MACROPHONE: AN AMERICAN ENGLISH TELEPHONE SPEECH CORPUS FOR THE POLYPHONE PROJECT

Jared Bernstein and Kelsey Taussig (SRI International)
Jack Godfrey (Texas Instruments)

SRI International
Speech Research and Technology Program
Menlo Park, CA, 94025

ABSTRACT

Macrophone is a corpus of approximately 200,000 utterances, recorded over the telephone from a broad sample of about 5,000 American speakers. Sponsored by the Linguistic Data Consortium (LDC), it is the first of a series of similar data sets that will be collected for major languages of the world in a cooperative project called Polyphone. It is designed to provide telephone speech suitable for the development of automatic voice-interactive telephone services. In particular, Macrophone contains training material for applications in transportation, scheduling, ticketing, database access, shopping, and other automated telephone interactions. In addition to being phonetically balanced, the spoken material refers to times, locations, monetary amounts, and interactive operations. The utterances are spoken by respondents into telephone handsets and recorded directly in 8-bit mu-law digital form through a T1 connection to the usual switched telephone network. The entire corpus will be made available by LDC in 1994. The paper describes the design of the linguistic materials in the corpus, and the process of solicitation, collection, transcription, and file preparation for the Macrophone corpus.

1. INTRODUCTION

During the second International Conference on Spoken Language Processing, held at Banff in October 1992, a group of scientists from around the world (COCOSDA, the Coordinating Committee for Speech Databases and Assessment) initiated a cooperative project known as ***Polyphone***. Its purpose was to create an ensemble of speech databases in as many of the world's culturally and commercially important languages as possible. These would be designed to support training and testing of technology for telephone speech recognition, sufficiently similar in content for meaningful performance comparisons across languages, and publicly available to encourage research and development by any interested parties. It was agreed that the expertise and technology required to execute such projects at reasonable cost would be shared among all participating groups. Ideally each project would be locally justified and funded; though participants acknowledged the difficulty of enlisting support from some sponsors, they viewed such a "database of databases" as precompetitive and an important community asset.

Macrophone, the American English contribution to Polyphone, will be the first such large database that is generally available. Just as its methods are similar to those first used in the Voice Across America project at Texas Instruments (Wheatley & Picone, 1991), we hope that its design and method of collection can serve as a model and facilitate similar work by others.

2. POLYPHONE SPECIFICATION

The early participants in Polyphone specified several targets for each corpus, including the following:

- Where feasible, sites will record directly from digital telephone lines (e.g., T1 or E1 or ISDN) in the local format (e.g., mu-law or a-law), and provide tables or software for users to convert to linear form.
- Every data set will contain at least 5000 callers and at least 125,000 utterances. A representative subset of approximately 500 calls will be designated for testing, divided into 5 sub-subsets of 100 for multiple tests.
- Calls will be designed to produce between 25 and 50 utterances (responses), including a small number of spontaneous utterances and one prompt inviting comment on the recording session. Each utterance will be stored in a file, preceded and followed by nominal silence.
- The utterances solicited will contain sentences for phonetic coverage, words spoken and spelled, digit strings or numbers in some form, and phrases containing numbers. The utterances will provide 100% phoneme and diphone coverage of the language, and will be designed to maximize triphone coverage, within practical limits.
- Every utterance will be transcribed orthographically, and a database of information on the callers and utterances will accompany the corpus of speech recordings.

The Macrophone corpus meets all these specifications and will be the first of the Polyphone data sets to be published. Other Polyphone efforts underway as of December 1993 include American Spanish, Dutch, Swiss French, and Japanese; several others are in the planning stage.

3. TASK FOCUS

The prospective applications for the Macrophone data partly determined the linguistic design of the material and the population of speakers to be recorded. Examples of the applications include:

- voice interactive systems to support telephone services like collect calls, third-party billing, or rate inquiries
- database information retrieval services that might provide schedule or availability information about transportation or other public services in a limited semantic domain
- systems for ordering theater or stadium tickets, or for making medical or other appointments
- systems for manipulating bank accounts or other financial resources.

4. MATERIAL DESIGN

4.1. Goal

The goal of the Macrophone project was to provide a basic set of common spoken material suitable for training and evaluation of speech recognition systems for telephone-based applications, particularly those that use names, places, times, and numbers in a North American context.

4.2. Sources

The material collected came from pools of prompt texts. Materials were selected from these pools automatically and combined into a prompting sheet that is mailed to a person. In the Macrophone corpus, 45 responses were solicited on each sheet. Of the 45 responses, 34 were read and 11 were spontaneous. The prompt material for the spontaneous utterances was designed to elicit particular responses or types or ranges of responses. The following describes the read and spontaneous material presented, with examples.

Read

3 digits strings: (nnn) nnn-nnnn; nnnn-nnnn-nnnn; and one identification number
3 natural numbers (2 with units): 236 years; 4.32 grams; 7000 tons
4 dollar amounts: $834; $73.27; $1,975.55
1 fraction: 1/4, 7/10, 1/16
2 places: Newark, New Jersey; Paris, France
6 application words: account; check; collect; ticket; visa
2 spelled words: A M B I G U O U S; C L E R K; R A N C H E R
1 date: Friday, January 1, 1993
1 time: 11:50; a quarter to twelve; 4:51
1 name at agency: Susan Crane of the U. S. Postal Service
3 name at street address: Larry Garcia, at 133 Elm St.
7 sentences (3 TIMIT, 2 WSJ, 2 ATIS)

Examples of the sentence types (TIMIT, WSJ, ATIS) are:

Will you please confirm government policy regarding waste removal?
The budget is a long way from completion, however.
I'd like to buy a coach class ticket for a flight from Columbus to San Jose.

For more information on the TIMIT sentences, see Lamel et al. (1986); for the WSJ sentence set, see Paul & Baker (1992); and for the ATIS materials, see Hirschman-MADCOW (1992).

Spontaneous

There were 11 prompts in the interaction that solicited spontaneous speech. Of these, ten were fixed and one was rotating. Some of the fixed questions provided additional demographic information about the speaker. Five of the fixed questions were designed to elicit an answer of yes or no. Six of the fixed questions were printed on each sheet:

Are you ready to start? (y/n)
Are you calling from your home phone? (y/n)
Do you speak any language besides English at home? (y/n)
Please name a major city in your state.
Would you be willing to participate in another study like this one?(y/n)
We would appreciate any comments you may have about this recording session. Please record your comments.

The four unprinted questions were:

Are you using a cordless phone? (y/n)
What is today's date?
What time is it now?
What is your date of birth?

The rotating question was taken from the set:

Please say any number from 1 to 100.
Please say any number from one thousand to one million.
What is your house number?

Prompt Pools

The utterances were selected from pools of prompts that were designed for this project. In particular, each pool was designed with knowledge of the material in the other pools; thus, for example, "Jones" was in the family-name pool, but was excluded from the street-name pool. The prompt pools were:

Place Names — All United States and Canadian cities with population over 150,000, and at least two cities in each state in the United States; all cities worldwide with over 2 million people, and at least one city from most major commercial nations.

Personal Names—A gender-balanced list constructed from the 600 commonest first names and the 600 commonest last names in the United States. Some common first and last names that are also common words were omitted.

Street Names — The most frequent 971 names compiled from the ZIP+4 directory BODY field by counting block faces. We deleted state names, city names, first and surnames that occur in our other lists, and we deleted the letter names A - Z.

Application Words — A list of 674 control words selected from existing and imagined telephone applications.

A fuller description of the materials that were actually collected from those people who called in will be distributed with the database including a statistical analysis of the phonological coverage achieved.

5. DATA FLOW

5.1. Overview

The data collection process has four distinct steps:

- material distribution,
- telephone signal collection,
- verification and transcription,
- package and delivery of the transcribed utterances.

This process is shown in Figure 1 and is described in the following sections.

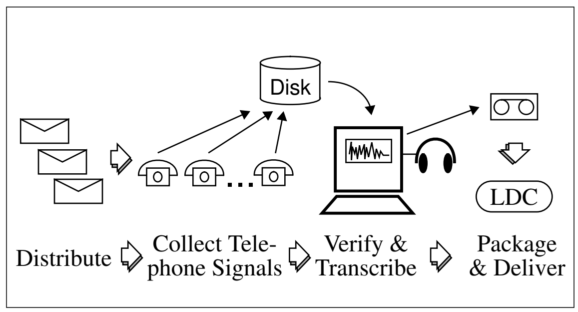

Figure 1. Data Collection Process

5.2. Hardware and Software for Phone Connection

SRI designed and implemented a set of systems for this data collection. Dialogic hardware resident in an IBM-compatible PC provided a digital connection to 10 telephone lines that are available toll-free to callers. The Dialogic/PC system concentrated the data and sent it to the disk of a small Sun Microsystems SPARCstation. As convenient, the data was moved to an archive disk from which the data was archived onto Exabyte tapes and pressed onto CD-ROMs for further manipulation. The labor-intensive part of the Macrophone corpus development was the transcription and verification of the spoken material. This was performed with special-purpose software running on Sun ELCs.

5.3. Time and Resources Used

The design and selection of material, set up and monitoring of the telephone collection system, the verification/ transcription of the utterances, and finally the delivery of the resultant files took about seven months of professional labor and about eight months of semi-skilled labor. After the linguistic material had been decided upon, the project yielded 200,000 utterance files with verified transcriptions and demographic headers in five calendar months.

5.4. Material Distribution

The data collection process began with the distribution of the material to prospective callers. The material was presented in the form of unique **prompting sheets** to guide the caller through the telephone interaction. The sheets were designed following a fixed format, but provide different read material. Each prompting sheet contained a different set of read material, and each caller received a different sheet. A sample prompting sheet is shown in Figure 2.

BEFORE YOU CALL

Please write your six-digit panel identification number in the space provided in the middle of this page. (It is printed at the top of your cover letter.)

PLEASE CALL 1-800-xxx-xxxx

A computer will answer and ask you the following questions.

Computer: *Thank you for calling SRI's voice recording system. Your voice will be recorded and used for research and development of speech technology. If you do not wish to have your voice recorded and used for these purposes, you may hang up now.*

The session will begin with a few questions. Your answers provide us with important information about vocal quality and speech patterns. All of your responses will be kept confidential.

Are you ready to start? (your response)
Are you calling from your home phone? (your response)
Do you speak any language besides English at home? (your response)
•
•
•

Please read your panel identification number which you filled in below:

____

(write your panel identification number here)

Thank you. You will now be asked to read each of the items in the right-hand column.

1.	*(a word)*	subtract
2.	*(a sentence)*	Could you give me a list of all afternoon flights from Los Angeles
3.	*(a telephone number)*	(379) 528-5883
4.	*(a place)*	Canton, China
5.	*(a spelled word)*	M, A, N, K, O, S, K, I
6.	*(a number)*	51,611 meters
7.	*(a name)*	Danny Payne, of 251 Ironwood Way
8.	*(a number)*	4
9.	*(a sentence)*	Military police was to keep the routes open and protect the settled areas.
10.	*(a time)*	9:22
11.	*(a word)*	south
12.	*(a sentence)*	"I hear that only people with money will be approved," she said.
13.	*(a dollar amount)*	$219
14.	*(a name)*	Penny Ward of the I R S
15.	*(a dollar amount)*	$6,588.12
16.	*(a date)*	Tuesday, March 20, 1990
17.	*(a sentence)*	The Statue of Liberty and Ellis Island are within the waters of New York Bay.
18.	*(a word)*	floor
19.	*(a dollar amount)*	$2
20.	*(a word)*	leave
21.	*(a credit card number)*	1495-1772-1515
22.	*(a sentence)*	The causeway ended abruptly at the shore.
23.	*(a name)*	Louise Bradford, of 143 Westfield Road
24.	*(a place)*	Kingston, Jamaica
25.	*(a fraction)*	1/8
26.	*(a dollar amount)*	$600
27.	*(a word)*	slower
28.	*(a spelled word)*	B, R, U, S, H
29.	*(a name)*	Dan Rich, at 388 Beacon
30.	*(a sentence)*	How can I get from the Tacoma airport to downtown.
31.	*(a number)*	17,905 kilograms
32.	*(a sentence)*	It's all psychological.
33.	*(a word)*	free

Thank you.

Would you be willing to participate in another study like this one? (your response)

We would appreciate any comments you may have about this recording session.
Please record your comments. (your response)

Thank you very much for participating in this data collection effort. You may now hang up the telephone.

Figure 2. Sample Prompting Sheet

Prospective callers were solicited through a market research firm that was able to select a sample from their panel of 400,000 U.S. households. Since no incentive was offered for placing the call, a conservative estimate of a 25% response rate was used.

Twenty thousand (20,000) prompting sheets were mailed, which resulted in 6700 calls, at a 33% response rate. The sheets were sent out as six separate mailings of 1000, 2000, 5000, 5000, 5000, and 2000 at approximately one-week intervals.

Calls typically started coming in the day following a mailing and peaked three days after a mailing. Although ten lines were available, all ten were never activated at once. The mailings and the response rate are shown in Figure 3. Each vertical bar shows the date and the number of prompting sheets in a mailing, and the lower line shows the number of calls received per day.

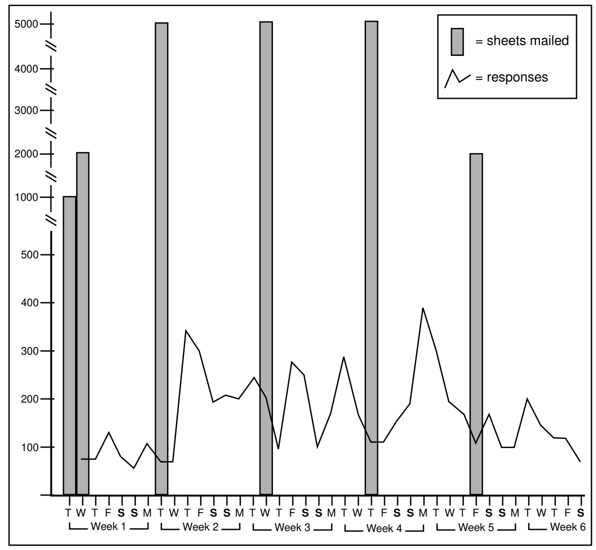

Figure 3. Mailings and Responses

The target population was specified to consist of equal numbers of males and females between the ages of 10 and 80, balanced between ages 20 and 60 and fewer in the 10-19 and 61-80 age groups. The target population was also specified to be geographically balanced according to the latest census figures. The sample population was selected by the market research firm to compensate for different expected response rates. Past experience indicated higher expected response from females than from males, lower response rates among young people, higher response rates among elderly people, and lower response rates from people with household incomes above about $40,000.

Figure 4 shows the mailings sent out (upper curve) and the responses received (lower curve) as a function of age. The low response rate among people aged 20-30 is due to income skew in the mailing sample, and partly due to low response rates among the people in that 20-30 age group who received sheets.

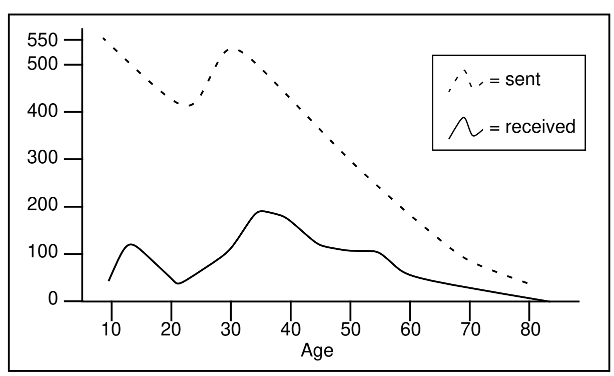

Figure 4. Response as a Function of Age

5.5. Telephone Signal Collection

The recipients of the prompting sheets were instructed to call a toll-free 1-800 number which connected them to one of 10 digital telephone lines set up to receive calls. All data was recorded directly from T1 digital telephone lines in 8-bit mu-law format using Dialogic hardware installed in an IBM-compatible PC. The PC, which operated under Interactive UNIX, was a 33 MHz 386 with 16 MBytes of RAM. Each response was written as a separate file to one of two 2-GByte disks of a Sun Sparcstation ELC. A completed call resulted in about 2 MBytes of data. The data collection system is shown in Figure 5.

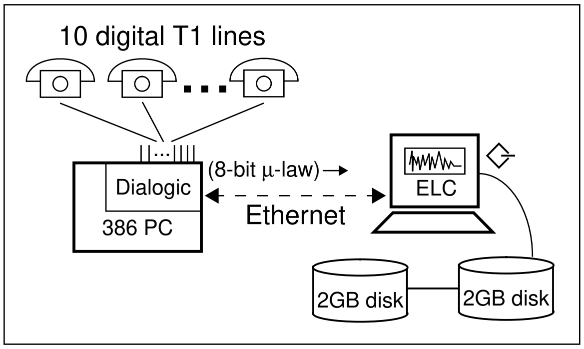

Figure 5. Data Collection System

Software was written to play out prerecorded prompts and record the interaction. As a half duplex system, the data collection system was only capable of either playing out a prompt or recording a response. Care was taken to truncate the end of the prompting text in an attempt to keep callers from responding before the system began recording. It was necessary to remove the written text from the printed sheets for a few of the prompts to force the participant to listen to the entire prompt before responding.

An average telephone call took about six minutes and resulted in about four minutes of collected speech (including two seconds of silence at the end of each utterance). Approximately one third of the collected data was silence.

5.6. Verification and Transcription

Verification of the read responses and transcription of the spontaneous responses was performed by temporary workers using SRI software written for Sun ELC computers.

The data verification occurred in two steps. Since each of the 20,000 sheets was unique, it was necessary to supply a unique sheet identifier to bring up the default transcription for the read material. The sheet identifier was in the form of a 10-digit telephone number and was transcribed in a first pass along with three other responses. These items were then propagated to the headers of all 45 speech files produced from that call. In addition to the demographic responses, a gender indication was included (decision made by the transcriber). The demographic information items are responses to the following prompts:

Do you speak any language besides English at home?
Are you using a cordless phone?
What is your date of birth?

The second step of data verification involves providing an orthographic transcription of each utterance. Each utterance waveform is displayed on the transcriber's console and played through the computer's audio port. Default transcriptions are provided for read data, previously transcribed demographic data, and predictable responses to spontaneous questions such as "Are you ready to start?", "What is today's date?", "Would you be willing to participate in another study like this one?".

Utterance files which did not contain any speech or contained truncated speech (approximately 7.5% of the total) were discarded. Utterances which were difficult to transcribe — those containing word fragments, mispronunciations, and other disfluencies — were set aside for a linguist to review. All other utterances were transcribed according to what was said; the transcriptions also included markings for non-speech events such as background noise, background speech, line noise, mouth noise, and verbal hesitations.

5.7. Package and Delivery

Each file was written as 8-bit mu-law with a SPHERE header. The headers contain information about the data as well as demographic information about the caller. All data files were written to Exabyte tape and shipped to LDC, where the files will be pressed onto CD-ROMs and made available to members of the Linguistic Data Consortium.

6. CONCLUSION

The Macrophone database collection project demonstrates that a large corpus of telephone speech can be solicited, collected, and prepared for use within a specified time and effort. The Macrophone data should be available from the Linguistic Data Consortium by summer of 1994.

ACKNOWLEDGMENTS

This work was supported by the Linguistic Data Consortium. Opinions, findings, conclusions and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Linguistic Data Consortium.

REFERENCES

1. L.Hirschman-MADCOW (1992): "Multi-Site Data Collection for a Spoken Language Corpus," *Proc. DARPA Speech and Natural Language Workshop*, M.Marcus (ed.), Morgan Kaufman. pp. 7-14.
2. L.Lamel, R. Kassel & S.Seneff (1986): "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop*, February 1986, pp. 100-109.
3. D.Paul & J.Baker (1992): "The Design for the Wall Street Journal-based CSR Corpus," *Proc. DARPA Speech and Natural Language Workshop*, M.Marcus (ed.), Morgan Kaufman. pp. 357-362.
4. B.Wheatley & J.Picone (1991) "Voice Across America," *Digital Signal Processing*, 1, pp. 45-63.