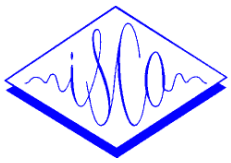

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

TELEPHONE SPEECH CORPUS DEVELOPMENT AT CSLU[1]

Ronald Cole, Mark Fanty, Mike Noel and Terri Lander

Center for Spoken Language Understanding
Oregon Graduate Institute of Science & Technology
P.O. Box 91000, Portland, OR 97291-1000 USA

ABSTRACT

This paper describes eight telephone-speech corpora at various stages of development at the Center for Spoken Language Understanding. For each corpus we describe data collection procedures, methods of soliciting callers, protocol used to collect the data, transcriptions that accompany the speech data, and the expected release date. The corpora are (or will be) available at no charge to academic institutions.

INTRODUCTION

The Center for Spoken Language Understanding (CSLU) collects and transcribes telephone-speech data to enable research activities at CSLU and elsewhere. Corpus development activities are performed by four full-time staff, aided by graduate students and part-time employees. In 1994, we anticipate collecting and transcribing speech from 10,000 callers in twenty languages. Corpus development activities are supported by industrial memberships and research grants.

Corpus development activities at CSLU include: (a) collecting telephone speech data in different languages; (b) transcribing speech at word and phonetic levels; (c) developing and documenting transcription conventions for each level; (d) measuring the level of agreement among transcribers; (e) developing interactive speech tools for labeling; (f) distributing the speech corpora to academic institutions free of charge; and (g) placing speech tools and labeling conventions in the public domain for use by others.

In this section, we present some general information about our corpus development activities. In the following sections we will describe individual corpora.

Data Collection. Telephone-speech data are collected over analog and digital telephone lines. Prior to November, 1993, speech data were collected over analog lines using several Gradient Technology Desklabs. Since November, 1993, the majority of our data has been collected using a 24 channel T1 line connected to three LINKON FC3000 Communication Boards. We are also using an Apple GeoPort Telecom Adapter connected to a Macintosh Quadra A/V to collect analog speech data for one of the corpora to be described.

Transcription. Each call is processed by one or more listeners. Responses are verified to determine that the caller followed instructions; some are also transcribed.

Transcription of corpora occurs at three different levels: non-time-aligned word level, time-aligned word level, and time-aligned phonetic level. Non-time-aligned word level transcriptions provide an orthographic representation of the utterance, including indications of extra-speech events such as breathes or lip smacks. Time-aligned word level transcriptions augment the transcription by aligned each word to the acoustic signal. Time-aligned phonetic transcriptions align phonetic symbols to the acoustic signal.

A precise description of the conventions used for all levels of labeling, including a complete list of all phonetic labels for each language, is presented in the CSLU conventions document [2].

Transcription Reliability. We are conducting experiments to determine the level of agreement among labelers. In these experiments, CSLU staff and professional phoneticians are using Worldbet [3] to transcribe the same intervals of speech. Initial results for English indicate overall agreement of approximately 70%.

Speech Tools. The OGI Speech Tools support data manipulation, analysis and display [4]. All corpus development activities are performed using these tools. They were developed at CSLU, then made portable and documented for distribution with support from NSF. The tools have been made available to the research community through anonymous ftp.

CORPORA

The first three corpora described in this section are considered to be complete and are now available from CSLU. They were collected over an analog telephone line using a Gradient Technology Desklab. The data

[1]This article is an updated version of a report published in [1]. We publish this updated report here to reach a wider audience.

were digitized at 8000 samples per second with a 14 bit resolution. All data are stored in the NIST wav file format, some with MIT shortpack compression. The remaining corpora are under development and estimated release dates are provided for each.

Spelled and Spoken Names Corpus

The Spelled and Spoken Names Corpus [5] contains utterances from 3667 calls. Callers were solicited through computer newsgroups and a public relations campaign initiated by OGI. The majority of callers were from the Pacific Northwest. The proportion of male to female callers is 1.15 to 1.

Callers spelled their name, both naturally and with pauses. They said their name. They said the city they were calling from and the city and state where they grew up. They said the alphabet with pauses. They answered two yes/no questions. Finally, they gave their address if they wanted a small reward for calling (recordings of addresses are not available, but excised number strings are—see OPERA section).

Documentation of the Spelled and Spoken Name Corpus includes a speaker-by-speaker log file containing non time-aligned word-level transcriptions of each utterance. Each utterance was transcribed by two separate listeners and any inconsistencies were resolved. The log also contains judgments of gender, age, connection quality, accent, and intelligibility. In addition, occurrences of extraneous speech, environmental noise, excessive breath, or line noise are indicated in the log file for each utterance.

A subset of the data was transcribed at the time-aligned phonetic level. The utterances were labeled by hand, then labels and time-alignments with the speech spectrogram were verified by an expert spectrogram reader. The phonetically labeled data include 100 alphabets, 2052 city names, 100 spoken names and 300 spelled last names.

Enhanced OGI Multi-Language Corpus

The OGI Multi-Language Telephone-Speech Corpus [6] consists of telephone-speech from 10 languages: English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The initial corpus included 900 calls—90 calls for each language.

Callers were solicited through computer newsgroups. Each caller was asked to respond to the following prompts: (1) What is you native language? (2) What language do you speak most of the time? (3) Please recite the seven days of the week; (4) Please say the numbers zero through ten; (5) Tell us something that you like about your hometown; (6) Tell us about the climate in your hometown; (7) Describe the room that you are calling from; (8) Describe your most recent meal.

In addition, unconstrained speech was obtained by asking callers to speak for one minute on any topic of their choice (hereafter "stories").

Each utterance was listened to by a native speaker of the language to verify that the caller responded appropriately. The native speaker also made judgments concerning the caller's gender, the caller's age, and the line quality.

The enhanced corpus is augmented with: (a) 200 Hindi calls; (b) speech files that were collected during the original collection but were not included in the original distribution; and (c) time-aligned phonetic transcriptions of over five hours of speech (up to 50 sec per call) in six languages: English, Japanese, German, Spanish, Hindi, and Mandarin. For the phonetic transcriptions, we have adopted the Worldbet labeling scheme, a set of orthographic symbols for multi-language transcription that correspond to IPA symbols. The rationale for using Worldbet and the symbol inventory for each language is provided in [2].

Stories Corpus

Collection for the OGI Multi-Language Corpus produced additional calls from English speakers not included in the Multi-Language Corpus. The Stories Corpus consists of up to 50 sec of spontaneous speech from the "stories" response of 692 English calls. All 692 calls have been transcribed at the non-time-aligned word level, 300 at the time-aligned word level, and 200 at the time-aligned phonetic level.

Twenty-one Language Corpus

CSLU plans to collect and verify calls from at least 200 fluent native speakers in 21 languages—Eastern Arabic, Cantonese, Czech, Farsi, French, German, Hindi, Hungarian, Japanese, Korean, Malay, Mandarin, Italian, Polish, Portuguese, Russian, Spanish, Swedish, Swahili, Tamil, and Vietnamese. Verification and global judgments will be performed by native speakers.

The protocol will be presented to the caller in their language. The English version includes the prompts: (1) What is your native language? (2) What language do you speak most of the time? (3) What language do you speak at home? (4) What other languages do you speak and understand? (5) How old are you? (6) What is your date of birth? (7) Are you male or female? (8) How long have you been in the United States? (9) In what city and state did you spend most of your childhood? (10) What is your zipcode? (11) What area code are you calling from? (12) What day is today? (13) What time is it? (14) Say a familiar telephone number; (15) How would you ask someone if they speak English? (16) Give us the greeting you usually use when answering the phone; (17) Describe the route you take to work or to the store; (18) Tell us something that

you like about your hometown; (19) Tell us about the climate in your hometown; (20) Describe the room you are calling from; (21) Describe your most recent meal.

Finally, two longer responses were elicited. The first on any topic of the caller's choice; the second about the caller. Collection has begun. The expected completion data is yet to be determined.

English Census Corpus

In conjunction with the U.S. Bureau of the Census, CSLU is collecting data to develop a prototype automated census system. Callers were solicited from Census Bureau employees, their family members, and family friends in the following cities: Dallas, Chicago, Boston, Charlotte, Atlanta, Philadelphia, Denver, Kansas City, Detroit, and Seattle.

Two protocols were used that differed in the wording of some of the prompts. Each protocol was recorded by both male and female speakers. In addition, male and female synthesized voices were used. Calls were assigned to the eight conditions in rotation.

The following information from the Census short form was solicited: (1) name, (2) sex, (3) birth date, (4) marital status (now married, widowed, divorced, separated, never married—choose one), (5) Hispanic origin (yes/no); if Hispanic: Mexican, Mexican-American, Chicano, Puerto Rican, Cuban or other (specify), (6) race: White, Black or Negro, American Indian (specify tribe), Eskimo, Aleut, Chinese, Japanese, Filipino, Asian Indian, Hawaiian, Samoan, Korean, Guamanian, Vietnamese or other (specify).

In addition, the caller was asked his or her native language, phone number and childhood city, and was asked to evaluate the questionnaire.

Each response will be transcribed at the time-aligned word level, including indications of filled pauses and other non-speech events. Each response will also be assigned a behavior code which characterizes the usability of the response. We are in the process of transcribing the calls that have been collected. We expect that the transcriptions will be completed and the corpus ready for distribution late in 1994.

Cellular Words, Numbers and Alphabet Corpus

This corpus will consist of up to 600 calls made from cellular phones. Each caller answers nine questions, says words that might be used in voice messaging applications, says a familiar phone number, and recites the letters of the English alphabet. Callers are being provided by a private company which helped fund the data collection.

The corpus is being collected using the Gradient Technology Desklab over an analog line. Non-time-aligned word-level transcriptions are being produced.

The caller is prompted with: (1) Are you calling from a cellular phone? (2) If you happen to know if you are calling from an analog or digital phone, please say which one; (3) Are you using a speaker phone? (4) What is your native language? (5) Where were you born? (6) Where did you spend your childhood? (7) What is the month day and year of your birth? (8) Please say your name; (9) Please say the name of the company or organization you are with; (10) Please say a familiar phone number, one digit at a time; (11) We would now like you to recite the English alphabet with a brief pause between letters, like this: A B C D E.

The caller was also prompted for the following words one at a time. Each word was presented in the carrier phrase "Say _____ now": Cancel, Change Greeting, Continue, Copy, Erase, Help, Listen, No, Operator, Pause, Replay, Rerecord, Reply, Resume, Review, Save, Send copy, Yes, Add, Dial , Call, Edit, Callback, Change, Delete, Phone book, Beginning, Choices, End, Directory assistance, Customer support, Next, Repeat, Replay message, Return call, Skip, Tutorial, Customer care, Verify, Scan, Messages, Message, List, Rewind, Fax, Voice, Print.

Currently, approximately 300 calls have been collected and transcribed. We estimate that the corpus will be ready for distribution August 1994.

Words, Numbers and Phrases Corpus

With support from Apple Computer, CSLU is collecting both analog and digital speech data for utterances related to voice messaging and voice control of computer applications. Callers are being provided both by Apple Computer and by CSLU through newspaper advertisements.

The protocol consists of two questions to help determine the caller's language background, followed by instructions to repeat 35 words or phrases given in the prompt. To increase the usefulness of the corpus, several sub-vocabularies, including first names, last names, digits, numbers and days of the week were inserted into the prompts. For example, the phrase "phone ⟨first name⟩" is expanded to 50 different phrases using 50 common first names. There are about 350 different phrases that will be recorded from different speakers.

The goal is to collect 1000 speakers using an Apple Macintosh Quadra A/V and 2000 speakers on the digital T1 system using the LINKON setup.

The protocol first asks the caller's native language and origin, then askes the caller to say the phrases: (1) play previous message again; (2) cancel my ten AM appointment; (3) make a meeting for today; (4) what is my street address; (5) quit; (6) forward this message to my wife; (7) set up a call with ⟨firstname⟩

and ⟨firstname⟩; (8) conference call ⟨lastname⟩ and ⟨lastname⟩; (9) who is at work; (10) stop; (11) what is the area code for this state; (12) add my son to the phone book; (13) remove number ⟨digit⟩ from the directory; (14) hello, what are my messages; (15) skip the next name; (16) help; (17) good-bye; (18) please send a car from the city; (19) dial ⟨number⟩; (20) delete my email tomorrow; (21) cancel; (22) read this text; (23) correct my balance; (24) call my daughter at eleven pm on ⟨day⟩; (25) erase all information; (26) no; (27) record extended phone book; (28) get my office; (29) transfer all calls to home at twelve o'clock; (30) use voice; (31) record urgent message; (32) yes; (33) find the operator; (34) call ⟨firstname⟩; (35) dial ⟨lastname⟩; (36) phone ⟨firstname⟩; (37) call ⟨number⟩; (38) phone ⟨number⟩;

The data collection is just beginning. We expect this corpus will be available September 1994.

OPERA Corpus

CSLU is collaborating with the International Computer Science Institute (ICSI) in Berkeley to develop speech corpora for Open Performance Evaluation of Recognition Algorithms (OPERA). These corpora will be distributed with designated training and test sets to all researchers who wish to compare recognition performance on a common task. Performance evaluation and summary of results will also be provided.

The first OPERA corpus, now under development, consists of numbers taken from three of the corpora described earlier: the Spelled and Spoken Words Corpus, the Cellular Words, Numbers and Alphabet Corpus, and the English Census Corpus. We estimate the final corpus will consist of about 10,000 different numbers.

Thus far, we have created numbers files from utterances in the Spelled and Spoken Names Corpus in which the caller provided their street address and zipcode. Speech intervals containing numbers found in street addresses, street names (e.g., "fifth") and zip codes were located manually, and new files were created containing just the numbers. A total of 2167 files have been created, from approximately 1300 different speakers. Each file has been transcribed at the non-time-aligned word level and at the time-aligned phonetic level.

AVAILABILITY

CSLU is dedicated to promoting progress in the field of computer speech recognition. To this end, corpora are made available at no charge to academic institutions. These data are available once they are completed. Portions of the Enhanced Multi-Language Corpus have been placed in the public domain.

For information on obtaining any of these corpora, the conventions document, or the speech tools, contact Mike Noel at noel@cse.ogi.edu.

ACKNOWLEDGMENTS

We are indebted to the organizations that helped fund the projects: U.S. Bureau of the Census, ONR, NSF, Linguistic Data Consortium, U S West, Digital Equipment Corporation, LINKON Corporation, and Apple Computer.

Much of the corpus development would have been impossible without the dedicated efforts of the labeling and transcribing staff. Many thanks are due to Terri Durham, Vince Weatherhill, Amie Wilson, Victoria Noel, Alexandra Guerra, Troy Bailey, Johan Schalkwyk, David Cole, Angela Noel and many others.

REFERENCES

[1] R. Cole, T. Lander and M. Noel, "Corpus development activities at the Center for Spoken Language Understanding," *Proceedings of the ARPA Human Language Technology Workshop*, C. J. Weinstein (ed.), Morgan Kaufmann: San Mateo, CA, 1994.

[2] Terri Lander, S. T. Metzler, *The CSLU Labeling Guide*, CSLU, Oregon, February, 1994.

[3] James L. Hieronymus, "Ascii phonetic symbols for the world's languages: Worldbet," *Journal of the International Phonetic Association*, 1993.

[4] CSLU. "OGI speech tools user's manual," Technical report, Center for Spoken Language Understanding, Oregon Graduate Institute, 1993.

[5] R. A. Cole, K. Roginski, and M. Fanty, "A Telephone Speech Database of Spelled and Spoken Names," *Proceedings of the International Conference on Spoken Language Processing*, Banff, Alberta, Canada, October 1992, pp 891–893.

[6] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," *Proceedings of the International Conference on Spoken Language Proceedings*, Banff, Alberta, Canada, October, 1992, pp 895–898.