Voice Across Hispanic America
Final Report

Yeshwant K. Muthusamy
Personal Systems Laboratory
Texas Instruments Incorporated
P.O. Box 655474, MS 238, Dallas, TX 75265.

February 26, 1996

1 Introduction

Voice Across Hispanic America (VAHA) was an effort to collect a large corpus of Spanish telephone speech from native speakers of American Spanish in various parts of the United States. We have attempted to make VAHA consistent (in design and content) with other large corpora like Voice Across America [4], Macrophone [1] and Dutch Polyphone [2].

The corpus contains 38,740 transcribed (read and spontaneous) utterances from 915 speakers. All data was collected in 8-bit ulaw format over T1 digital telephone lines.

2 Project Overview

The project was divided into the following parts: (i) corpus design, (ii) prompt sheet design, (iii) subject recruitment, (iv) data acquisition, validation and transcription, and (v) file preparation for shipment to LDC.

3 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.

3.1 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.

3.2 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 newswire 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.

4 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. A total of 10,000 prompt sheets were generated.

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

4.1 Read Speech

The prompt sheets contained the following items to be read:

- 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 Lpez del Departamento 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.

4.2 Spontaneous Speech

The prompt sheets also contained the following questions and requests:

- 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. Figures 1 and 2 show both sides of a sample prompt sheet. Each prompt sheet was printed double-sided on a single sheet of paper.

5 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.

A market research firm familiar with the Hispanic population in the U.S. contracted to recruit participants for the project. The original intention, as expressed in their statement of work, was for VAHA participants to come predominantly from the southwestern United States (Texas, New Mexico, Arizona, California, and surrounding areas), the contractor's area of operations. This would result in a fairly consistent variety of Spanish with its origins mostly in the northern "high plains" states of Mexico, although in recent years larger numbers of immigrants from the Central American countries of Guatemala, El Salvador and Nicaragua have settled there as well. It was thought that technology-oriented researchers would prefer less rather than more dialectal variation in such a corpus; this premise, which was arguable in any case, became moot when the marketing firm was unable to recruit more than a fraction of the number of callers promised. The principal reason cited for this failure was the Hispanic population's widespread suspicion of any effort to gather identifying information about individuals.

After several unsuccessful campaigns, this approach was abandoned, the contract cancelled, and subsequent recruiting was done by volunteers using email, computerized lists, and word of mouth, with no restriction as to the participants' place of residence in the United States. The largest single source of callers in this second phase was the LDC's solicitation list for the CALL HOME and CALL FRIEND projects. In this effort, speakers of Latin American Spanish were being offered payment to allow their phone conversations to be recorded for a government database. By simply offering an extra $5.00 to make a VAHA call, LDC was able to recruit over 200 calls from speakers of many different Latin American dialects residing in the U.S.

Due to the fact that the recruiting mechanics were controlled initially by a contractor, and later by the procedures already in place at LDC, there were several differences in the results which should be noted:

Identificación de participante *N°nnnnn*

Voz a Través de Hispano-América
Hoja de la sesión

Para participar en el proyecto **Voz 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* **Voz 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**

Figure 1: Side 1 of a sample prompt sheet.

1. ....(668) 917-7119
2. ....active las llamadas de espera
3. ....El mercado de valores de Nueva York cierra en alza.

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 Marzo 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, deletrée 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 2: Side 2 of a sample prompt sheet.

- As mentioned, the market research firm solicited subjects mainly from the southwestern U.S. (Texas, New Mexico, Arizona, Colorado, Southern California); LDC obtained subjects from all over the U.S.
- Both efforts solicited subjects with the following educational levels: *did not complete high school, completed high school, had some college* and *completed college.* However, almost all the subjects solicited by the market research firm had lower educational levels (high school or less) than those solicited by LDC.
- The market research firm did not ask for city of origin in their subject interviews, for fear of scaring away potential subjects; LDC did obtain this information from their subjects.
- The market firm used 3 age categories: *less than 45*, *45–59*, *60 and above*; LDC asked for, and usually obtained, specific ages.

As the subjects were recruited, prompt sheets, with accompanying cover letters, were sent to them by mail. Subjects were rewarded for their time with either $5.00 in cash or a free 20-minute phone call within the U.S.

As might be expected, a substantial number of participants did refuse to divulge their age, education or income information. Given the difficulty and cost of collection, we have included their recordings and any available information, even where incomplete.

6 Data Acquisition

The data collection hardware consisted of the InterVoice RobotOperator, an IBM PS/2 machine running OS/2 with a proprietary digital interface. The RobotOperator was 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 answered each call, played pre-recorded Spanish prompts, and recorded the subject responses. The recording was done at 8 kHz in μ-law format. The automatic utterance detector/endpointer was set to avoid truncating speech. As a result, some of the utterances on noisy lines or with excessive background noise need to be truncated by hand.

The machine also kept a log of all incomplete calls. At the end of each day, the data files were automatically moved to a disk on a Unix workstation for further processing.

7 Validation and Transcription

The calls were validated and transcribed by three trained transcribers who were educated native speakers of Chilean, Nicaraguan, and Puerto Rican Spanish, respectively. The Chilean speaker, who was the most experienced and the only one who worked throughout the project, performed all quality checking and arbitrated all questions raised by the other transcribers. TI professional staff were responsible for communicating and enforcing the transcription conventions, and provided expertise on questions of linguistics or engineering (e.g., sources of audio distortion) as needed.

The following two-pass validation approach was adopted to process the calls.

- **Preliminary Validation.** For each caller, the validators listened to just the file containing the spoken CIN and noted it down, along with their judgment of the caller's sex. An automatic determination of the number of responses provided by the caller was also made and the call was either set aside (incomplete call) or passed to the second validation stage. In the latter case, the CIN was used to automatically retrieve the corresponding prompt sheet and provide default transcriptions to all the read items of the call. The CINs were relayed back to the market research firm or the LDC. This procedure provided 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 listened to all the utterances of each call, made judgments about the quality and content of the speech, noted down all non-speech events (e.g. background noise, line noise) and provided orthographic transcriptions of the

spontaneous speech responses. Calls from non-native speakers were removed from the corpus. There were 27 such calls.

8 File Preparation

The speech files were converted from a TI-internal signal-object format to the NIST sphere format. All the transcription and validation information was written to the header of the files. A sample NIST file header is shown below.

```
sample_rate 8000
sample_coding ulaw
channel_count 1
sample_n_bytes 1
sample_sig_bits 8
line_id 1
sample_byte_format 1
signal_quality background_noise
speaker_rate normal
speaker_articulation normal
speaker_effort normal
database_name vaha_00
data_collection_system_name ivi_ros2_1
speaker_id 000058
speaker_mode read
speaker_accent spanish
signal_condition complete
speaker_quality normal
speaker_sex female
speaker_age adult
sample_count 21120
session_date Jul_27_94
prompting_text agregue el nombre
transcription agregue el nombre [background_noise]
```

Most of the "speaker_" fields and the "signal_condition" and "signal_quality" fields contain subjective judgments made by the transcribers. They are described in greater detail in the transcription conventions document included with the corpus.

Tables 1, 2, 3 and 4 show the distribution of the 915 calls (570 female and 345 male) by state of origin, age, education level and income range, respectively. The distributions by income range and age group appear to be skewed towards the lower end, while the distribution by education level favors the higher end, mainly due to the higher education levels of the speakers solicited by the LDC.

Table 1: Distribution by state of origin—top 10 states (U = *unknown*).

TX	*CA*	*PA*	*AR*	*FL*	*WI*	*NY*	*IL*	*MA*	*U*
265	205	90	29	23	20	17	15	13	153

Table 5 shows a sampling of the detailed speaker demographic information included with the corpus documentation. "UNK" in a character field or "0" in a numeric field indicate that the information was not divulged by individual participants, and *SWUS* indicates "Southwestern United States"; SWUS was assigned to all callers recruited by the marketing firm unless other information was received, so it must

Table 2: Distribution by age group

Age	*<45*	*45-59*	*60+*	*U*
#Subj.	580	101	47	187

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

Edu.	*<HS*	*HS*	*someCOLL*	*COLL*	*U*
#Subj.	150	184	95	309	177

be viewed as a tentative *a priori* classification. Finally, the *City of Origin* column contains the answer to the question "Where did you grow up?" which did not always include a city name, or sometimes had more than one city name.

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

Inc.	*<20*	*20-45*	*46-74*	*>75*	*U*
#Subj.	209	107	30	15	554

9 Summary

In all, there were 915 complete calls (570 females and 345 males) from a total of 6500 solicitations. The total solicitation yield was 14.1%. There were a total of 38,740 utterances collected, out of which 31,066 were read and 7,674 spontaneous.

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. Versterijlen, 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.

[4] B. Wheatley and J. Picone. Voice Across America: Toward robust speaker-independent speech recognition for telecommunications applications. *Digital Signal Processing*, 1(2):45–63, 1991.

Table 5: Sample from demographic information included with the corpus. **Age codes:** 0 = *unknown*, 1 = *less than 45*, 2 = *45-59*, 3 = *60 and above*; **Education codes:** 0 = *unknown*; 1 = *less than high school*, 2 = *high school*, 3 = *some college*, 4 = *college graduate*; **Income codes:** 0 = *unknown*, 1 = *less than $20K*, 2 = *20–45*, 3 = *46–74*, 4 = *$75K and above.*

Call#	*CIN*	*Sex*	*Age*	*State*	*Country of Origin*	*City of Origin*	*Education*	*Income*
000149	10269	F	1	CA	SWUS	UNK	4	0
000462	10273	F	2	CA	SWUS	UNK	4	1
000376	10301	M	2	CA	SWUS	UNK	1	1
000067	10312	F	0	CA	SWUS	UNK	2	0
000270	10325	F	2	CA	SWUS	UNK	1	0
000889	15529	F	2	MA	Uruguay	Montevideo	4	0
001098	15533	M	0	UNK	UNK	UNK	0	0
001002	15535	F	1	MD	Argentina	Buenos Aires	4	0
000962	15537	M	3	CA	Uruguay	Montevideo	4	0
001009	15558	M	1	AR	Costa Rica	Costa Rica	3	0
001003	15559	M	1	AR	Honduras	La Ceiba	2	0
000938	15560	F	1	AR	Guatemala	Guatemala	4	0
001044	15561	M	1	AZ	Argentina	Santiago de Estero	4	0
001011	15564	F	2	AZ	Dominican Republic	Santo Domingo	4	0
001025	15567	M	1	IL	Ecuador	Chicago	4	0
000937	15568	F	1	AR	Costa Rica	San Jose	4	0
000999	16117	M	2	PA	Mexico	Atizapan de Zaragoza and Mexico City	0	4
000977	16121	M	1	VA	Peru	Lima	4	0
000882	16122	F	1	MA	Peru	Lima	4	0
001346	16557	F	0	UNK	UNK	UNK	0	0