shows the distribution of Fisher subjects according to these regional and dialect categories.

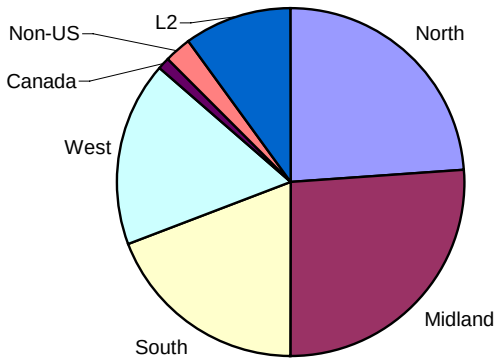

Figure 1: Regional/Dialect Distribution of Fisher Speakers

The distribution of calls by topic depends upon several factors including the days of the week on which it appeared. We used an initial list of forty topics through most of the first phase of collection cycling through that list nearly nine times. We introduced a second set of sixty topics toward the end of 2003 so that each new topic

appeared only once. There are on average 400 calls for each daily topic from the first list and a few dozen for the second topic list. However, even within these lists there is considerable variation in the number of calls per topic. For example a topic that appeared eight times of which three were typically slow weekend days would have had fewer on-topic calls (as few as 207) than another used nine times of which only two were weekend days (as many as 648). The major cause of this variation seems to be the availability of subjects and their tendency to answer the phone more reliably during the week than they do during weekends.

QTr: The Quick Transcription Specification

Naturally, in order to support ASR research, it was necessary to transcribe the thousands of calls collected under Fisher. However, normal rates for the careful transcripts of conversational telephone speech average 20xRT or 20 hours for each hour of conversation per channel. Using such a transcription specification would have required over 100,000 hours of human effort. Instead LDC and EARS sites developed Quick Transcription (QTr) Specifications that require only 6 hours of effort for every hour of speech. One of these specifications was developed at LDC; the other was developed by BBNT and based upon a commercial transcription service provider, WordWave, International (WordWave 2004). The LDC variant relies upon automatic segmentation of the conversational audio into utterances of two to eight seconds in duration. Transcriptionists then make a single pass over the audio creating a verbatim record of the conversation. They make no special effort to provide capitalization, punctuation or special indicators of background noise, mispronunciations, nonce words or the like. Transcriptionist do tag non-lexemes from a small standard lists and marking acronyms read as a sequence of letters. The BBN/WordWave variant transcribes the entire conversation without initial segmentation but then applies forced alignment to yield time-stamps for each word. Although these approaches were expected to yield lower quality transcripts it was anticipated that the much greater volume of data available would compensate. In fact, initial experiments among EARS sites suggest that transcripts created this way are, for purposes of training ASR systems, are of equal value to the considerably more expensive transcripts created under the HUB-5 specification, for example. As a result EARS funded large scale transcription using the QTr specification with LDC producing 200 hours and BBN/WordWave producing 1735 hours.

Conclusion

This paper began by describing the DARPA EARS program and its intensive need for large volumes of high quality broadcast news and conversational telephone speech in English, Chinese and Arabic with transcriptions and annotations to support the EARS Metadata Extraction task. Next we focused on the particular need for a large collection English conversational telephone speech which features short conversations from a large number of subjects on assigned topics in order to maximize interspeaker variation and vocabulary coverage. We then described LDC's activities in collecting Fisher English data during 2003 including the yields of the collection Finally, we outlined the Quick Transcription specification that allowed 2000 hours of Fisher audio to be transcribed in less than one year. Fisher data is already in use within the DARPA EARS program as training material and in the RT-03 Rich Text Evaluation coordinated by the National Institute of Standards and Technology (NIST 2004). The Linguistic Data Consortium plans to begin publishing Fisher data for general use in 2004.

References

Cieri, Christopher, David Miller, Kevin Walker, From Switchboard to Fisher: Telephone Collection Protocols, their Uses and Yields, Proceedings of EuroSpeech 2003.

EARS, 2004, Web Page http://www.darpa.mil/ipto/programs/ears/

Labov, William, 2004, Phonological Atlas of North America, http://www.ling.upenn.edu/phono_atlas/home.html

Linguistic Data Consortium, 2004, Catalog, http://www.ldc.upenn.edu/Catalog

National Institute of Standards and Technologies, 2004, Benchmark Tests Web Site, http://www.nist.gov/speech/tests/

WordWave International, 2004, Web Page, http://www.wordwave.co.uk/