VOICE ACROSS HISPANIC AMERICA: A TELEPHONE SPEECH CORPUS OF AMERICAN SPANISH

Yeshwant Muthusamy, Edward Holliman, Barbara Wheatley, Joseph Picone† and John Godfrey‡

Systems and Information Sciences Laboratory, Texas Instruments, Dallas, TX
† Dept. of Electrical Engineering, Mississippi State University, Starkville, MS
‡ Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA
yeshwant@csc.ti.com

ABSTRACT

As part of the **Polyphone** project, Texas Instruments is in the process of collecting and developing a corpus of telephone speech in American Spanish. The corpus, called Voice Across Hispanic America (VAHA), will attempt to provide balanced phonetic coverage of the language, in addition to containing widely used vocabulary items such as digits, letter strings, yes/no responses, proper names, and selected command words and phrases used in automated telephone service applications. The speakers are native speakers of Spanish living in the United States. The collection and development of the corpus is expected to be completed by June 1995. So far, we have collected about 500 speakers from various parts of the U.S.

In this paper, we describe the design issues in various aspects of the project, such as subject recruitment, corpus and prompt sheet design, the data acquisition system, and validation and transcription. We conclude with a brief statistical profile of the data collected.

1. INTRODUCTION

There is increasing interest in the speech research community in extending speech recognition research and system development beyond a single language. To transfer technology developed for one language into other languages, or to develop technology in multiple languages simultaneously, we need comparable speech corpora in a number of languages. The **Polyphone** project, an international cooperative effort initiated in 1992 through COCOSDA[1], addresses just this need. The goal of the Polyphone project is to make publicly available telephone-based speech corpora in as many of the world's major languages as possible. To date, corpora in American English (the Macrophone Corpus [1]) and Dutch [2] have been completed.

As part of the Polyphone project, Texas Instruments is in the process of collecting and developing such a corpus of telephone speech in American Spanish. The corpus, called Voice Across Hispanic America (VAHA), will attempt to provide balanced phonetic coverage of the language, in addition to containing widely used vocabulary items such as digits, letter strings, yes/no responses, proper names, and selected command words and phrases used in automated telephone service applications. The collection and development of the corpus is expected to be completed by June 1995.

2. WHY AMERICAN SPANISH?

The last decade has seen a steady increase in the Hispanic population in the United States. Spanish is now spoken by a significant portion of the population in several parts of the country. As a result, telephone companies and business organizations in the U.S. are beginning to recognize the need for automated telephone services in Spanish, such as spoken speed dialing, spoken number dialing and voice-messaging. Development of such services requires a large corpus of training data, collected over commercial telephone lines. The VAHA corpus will satisfy this important need.

3. SUBJECT RECRUITMENT

The subject population consists of men and women between the ages of 16 and 70 whose primary language is Spanish and who reside in the U.S. An effort is being made to collect equal numbers of subjects of both sexes, in four age categories and at four broad educational levels. The age categories are: 16 to 30, 31 to 45, 46 to 59 and 60 and above. The educational levels are: *did not complete high school, completed high school, had some college* and *completed college.*

[1] Coordinating Committee for Speech Databases and Assessment.

0-7803-2431-5/95 $4.00

The recruitment of the subjects is being handled by a market research firm familiar with the Hispanic population in the U.S. As the subjects are recruited, prompt sheets, with accompanying cover letters, are sent to them by post-office mail. Subjects are rewarded for their time with either $5.00 in cash or a free 20-minute phone call within the U.S.

Demographic information (age group, level of education, income range) is elicited from the subjects either during solicitation or in some cases (e.g., blanket mail-outs to known Hispanic groups or organizations) after they completed their call. Some of the subjects have not yet responded with the information, while others have refused to do so. Thus, the speaker statistics for this corpus will be incomplete.

4. CORPUS DESIGN

The corpus was designed to provide adequate coverage of vocabularies of interest (credit-card and phone numbers, dates, times, yes/no responses, proper names, alphanumeric strings), command words and phrases useful in automated telephone service applications, as well as phonetically rich data (phonetically balanced sentence sets). The numeric items were automatically generated. The command phrases and the phonetically balanced sentences merit further explanation.

4.1. Command Words and Phrases

A total of 79 command words and phrases were chosen from the following telephone service domains (examples in Spanish with English translations in parentheses):

- **spoken speed dialing** [e.g., llame a casa (call home), agregue el nombre (add name)]
- **spoken number dialing** [e.g., marque el numero (dial number), numero internacional (international number)]
- **voice-messaging** [e.g., guarde el mensaje (save message), escuche el mensaje (listen to message)]
- **telecommunication services** such as call waiting, call forwarding, call block, call trace, three-way calls, collect calls, directory assistance, etc.

4.2. Phonetically Balanced Sentences

The phonetically balanced sentences were selected from the Spanish portion of the UN Parallel Text Corpus (available from the Linguistic Data Consortium), and from Reuters and Associated Press Spanish news-wire text. These text corpora were manually checked to remove sentences that dealt with unpleasant subjects (e.g., murder, genocide, etc.), contained long, unwieldy or unfamiliar words, or were ungrammatical or incomplete. This selection process yielded a total of 53,314 "acceptable" sentences. The candidate sentences were then processed using an entropy balancing algorithm, similar to the one used in [3], to select 13,338 phonetically rich sentences.

5. PROMPT SHEET DESIGN AND GENERATION

The prompt sheet was designed to obtain samples of the complete range of vocabulary items described above from each subject. Each sheet consisted of 45 items: 36 read items and 9 elicited spontaneous speech responses. There was overlap between prompt sheets, but no two were identical.

Each subject receives a unique prompt sheet, identified by a 5-digit caller identification number (CIN). An on-line database of prompt sheets provides easy access to individual items during the validation and transcription process.

5.1. Read Speech

- 1 5-digit caller identification number
- 12 command words and phrases
- 12 numeric items
  - 4 telephone numbers
  - 2 credit card numbers
  - 1 quantity item (e.g., 156 kilogramos, 52 minutos, 31 yardas)
  - 1 dollar amount (e.g., $15.95)
  - 1 date
  - 1 list of 6 digits
  - 1 unsegmented 8-digit string
  - 1 unsegmented 8-character alphanumeric string (e.g., LQ14PZ6R)
- 1 spelled word
- 2 'name at agency' type phrases (e.g. Maria Lopez del Departmento del Trabajo), and
- 8 phonetically rich sentences

All the numeric iems were generated automatically, making sure that the distribution of digits and letters within and across prompt sheets was approximately uniform. The 'name at agency' type phrases were generated using an alphabetized list of names obtained from telephone directories.

The 79 command phrases were replicated 1,519 times and randomized to yield 120,000 command phrases (12

per prompt sheet for 10,000 sheets). Care was taken to ensure that each block of 12 phrases did not contain any repetitions. The set of 13,338 phonetically balanced sentences was replicated 6 times and randomized to provide a final pool of 80,028 sentences. This pool was the source of 8 sentences per prompt sheet.

5.2. Spontaneous Speech

- 4 questions eliciting yes/no responses (English translations):

  Are you ready?
  Are you calling from a push-button phone?
  Did you find this session difficult?
  Would you be willing to participate in a similar project in the future?

- 5 prompts and questions that elicit spontaneous speech responses (English translations):

  What is the time?
  Please say a familiar phone number.
  Please say a familiar name.
  What language or languages do you speak at home?
  How often do you travel?

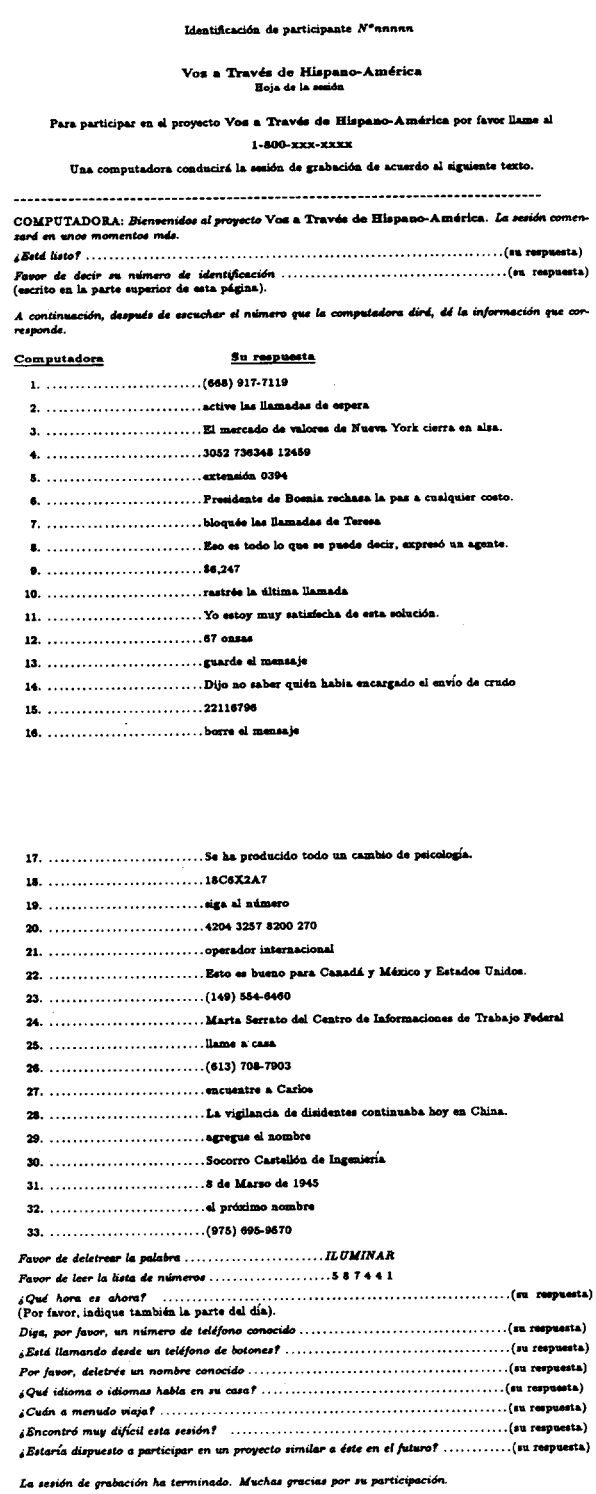
The spontaneous speech prompts were the same across all prompt sheets. The time item was elicited as a spontaneous response to avoid dealing with the myriad of ways in which time is said in Spanish. Figure 1 shows both sides of a sample prompt sheet. Each prompt sheet was printed double-sided on a single sheet of paper.

6. DATA ACQUISITION

The data collection hardware consists of the InterVoice RobotOperator, an IBM PS/2 machine running OS/2 with a proprietary digital interface. The RobotOperator is programmed to automatically monitor all of the 8 T1 (digital) phone lines connected to the toll-free 800 number that the subjects call. The machine answers each call, plays pre-recorded prompts (in Spanish) and records the subject responses. The recording is done at 8 kHz in μ-law format. The machine also keeps a log of all incomplete calls. At the end of each day, the data files are automatically moved to a disk on a Unix workstation for further processing.

7. VALIDATION AND TRANSCRIPTION

The calls are validated and transcribed by two trained transcribers who are educated native speakers of Spanish. The following two-pass validation approach has been adopted to process the calls.



Identificación de participante *N°nnnnn*

Vos a Través de Hispano-América
Hoja de la sesión

Para participar en el proyecto Vos a Través de Hispano-América por favor llame al
1-800-xxx-xxxx
Una computadora conducirá la sesión de grabación de acuerdo al siguiente texto.

COMPUTADORA: *Bienvenidos al proyecto* Vos a Través de Hispano-América. *La sesión comenzará en unos momentos más.*
¿Está listo? .....(su respuesta)
Favor de decir su número de identificación .....(su respuesta)
(escrito en la parte superior de esta página).

A continuación, después de escuchar el número que la computadora dirá, dé la información que corresponde.

Computadora	Su respuesta
1.	(668) 917-7119
2.	active las llamadas de espera
3.	El mercado de valores de Nueva York cierra en alza.
4.	3052 736348 12459
5.	extensión 0394
6.	Presidente de Bosnia rechaza la paz a cualquier costo.
7.	bloquée las llamadas de Teresa
8.	Eso es todo lo que se puede decir, expresó un agente.
9.	86,247
10.	rastrée la última llamada
11.	Yo estoy muy satisfecha de esta solución.
12.	67 onzas
13.	guarde el mensaje
14.	Dijo no saber quién había encargado el envío de crudo
15.	22116796
16.	borre el mensaje
17.	Se ha producido todo un cambio de psicología.
18.	18C6X2A7
19.	siga al número
20.	4204 3257 8200 270
21.	operador internacional
22.	Esto es bueno para Canadá y México y Estados Unidos.
23.	(149) 554-6460
24.	Marta Serrato del Centro de Informaciones de Trabajo Federal
25.	llame a casa
26.	(613) 708-7903
27.	encuentre a Carlos
28.	La vigilancia de disidentes continuaba hoy en China.
29.	agregue el nombre
30.	Socorro Castellón de Ingeniería
31.	8 de Marso de 1945
32.	el próximo nombre
33.	(975) 695-9570

Favor de deletrear la palabra .....*ILUMINAR*
Favor de leer la lista de números 5 8 7 4 4 1
¿Qué hora es ahora? .....(su respuesta)
(Por favor, indique también la parte del día).
Diga, por favor, un número de teléfono conocido .....(su respuesta)
¿Está llamando desde un teléfono de botones? .....(su respuesta)
Por favor, deletree un nombre conocido .....(su respuesta)
¿Qué idioma o idiomas habla en su casa? .....(su respuesta)
¿Cuán a menudo viaja? .....(su respuesta)
¿Encontró muy difícil esta sesión? .....(su respuesta)
¿Estaría dispuesto a participar en un proyecto similar a éste en el futuro? .....(su respuesta)

La sesión de grabación ha terminado. Muchas gracias por su participación.

Figure 1: Both sides of a sample prompt sheet.

Table 1: Distribution by state of origin

TX	*CA*	*IL*	*NM*	*PR*	*FL*	*NY*	CO CT NV OH PA
259	209	9	8	5	2	2	1

Table 2: Distribution by age group (R = *refused* and U = *unavailable*).

Age	*16-30*	*31-45*	*46-59*	*60+*	*R*	*U*
#Subj.	176	183	76	42	4	18

- **Preliminary Validation.** For each caller, the validators listen to just the file containing the CIN and note it down, along with their judgment of the caller's sex. An automatic determination of the number of responses provided by the caller is also made and the caller is either discarded (incomplete call) or set aside for the second validation stage. In the latter case, the CIN is used to automatically retrieve the corresponding prompt sheet and provide default transcriptions to all the read items of the call. The CINs are relayed back to the market research firm. This procedure provides them with valuable feedback about the call inflow and the information needed to make reminder solicitation calls, if necessary.
- **Full Validation and Transcription.** In this stage, the validators listen to all the utterances of each call, determine the caller's dialect, make judgments about the quality and content of the speech, note down all non-speech events (e.g. background noise, line noise) and provide orthographic transcriptions of the spontaneous speech responses.

8. CURRENT STATUS

To date, 536 complete calls (364 females and 172 males) have been collected, validated and transcribed from the first 2500 solicitations. This 21% yield is disappointing, though not entirely surprising, since the registration and recording protocols may easily evoke suspicion. We are taking steps to mitigate this problem, but cannot yet estimate the rate of return, nor, therefore, the final number of callers.

Tables 1, 2, 3 and 4 show the distribution of these calls (345 female and 154 male; demographic information for the other subjects is not available) by state of origin, age, education level and income range, respectively. The distribution across education levels and age groups is uniform, while the income range distribution favors the lower ranges.

Table 3: Distribution by education level (HS = *high school*, $COLL$ = *college*).

Edu.	*<HS*	*HS*	*>HS*	*COLL*	*R*	*U*
#Subj.	148	140	64	129	0	18

Table 4: Distribution by income range ('000 $)

Inc.	*<20*	*20-45*	*45-75*	*>75*	*R*	*U*
#Subj.	222	111	32	15	101	18

9. ACKNOWLEDGEMENTS

This work was supported by the Linguistic Data Consortium. Opinions expressed in this paper do not necessarily reflect the views of the Linguistic Data Consortium.

10. REFERENCES

[1] J. Bernstein, K. Taussig, and J. Godfrey. Macrophone: An American English telephone speech corpus for the Polyphone project. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 94*, Adelaide, South Australia, April 1994.

[2] M. Damhuis, T. Boogaart, C.in't Veld, M. Versteijlen, W. Schelvis, L. Bos, and L. Boves. Creation and analysis of the Dutch Polyphone corpus. In *Proceedings International Conference on Spoken Language Processing 94*, Yokohama, Japan, September 1994.

[3] T. Staples, J. Picone, and N. Arai. The Voice Across Japan database – the Japanese language contribution to Polyphone. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 94*, Adelaide, South Australia, April 1994.