MULTILINGUAL SPEECH DATABASES AT LDC

John J. Godfrey

Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104

ABSTRACT

As multilingual products and technology grow in importance, the Linguistic Data Consortium (LDC) intends to provide the resources needed for research and development activities, especially in telephone-based, small-vocabulary recognition applications; language identification research; and large vocabulary continuous speech recognition research.

The POLYPHONE corpora, a multilingual "database of databases," are specifically designed to meet the needs of telephone application development and testing. Data sets from many of the world's commercially important languages will be available within the next few years.

Language identification corpora will be large sets of spontaneous telephone speech in several languages with a wide variety of speakers, channels, and handsets. One corpus is now available, and current plans call for corpora of increasing size and complexity over the next few years.

Large vocabulary speech recognition requires transcribed speech, pronouncing dictionaries, and language models. To fill this need, LDC will use the unattended computer-controlled collection methods developed for SWITCHBOARD to create several similar corpora, each about one-tenth the size of SWITCHBOARD, in other languages. Text corpora sufficient to create useful language models will be collected and distributed as well. Finally, pronouncing dictionaries covering the vocabulary of both transcripts and texts will be produced and made available.

1. MULTILINGUAL DATABASES

In its nearly two year history, LDC has assembled substantial resources for linguistic research on English: more than a half billion words of text, many hundreds of hours of speech, syntactically tagged corpora, and the beginnings of a multipurpose lexicon of English syntactic features, word senses and pronunciations. With one third of its members from outside the US, and with increasing interest everywhere in expanding the scope of linguistic technologies to other languages, LDC is increasingly called upon for resources in languages other than English.

Other papers in this session describe the efforts under way to secure large and useful bodies of text in other languages, and to develop lexicons or pronouncing dictionaries. This paper will focus on speech corpora, and for the most part on those which are expected to be available in the next year or two.

1.1. Telephone applications

This is currently the leading edge of commercial interest, because of the simplicity of most telephone applications, their large scale, and the international and multilingual nature of telecommunication systems. With the number of telephone companies and their vendors who belong to LDC, this is and promises to remain a high priority. The first offerings will be a series of data sets from an international project known as POLYPHONE, in which LDC has played a leadership role. This multilingual "database of databases" is designed to meet the need for:

- adequate training data for the most common telephone applications of SR;
- public evaluation data for each language;
- cross-language comparability of performance;
- a legitimate testbed for language portability.

Specifications of POLYPHONE databases. The idea of a multinational, multilingual, distributed data collection project was first discussed at a meeting of the Coordinating Committee on Speech Databases and Assessment (Cocosda) in October 1992. In addition to its obvious commercial value for developers of telephone speech recognition, such a database was thought to be "precompetitive," in the sense that it involves no new knowledge or advanced capabilities, and is designed more to support general technology research rather than product development. To make wide participation possible, the plan was to keep the cost of collection in the range of one or two dollars per utterance.

In the ensuing months, prospective participants sought funding and exchanged e-mail, culminating in a meeting in April 1993 at ICASSP where a set of broad specifications was proposed. Each POLYPHONE data set would consist of 125,000 to 200,000 utterances, both read and spontaneous, recorded digitally and without human intervention from at least 5000 callers. The callers were to be drawn in roughly equal proportions from both sexes, from three categories of education, and from three or more age brackets. The utterances would include digits, strings of digits, spelled words, names of persons and places, plus a variety of application-oriented words.

The exact content of the vocabulary in each language was left to the local projects and their sponsors, since they might wish to choose words or phrases for particular word-based recognition applications such as banking, catalog shopping, speed dialing, etc. But every data set must also include several sentences or phrases by each caller which, in the ensemble and taken with the other read items, guarantee balanced phonetic coverage of the language in terms of triphones. Sites may accomplish this by choosing material from well-known sets of phonetically balanced sentences, by generating phonetically balanced phrases, or by selecting sentences from very large text collections by a procedure which optimizes phoneme, diphone, and/or triphone coverage. The selection criteria and phonetic statistics should be documented with each database.

All utterances are to be audited and transcribed orthographically; criteria and conventions have been suggested for accepting or rejecting utterances, and for marking unusual speech and nonspeech events, so that a degree of uniformity across datasets can be expected. Participants are encouraged to collect directly from digital lines wherever possible, leaving the data in the original *a*-law or *mu*-law format. The goal is to have a collection which is in some sense representative of the acoustics of the national telephone network, to minimize artifacts of collection, and to sample the population of potential users of telephone-based speech recognition applications. Thus a certain bias toward more affluent or educated users is specifically permitted.

The American English contribution, collected for LDC at SRI International [1] and described in another paper in this session, will be the first POLYPHONE corpus to be published. Others are in various stages of development:

- A Dutch version, co-sponsored by the Speech Expertise Centre (SPEX) and the national telephone company and supervised by Prof. Louis Boves, is partly collected and being transcribed. The PC platform, telephone interface, and commercial application programming software used in this project cost less than $20,000.
- A Flemish version is planned, under the supervision of Prof. Dirk van Compernolle at Louvain University in Belgium. If funded, this project will take advantage of the reusability of much of the Dutch material and software for that language.
- A Spanish version is now in progress at Texas Instruments (TI) in Dallas. The participants will be predominantly native speakers of Southwestern American Spanish. The collection platform is an InterVoice Robotoperator, a commercial product with user modifiable software that interfaces to a T1 telephone line. Pilot data collection is complete, and full scale collection will begin soon. The project title, "Voice Across Hispanic America," harks back to the "Voice Across America" effort carried out at TI in 1989 [2], [3], which was, in many ways, the ancestor of all these automated telephone data collection efforts.
- "Voice Across Japan," a project at the TI laboratories in Tsukuba, Japan, is also in progress [4]. The design and planning of this corpus predate POLYPHONE, and thus there are differences in some parameter choices – more speakers, fewer utterances per speaker, for example. Nevertheless, the resulting database will be generally quite similar to the other POLYPHONE data sets, and TI has expressed willingness to release it through LDC.
- A Swiss French corpus is being collected under the direction of Prof. Gerard Chollet at the Institute Dalle Molle d'Intelligence Artificielle (IDIAP), with sponsorship from the Swiss national telephone company. The platform will be similar to the one used in the Netherlands.
- The Italian telephone laboratory CSELT is also collecting a telephone speech corpus of which the POLYPHONE data will be a subset. Pilot data collection took place in December and January.
- The Taiwanese consortium SIGSLP, which includes the national telephone laboratories as a member, is committed to carrying out a POLYPHONE collection in Mandarin, though funding was not received on the first try.
- Proposals are also under active consideration in other countries, including Denmark and Australia.
- Sponsorship is being sought for German and for standard French.
- LDC may also sponsor collection of a separate POLYPHONE corpus from speakers of English as a second language.

To our current knowledge, the costs of collection were as predicted, but intellectual property rights were a serious concern. In some cases, LDC must still negotiate for the right to distribute the individual POLYPHONE corpora, but most of the sponsors seem willing to allow this, at least within a few months of completion of the project. Others may wish to distribute on their own terms.

Each of these corpora, with transcriptions and supporting database tables and documentation, will occupy about 6 to 10 CD-ROMs; the 200,000 American English utterances, for example, average about 4 seconds per file (with some silence around each), so they will amount to about 6.4 Gbytes of (8-bit *mu*-law) sampled data at 8 kHz. Subsets of talkers will be marked for development and evaluation testing.

1.2. Language Identification Research

In addition to government funded research, there is clearly a basis for commercial interest in this area, since a language ID algorithm can serve as a gateway to any other telephone-based application in much the same way as speaker ID can. In order to be useful for language ID research, data must be gathered in such a way that selections are otherwise indistinguishable by virtue of channels, talkers, environmental variables, or other artifacts of collection. This means that most data sets collected for any other purpose will almost inevitably be inappropriate, especially if each language comes from a different location, since even slight channel differences will betray the language.

OGI corpus. The first publicly available data intended for language ID research is a collection of prompted telephone responses collected at the Oregon Graduate Institute [5], now available from LDC. It contains speech in eleven languages from about 90 native speakers each. They were recorded at a single site in the US over conventional long distance telephone lines, using a PC, an A/D converter, and a telephone interface.

The languages are English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. The speech is a mixture of brief responses to questions (e.g., days of the week) and short monologs (up to a minute) on unrestricted topics. Up to two minutes of speech was collected from each caller, and there are about 90 callers for each language. About ten percent of the calls are transcribed phonetically and aligned in time with the signal; the remainder have been audited to check for the contents, but not transcribed.

The LDC version of the OGI Multilingual corpus on two CD-ROMs will have a suggested division by callers into training (50 callers), development test (20), and evaluation test (20) subsets for each language. This division is, in fact, being used by NIST in a government technology evaluation program. Future data of this type will be published as it becomes available.

New Corpus in 1995 Another resource which will be useful for language ID is the CALL HOME corpus, described in more detail in the next section. CALL HOME data will resemble SWITCHBOARD in being spontaneous two-way conversational speech with each side being recorded separately. The calls will be between native speakers of many languages besides English, and although all will be initiated from within the US, many will be international calls. Since country-specific channel information might betray the language, the complete CALL HOME conversations will probably not be usable for language identification research. However, the domestic sides, i.e., at least half of each call, should be unbiased and thus appropriate for this purpose.

1.3. Large Vocabulary Speech Recognition (LVSR)

The issue of porting LVSR technology from one language to another is attracting increasing interest. For example, the SQALE project, recently begun in Europe, will apply the ARPA evaluation paradigm next year to three sites, each of which must develop speech recognition capability in at least two languages. Not only is there much interest in the "portability" of speech recognition technology across languages, but there are also research systems which use the speech recognition "engine" for other more limited tasks whose dependence on language models is minimal, such as speaker recognition, word spotting, and other applications. How language independent are these technologies? Only with comparable data across several different languages can such issues be addressed.

CALL HOME The CALL HOME corpus will consist of telephone conversations gathered somewhat in SWITCHBOARD style [6], that is:

- automatically, with computer prompting but no human intervention;
- digitally, with no A/D conversion except at the speakers' locations;
- fully duplex, with each side recorded separately;
- from anywhere in the US, using an 800 number;
- transcribed verbatim, at least up to 10 minutes;
- time aligned between signal and transcript, at least at speakers' turns.

In contrast to SWITCHBOARD, the recordings will be:

- unprompted as to topic;
- international as well as domestic;
- limited to one call per participant;
- uncontrolled as to who is called;
- up to 30 minutes in length.

Within the next year, several hundreds of these calls will be recorded and transcribed in Mandarin, Japanese, Spanish, and English; hundreds more will be collected in other languages and used for language identification research as described in the last section. If the collection paradigm proves successful, transcription will go forward on these and other languages in following years.

Text Collections. To be widely useful, LVSR data sets must include not only speech and transcripts but a language model (or texts from which to construct one), and, in most cases, a pronouncing dictionary. The simplest, and in fact perhaps the only practicable means of providing the amount of text required to build useful language models in several different languages is to acquire newspaper or newswire texts in bulk. LDC plans to make available on CD-ROM, therefore, tens of millions of words in Japanese and Mandarin, and perhaps 100 million in Spanish. The majority of this will be acquired by daily spooling of newswire services. Apart from the use of the Standard Generalized Markup Language (SGML) for demarcation of the higher level units, details of the formats in which these texts will be distributed is still open for discussion.

Lexicons. Pronouncing dictionaries will be produced for the main CALL HOME languages, i.e., those for which transcripts are produced. The current design calls for each lexicon to cover the pronunciation and part of speech of at least all the words used in the CALL HOME transcripts, the words in the text corpora used for language models, and any accidental gaps in the "core vocabulary" of the language after that. The definition of the "core," and any other information that might be provided, will depend on the language and on what resource materials are available at the time the lexicon is developed. A more detailed description of this project is given in the companion paper in this session by Mark Liberman.

References

1. J. Bernstein, K. Taussig, and J. Godfrey, "MACROPHONE: An American English Telephone Speech Corpus For the POLYPHONE Project," *Proceedings ICASSP-94*.
2. B. Wheatley and J. Picone, "Voice Across America: Toward Robust Speaker Independent Speech Recognition For Telecommunications Applications", *Digital Signal Processing: A Review Journal*, vol. 1, no. 2, pp.145-64, April 1991.
3. J. Picone and B. Wheatley, "Voice Across America: A Step Towards Automatic Telephone Transactions," *Voice Processing Magazine*, pp. 146-47, February 1991.
4. T. Staples, J. Picone, K. Kondo, and N. Arai, "The Voice Across Japan Database: The Japanese Language Contribution To POLYPHONE," *Proceedings ICASSP-94*.
5. Y. Muthusamy, R. Cole, and B. Oshika, "The OGI Multi-Language Telephone Speech Corpus", *Proceedings ICSLP-92*, pp. 895-898.
6. J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proceedings ICASSP-92*, pp. I517-I520. 895-898.