# Data Collection and Modeling For Speech Recognition

**Final Report**


**Abstract**

**The Hispanic-English Corpus (1.0) contains approximately 30 hours of conversational speech data from non-native speakers of English. Approximately 24 hours of the data are closely transcribed. For the purpose of collecting this data, Entropic designed and implemented a system infrastructure that allows for simultaneous recordings of wide-band and telephone-bandwidth speech. Funded by the US Government, this data collection effort yielded a conversational speech database designed to support speech recognition technology for telephone speech and to facilitate research on the characteristics of non-native English. This report describes the protocol Entropic developed for collecting and processing the data and the resulting Hispanic-English speech database.**


## 1.0 Introduction

Recent advances in automatic speech recognition technology suggest a strong correlation between recognizer performance on the one hand, and the quality and nature of training and testing data on the other. Especially in the domain of ASR research for telephone speech, there is a growing demand for large quantities of carefully transcribed conversational speech data. Moreover, none of the existing conversational speech corpora include much data suitable for developing ASR systems capable of handling non-native English.

The major reason for the current lack of suitable conversational data is production cost. Collection and processing such data is resource and labor intensive, requiring not only a costly hardware configuration, but also a skilled labor force for processing the data.

During the period between September 1996 and May 1998, Entropic Research Laboratory, Inc. designed and implemented a protocol for collecting spontaneous conversational speech data directly over the telephone network (via digital T1 telephone lines). In addition to creating the infrastructure for making simultaneous wide-band/narrow-band recordings, Entropic tested and selected suitable conversational topics, and developed a set of procedures for collecting and processing the data. Once the hardware configuration was completed and tested, these procedures were implemented to create a corpus of conversational non-native data that can be used in training and evaluating telephone speech recognition systems as well as in developing speech applications for computer-aided language Learning (CALL).

The resulting database comprises a corpus of approximately 30 hours (about 180000 words) of spontaneous, conversational speech data from 22 Hispanic speakers of English. The transcribed data are delivered along with this report on 17 CD-ROMS. The Hispanic-English database possesses a set of unique features that make it valuable for advanced speech recognition research on non-native English.

- Conversations are recorded simultaneously on four separate channels.
- Conversations are spontaneous and natural. Conversational topics are engaging and task--oriented in nature, covering a broad range of grammatical structures and pragmatic tasks.
- The collected data exhibits various levels of non-native English proficiency, which makes it ideal for modeling disfluencies and infelicities typical of language learners.
- Roughly 80% of the data have been closely transcribed so far. Transcribers used Entropic's *Annotator* for the transcription task, a software tool that facilitates detailed, time-aligned labeling of speech data. Special symbols identify mispronunciations, hesitations, and other characteristics of non-native English.
- In addition to the conversations, each of the 22 speakers recorded a total of 100 read utterances, divided equally between Spanish and English.
- Prior to participating in the data collection, all participants took an English proficiency test. (A breakdown of test scores is provided in Table 2 below.)

The following report provides a detailed description of the data collection protocol that was devised by Entropic to produce this database. Additional relevant materials are included in the Appendix to this document.

# 2.0 Database Overview

The Hispanic-English Speech corpus currently consists of a set of 17 CD-ROMS. Disks 1 through 5 contain conversational speech data. Disk 5 further includes the recordings of read speech from the first 3 speaker pairs. Since disk 6 through 16 were delivered incrementally (one speaker pair at a time), the 100 sentences are included in the double disk set for each session. Each CD-ROM provides a README. An earlier version of this report dated is included on Disk 1. A copy of this report is included on Disk 16.

## 2.1 Filename Conventions

All materials pertaining to a particular speaker are identified by a unique speaker ID, a string consisting of the speaker's first initial and the first two letters of the speaker's last name. This naming convention was adopted to preserve speaker anonymity but did lead to one duplicate (there are two speakers in the database with the id "ahe"). Demographic information about the speaker along with other recording information is contained in the header of the ESPS data files (conversations and read speech) The contents of the header can be viewed with the ESPS "psps" utility:

```
> psps -vD ahe1.ss009.sd

--Generic Part of Header--
Age: 33
Gender: F
Geographic origin: San Julian, Argentina
Microphone: Shure SM-10a
Name: A HE
Native language: Spanish
Prompt string: Las investigaciones llegaron a la verdad.
Prompt type: printed
Record program: sgrecord2
Recording location: erl ca
max_value: 0
record_freq: 16000
samples: 320000
start_time: 0
```

For a summary of speaker demographics, see Table 2 below.

### 2.1.1 Conversational Speech

Data files and corresponding label files for conversational data are stored in subdirectories consisting of a speaker-pair ID and a session number. The speaker-pair ID is a combination of two speaker IDs. The first three letters identify the speaker on Channel A; the last three letters identify the speaker on Channel B. An effort has been made to maintain filename consistency throughout the database. Speaker IDs and corresponding channels are also encoded in the ESPSA file header.

Data and label files are contained in numbered session subdirectories and are identified by the following filename extensions:

| | |
|---|---|
| **.wb.sd** | 2 channel wide-band archive in ESPS format |
| **.nb.sd** | 2 channel narrow-band archive, converted from mu-law to ESPS format |
| **.nb.lab,.wb.lab** | label file containing time stamped orthographic transcriptions narrow-band and wide-band archive files |

Note that the two files ending in **.wb.lab** and **.nb.lab** are identical. The transcriptions were done with Entropic's *Annotator*, which creates a separate label file for each archive. Reviewing the data in the *Annotator* display mode requires the existence of a separate file for each archive in the user's home directory. When copying these files from disk, we suggest linking them symbolically or collapsing them into one single label file. Unless the Annotator is used to review the data, it suffices to work with one copy of the label files.

### 2.1.2 Read speech

The Spanish and English sentences are included on Disk 5 and in subsequent disks (for subsequent speaker pairs) under a subdirectory called "100sentences." Waveforms and transcriptions are stored in numbered subdirectories for each speaker. In most cases Spanish and English Data directories are identified by the extension 1 and 2 respectively. How-

ever, one or two subjects started with the English prompts, which means that the numbering is not enirely consistent.

Recordings of English and Spanish sentence prompts are identified as follows:

**speakerid1.ss<digit>.sd** waveforms of Spanish utterances
**speakerid2.es<digit>.sd** waveforms of English utterances

The sentence IDs, **es<digit>** for English and **ss<digit>** for Spanish, refer to the corresponding indices used in the prompt files. Sentence prompts are identical for all speakers. They are contained in two separate ASCII files, **prompts.engl** and **prompts.span.** The written prompt for each recorded utterance is stored in the header of the corresponding data file along with demographic information about the speakers.

## 2.2 Distribution of Conversational Speech Data

Each recording session was divided into several sub-sessions in order to keep the length of data files within manageable limits (between 130 and 150 mb for wide-band archives) and to provide breaks for the talkers. Although an attempt was made to create sub-sessions of equal length (ideally five 36 minute sub-sessions per one 2 hour and 40 minute session), this was not always possible. Some sessions had to be concluded prematurely due to unexpected technical problems with the recording software. In most cases, such problems affected one channel of the wideband recordings and the session was edited to the length of the shortest wideband file. If the problem was caught right away, a new session could be started and the "lost time" made up. In some cases, the problem was not detected until after the end of the session and the session turned out to be cut short. Also, the first session is significantly shorter. On the other hand, all of the remaining sessions exceed 2h 40m so that the average length turned out to be 2h45m.

Table 1 shows a breakdown of transcribed recording sessions according to the number of sub-sessions, session length in minutes, word count per session, and conversational topics. Two entire sessions and part of session 3 remain untranscribed.

**Table 1**

| Filename | Length | #of words | topic(s) | label file | disk |
|---|---|---|---|---|---|
| mroahe1 | 38 min | 5179 | scruples | yes | 1 |
| mroahe2 | 38 min | 4503 | pict. seq. | yes | 1 |
| **Session 1** | **1h16min** | **9682** | | | |
| hfrero1 | 8 min | 1325 | scruples | yes | 2 |
| hfrero2 | 35 min | 4789 | story cmp. | yes | 2 |
| hfrero3 | 23 min | 3230 | Mars, pict. seq. | yes | 2 |
| hfrero4 | 42 min | 6268 | scruples | yes | 2 |
| hfrero5 | 35 min | 5209 | scruples | yes | 3 |
| hfrero6 | 30 min | 3845 | story cmp. | yes | 3 |

| Filename | Length | #of words | topic(s) | label file | disk |
|---|---|---|---|---|---|
| **Session 2** | **2h53min** | **24666** | | | |
| gbaaes1 | 34 min | 2760 | scruples | yes | 4 |
| gbaaes2 | 42 min | 4466 | Mars, pict. seq. | yes | 4 |
| gbaaes3 | 31 min | [~2500] | scruples | no | 4 |
| gbaaes4 | 36 min | [~3000] | story cmp. | no | 4 |
| gbaaes5 | 44 min | [~4000] | scruples | no | 5 |
| **Session 3** | **3h07min** | **~16500** | | | |
| lblelo1 | 42 | 2523 | Mars, story cmp. | yes | 6 |
| lblelo3 | 46 | 2796 | story cmp. | yes | 6 |
| lblelo4 | 43 | 2615 | story cmp. pict.seq. | yes | 7 |
| lblelo5 | 46 | 2789 | scruples | yes | 7 |
| **Session4** | **2h57m** | **22017** | | | |
| pramgo1 | 10 | 1692 | Mars | yes | 8 |
| pramgo2 | 47 | 7043 | story cmp. | yes | 8 |
| pramgo3 | 47 | 6443 | story cmp. | yes | 8 |
| pramgo4 | 44 | 6792 | scruples | yes | 9 |
| pramgo5 | 43 | 6193 | scuples | yes | 9 |
| **Session5** | **3h08m** | **28163** | | | |
| ealfto1 | 43 | 5728 | Mars, story cmp | yes | 10 |
| ealfto2 | 41 | 5001 | pict. seq | yes | 10 |
| ealfto3 | 45 | 6593 | scruples | yes | 11 |
| ealfto4 | 46 | 6713 | scruples | yes | 11 |
| **Session6** | **2h55m** | **24035** | | | |
| ahejhe1 | 41 | 6418 | mars, story cmp. | yes | 12 |
| ahejhe2 | 46 | 6486 | pict. seq. | yes | 12 |
| ahejhe3 | 17 | 2640 | scruples | yes | 13 |
| ahejhe4 | 49 | 6626 | scruples | yes | 13 |
| **Session7** | **2h33m** | **22170** | | | |
| rgoghe1 | 32 | 3239 | mars | yes | 14 |
| rgoghe2 | 52 | 4910 | story cmp. | yes | 14 |
| rgoghe3 | 55 | 4621 | story cmp, pict. seq. | yes | 15 |
| rgoghe4 | 16 | 1385 | pict. seq. | yes | 15 |
| rgoghe5 | 29 | 2583 | pict. seq. | yes | 15 |

| Filename | Length | #of words | topic(s) | label file | disk |
|---|---|---|---|---|---|
| **Session8** | **3h04m** | **16738** | | | |
| acunma1 | 43 | 5856 | Mars | yes | 16 |
| acunma2 | 45 | 5769 | story cmp | yes | 16 |
| acunma3 | 45 | 6052 | pict. seq. | yes | 17 |
| acunma4 | 28 | 3549 | scruples | yes | 17 |
| **Session9** | **2h50m** | **21226** | | | |
| **TOTAL** | **24h43m** | **185197** | | | |

Wordcounts in column three are calculated on the basis of the transcription files and refer only to words actually spoken (not including time markers and transcription symbols). Word counts in square brackets are estimates based on the length of the corresponding archive file. They are included in the total count. Note that the README file for disks 5 and 6 does have erroneous wordcounts. The numbers in the table above represent the accurate wordcount information.

For a detailed description the conversation topics see section 4.2.1. below.

# 3.0 Subject Recruitment

Participating subjects were paid volunteers who were recruited from the local Hispanic Community. All were adult native speakers of Spanish as spoken in South and Central America. The criterion for selection was a minimum of one year of residence in the US, as well as a basic ability to understand, speak, and read English.

Prior to participating in the data collection, subjects signed a written consent form in which they agreed to the publication of the data (Appendix). They were informed about the nature of the communicative tasks and the duration of the recording session (which lasted on average 5 hours for a 2 hour and 40 minute recording session). In addition, they were instructed by a Spanish consultant on how to use the recording interface, how to wear the microphone, and how to place the telephone handset correctly. Subjects were paid $100.00 for participating in the recording session.

As part of the recruiting process, the subjects' proficiency in English was tested. Entropic used a telephone-based, automated English proficiency test developed by Ordinate Corporation (1998). The test measures the test-taker's ability to comprehend and produce (or reproduce) spoken US English at a normal conversational speaking rate. Test responses were graded automatically by a computer-based system that assigns a combined score for lexical accuracy, fluency, pronunciation, and listening comprehension. Results are reported on a scale from 2 through 8, with most native speakers scoring above 7.5. The subject population is fairly evenly distributed over this scale with a slight bias toward higher scores (the median proficiency score is 5.9).

Table 2 provides a breakdown of subjects according to gender, geographical origin, and test scores.Table 2 provides a breakdown of subjects according to gender, geographical origin, and test scores.

| Spkr ID | G | C | S |
|---|---|---|---|
| aes | F | Mexico | 6.0 |
| ahe | F | Argentina | 5.0 |
| ero | F | Argentina | 7.8 |
| gba | F | Chile | 3.8 |
| hfr | M | Argentina | 6.9 |
| mro | F | Argentina | 6.6 |
| elo | M | Mexico | 6.4 |
| lbl | M | Nicaragua | 5.9 |
| fto | M | Peru | 5.1 |
| eas | M | Argentina | 7.2 |
| jhe | M | Cuba | 4.9 |
| ahe | M | Cuba | 4.3 |
| mgo | F | Argentina | 7.2 |
| pra | F | Argentina | 7.1 |
| rgo | M | Mexico | 5.9 |
| ghe | F | Mexico | 3.5 |
| bav | M | Nicaragua | 7.4 |
| ilo | M | ElSalvado | 3.0 |
| acu | F | Peru | 5.3 |
| nma | F | Mexico | 5.6 |
| rar | M | Nicaragua | 7.8 |
| kpa | F | Peru | 4.5 |

Table 1: Speaker Demographics: C- country of birth, G- gender, S- proficiency test score.

# 4.0 Data Collection and Processing Procedures

The Hispanic-English database covers two different types of speech: wide-band recordings of read speech and four channel recordings of spontaneous conversational speech. Since methods and procedures differ in each case with regard to material design, recording set-up, and data processing, we describe them separately for each data type.

## 4.1 Read Speech Corpus

The read speech corpus comprises a total of approximately 2200 read utterances (50 English and 50 Spanish utterances per speaker). The data was digitized at 16 bits per sample and a sampling rate of 16kHz.

### 4.1.1 Selection of Input Prompts

The prompts used for the recordings in English were selected from the TIMIT database (LDC, 1991). The Spanish sentence prompts represent a subset of the materials used in the Latino-40 database (LDC, 1995). The items were selected according to the following criteria:

- brevity (sentence length under ten words)
- absence of low-frequency lexical items
- relatively simple, straightforward syntax

Within each language, the prompts were arranged in an order of increasing length. The TIMIT selection includes two shibboleth sentences: "Don't ask me to carry an oily rag like that" and "She had your dark suit in greasy wash water all year." A complete listing of sentence prompts for both languages is included in the Appendix.

### 4.1.2 Recording Procedure for Read Speech

Each of the participating talkers began by recording 50 Spanish and 50 English sentences. The recordings were made on two Silicon Graphics Indy (SGI) work stations with native audio. Subjects interacted with the computer via a software interface that displayed a written sentence prompt and recorded the speech input. Subjects controlled the pace of the recordings and had the ability to check their input and redo a recording. Prior to the recording session, subjects were instructed in Spanish on how to use the recording interface and how to wear the microphone correctly. Recordings were made with a high-quality noise-canceling, head-mounted microphone (Shure SM10A). A Rane pre-amplifier was inserted between the microphone and the SGI Indy line-input jack. The recordings were digitized in 16-bit samples at 16 kHz and stored in ESPS format. The input gain level was calibrated once at the beginning of the recording session.The recordings were made during off-hours at Entropic's Menlo Park office so as to reduce extraneous noise and to secure a quiet, unintrusive recording environment. Recordings were made simultaneously, but each subject was sitting in a separate room to avoid any interference between the two recordings.

### 4.1.3 Data Verification

Two slightly different verification procedures were used for the English and Spanish data.

The Spanish data was verified by one reviewer who marked utterances that contained mispronunciations, omissions, or insertions of words not found in the written prompt. These utterances were subsequently eliminated from the data set.

In verifying the non-native English utterances, an attempt was made to distinguish between systematic mispronunciations due to accent on the one hand, and genuine reading errors that are of no systematic value for studying the properties of non-native English on the other hand. Examples of such errors include repetitions, self-corrections, stuttering, deletions, and insertions. Utterances containing the first type of errors were accepted as "good," those containing errors of the second type were marked as "bad."

The English utterances were verified by two reviewers independently. The reviewers were asked to flag any utterances that struck them as problematic (for whatever reasons). The set of defective English utterances was retained in a separate directory and is included in the database, so that reverification for different training purposes remains possible.

## 4.2 Conversational Speech

### 4.2.1 Material Design

In designing suitable materials for generating conversations for language learners with disparate proficiency levels, Entropic pursued a twofold goal: topics should engage the speakers in a collaborative, problem-solving activity while at the same time stimulating the production of a broad range of grammatical and lexical structures and pragmatic attitudes. Furthermore, communicative tasks should vary in difficulty such that a selection could be made that would match the linguistic skills of any given subject. After reviewing and testing a number of potentially suitable materials, three types of tasks were selected to be used in the database. The selected topics draw on tasks and exercises commonly used in foreign language instruction and by teachers of ESL [(Mackey, 1994) and can be classified into three categories:

**Task Type 1: Picture Sequencing**

**Description:** Each subject received half of a randomly shuffled set of cartoon drawings. Subjects were instructed that the pictures were part of a story and that they were to reconstruct the original narrative sequence with the help of their partner, who held the remaining drawings.

**Grammatical Structures elicited:** Negatives, yes/no questions wh-questions, present tense, SVO, -ing forms, nouns, pronouns, singular and plural forms.

**Pragmatic Task:** Descriptive

**Skill Level:** Basic, no reading knowledge of English required.

**Task Type 2: Story Completion**

**Description:** Subjects were given two identical copies of a set of drawings depicting isolated scenes from a larger narrative context. Contrary to the picture sequencing task, the scenes were entirely unrelated. Beneath each drawing, three written prompts invited the speakers to comment on the following questions: (1) what is going on here? (2) what happened before? and (3) what is going to happen next?

**Grammatical Structures elicited:** Questions, -ing forms, wh- questions, regular and irregular past tense, present and future tense, object pronouns.

**Pragmatic attitudes:** Narrative, descriptive

**Skill Level:** Basic to intermediate, some reading comprehension required

### Task Type 3: Conversational games

**Description:** Two conversational games were used. The first one was a commercially available game called "Scruples." A deck of randomly shuffled cards were distributed to the participants. Each card displayed a written question posing a kind of moral dilemma or conflictual situation. The problems tend to be posed in terms of an hypothetical situation: "Suppose you hit someone's parked car while backing out of your driveway. Do you leave a note for the owner of the car?" Subjects were instructed to alternate posing the questions and to negotiate an agreement on how to resolve the conflict.

The second game, called "Going to Mars," placed speakers in a hypothetical situation, in which they were called upon to agree on five professionals to take along on a space colonization mission. Subjects were given a list of ten professions (e. g., a teacher, a policeman, a priest, etc.) and were asked to come to an agreement with regard to their choices.

**Grammatical Structures elicited:** Subordinate conjunctions, hypothetical constructs, wh-questions, negations, all tenses.

**Pragmatic attitudes:** Argumentative

**Skill Level:** Intermediate to advanced, requires solid reading comprehension

Even though Scruples was clearly the most difficult and complex task, it turned out to be the most popular one among subjects. Most likely, the popularity of this conversational game was due to the engaging and provocative nature of the issues. Contrary to initial expectations, even speakers with relatively low proficiency scores were remarkably creative in handling this task. When they did not fully understand a question, they asked their partner for clarification of unknown lexical items. When both were unsure about the meaning of a question, they moved to the next one. As it turned out, compared to the two other tasks, Scruples generated the most lively and interesting conversations.

### 4.2.2 Hardware Infrastructure

The conversational data was simultaneously recorded on four channels. Two of the channels were collected directly over the telephone network with the remaining two channels recorded locally on two Silicon Graphics Indy workstations. The telephone speech was digitized using a Dialogic D/240SC-T1 card that ran on a DEC Alpha 1000 4/233. The T1 line was carrying an ISDN-PRI signal, providing 23 B channels. Two of these channels where used to place phone calls to telephones in two separate offices and to record the incoming speech of the two T1 channels into separate files. The wide-band channel for each subject was recorded locally on an SGI machine with the same hardware configuration that was employed in recording the read speech. The telephony data was recorded in raw mu-law format. The wide-band speech was digitized at a 16kHz sample rate with 16 bits per sample in ESPS format.

### 4.2.3 Recording Software

A top level script was written to start both the "sgrecord" processes on the local SGI machines and the telephony software on the DEC Alpha. The telephony software was programmed using Dialogic's API. The application placed phone calls to two separate telephone numbers and internally routed the incoming signals to the other channel such that both speakers were able to hear each other and converse with one another. Since wide-band recordings were made simultaneously on the workstations, subjects had to be physically present in the office during the recordings.

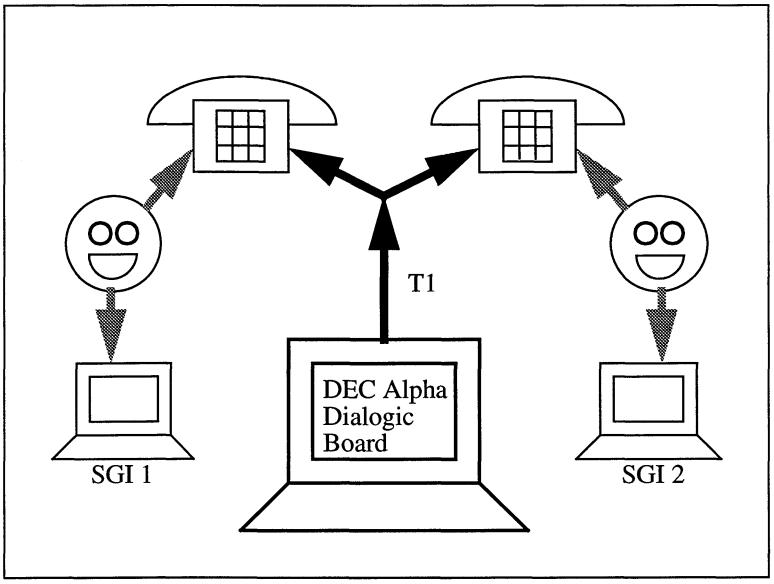


**FIGURE 1. Data Flow**

### 4.2.4 Technical Problems

After the hardware set-up was completed, it was tested extensively to eliminate potential sources of recording errors. As it turned out, the first few test recordings were substan-

tially distorted by echo, which appeared as an interfering cross-talk between the two telephone channels. The Automatic Gain Control, which is a default setting in the Dialogic recording interface, was identified as the major source of this echo. Moreover, the Automatic Gain Control setting made this echo appear much stronger then it really was. Disabling AGC significantly improved the signal quality of the recordings, even though some echo remained.

It would have been possible to eliminate the echo distortion by developing echo cancelling algorithms directly from the corresponding wide-band recordings and use the information from these recordings to cancel the echo in the telephone channel. However, it turned out to be simpler and much more effective to take advantage of the existing echo cancelling features offered by long distance carriers. When phone calls are rerouted to a long distance relay, the carrier takes care of all echo cancellation. To exploit this service, Entropic established two dedicated long-distance accounts with Sprint in Monterey, CA, which included call forwarding features. The Dialogic API was programmed to place phone calls to these numbers and then use the call-forwarding feature to reroute the calls back to Entropic's office.

Several long distance carriers and locations were tested to determine the best output quality. However, it turned out to be impossible to eliminate the echo completely; some cross-channel interference remained. It is most audible when both talkers speak simultaneously, and it manifests itself in the form of a certain "fuzziness" of the signal and a visible distortion of the waveform. In some of the material, the effects of this distortion are rather prominent. In order to deal with this problem, a special transcription symbol was devised to mark affected dialog turns [xta/xtb].

## 4.3 Data File Synchronization

The recording process yielded four channels of unequal length (one narrow-band telephone and one wide-band channel per speaker). Channels vary in length, because the software interface controlling the recordings had to initiate recording processes sequentially on three different machines. On average, approximately 12 to 14 seconds elapsed between the start of the local wide-band recording processes (SGI) and the telephone recording, due mainly to the long distance call-forwarding feature used in the recording set-up.

For the purpose of transcribing the data in such a way that one time-aligned annotation file covers both telephone and wide-band recordings, the channels had to be edited and synchronized. Editing turned out to be laborious, since the data files were too large to use the cut-and-paste features provided with *xwaves+*. Instead, editing had to be done via the command line by copying selected records from the input data file to a designated output file, which requires considerable resources in CPU time and disk space.

The synchronizing procedure is summarized below in the following steps:

1. Recordings were copied from three different machines to a 2 GB partition for further processing, and the unedited materials were backed up on CD-rom.

2. The ESPS utility "mu2esps" was used to convert the two telephone channels to ESPS format.
3. After the conversion, both the wide-band and the narrow-band waveforms for each speaker could be displayed with *xwaves+* for comparison. A rough synchronization was done by cutting off the first 10-13 seconds from the wide-band channel.The beginning time of the edited wide-band channel was then set to zero to allow for fine-tuning the synchronization.
4. To refine the synchronization, wide-band and telephone channels for each speaker were time aligned and zoomed in four times. A clear-cut word onset at the beginning of the telephone channel was used as a reference point for synchronization, and the distance from the beginning of the file to the reference point was measured in milliseconds. The procedure was repeated for the wide-band recording. The difference between the two numbers was cut from the onset of the wide-band recording, and the beginning time of the edited file was set again to zero.
5. The resulting synchronization was checked for accuracy by spot checking the alignment of acoustic events throughout the two files. In most cases, two passes at editing the wide-band recordings proved to be sufficient. Occasionally, however, a third pass at editing was necessary to improve the result. Synchronization was considered satisfactory when the discrepancy between wide-band and narrow-band channels was less than 10 ms.
6. After channels AA and BB were synchronized, the wide-band and narrow-band channels were joined into two dual-channel waveforms (AB) using the ESPS "mux" utility. The resulting output files were cut to the same length.

## 4.4 Transcription Procedures

### 4.4.1 Objectives

In order to transcribe the data, Entropic developed a protocol designed to meet the following objectives:

- provide an accurate, time-aligned labeling of acoustic events that meets the requirements of advanced speech recognition research
- provide a level of detail that does justice to the unique features of non-native conversational speech
- encourage readability of the transcriptions
- incorporate quality control measures that secure consistency across transcriptions

The conventions for transcribing the data were developed in collaboration with the Center for Language and Speech Processing at Johns Hopkins University. They draw on existing sources (SWITCHBOARD, CallHome) that were tailored to the specific nature of the material at hand.

### 4.4.2 General Principles

Transcriptions reproduce the content verbatim without attempting to edit or "correct" grammatical or other errors.

Wherever possible, transcriptions preserve the integrity of a dialog "turn" to facilitate modeling non-native disfluency over long intervals. Separate turns are defined by two criteria: (a) the thrust of the conversation is shifted from A to B, and (b) this shift marks the end of a meaningful segment of conversation. The beginning and the end of each turn is marked by a time stamp (+beg. seconds end seconds) that refers to both wide-band and narrow-band recordings. Contrary to common practice (e.g. , in SWITCHBOARD), dialog turns are not broken up simply because of an interruption by the second speaker. Instead, interruptions and/or simultaneous speech is transcribed sequentially. The exception is when simultaneous speech creates audible cross-channel interference in the telephone channel, in which case, the affected turn is marked by [xta] and [xtb] respectively.

Example:

```
+ 477.780562 478.539625 A:      Yeah.

+ 481.163375 489.150188 B:      I mean, it >>w<< would be different if it said,
                                hm, no, no, I wouldn't >>go<<. How about you?

+ 484.480250 514.909250 A:     [xtb] Sorry? Yeah, I, I think it takes away
                                much of the fun. I mean, if you go to a place,
                                and you, you plan to have a nice time, if you are
                                >>thinking<<, you know, you are eating at the
                                expense of people being exploited or taken advan-
                                tage, I mean, and you pay the bill, you say, you
                                know this is ((we want)) expensive but +it's+
                                cheap because these people is exploited then
                                takes @away@ some of, quite a bit of enjoyment of
                                having a nice meal, I think.

+ 491.113875 520.470312 B:     [xta] Right. Exploited is not good, yeah. Right,
                                right, @right, I agree@. {laugh} Yeah, it's not
                                right, anyway its' not right in the first place.
                                So...
```

A set of special symbols was developed to flag unusual speech characteristics, non-lexical items and channel specific events, such as the cross-channel interference mentioned above. For details, see the transcription manual included in the Appendix.

Transcriptions were generated with Entropic's *Annotator,* a software tool that allows for a convenient dual-channel display of the data and for precise time-stamping and labeling of acoustic events in a stream of sample data. In contrast to conventional transcription from tape recorders, the *Annotator*, displays the waveform, lets the transcriber execute a variety of customized play, view, and editing commands, and create the corresponding annotation in an Emacs buffer.

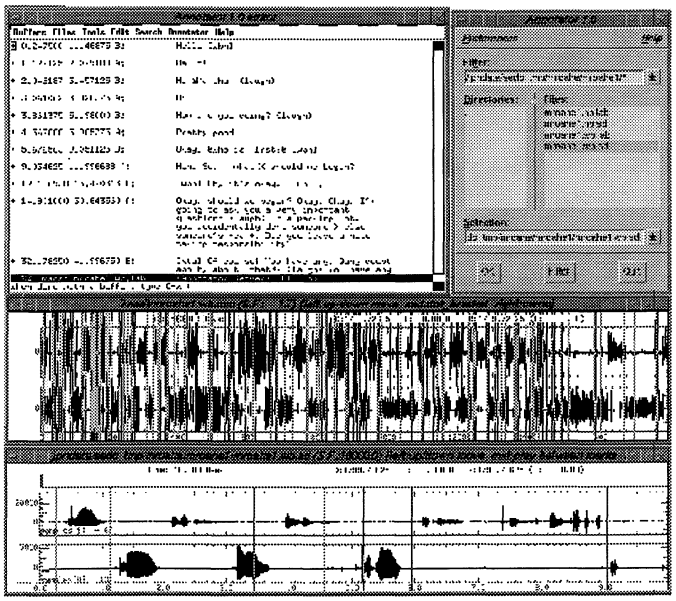

**FIGURE 2.**

The he transcriptions were done by two sets of transcribers (3 and 3) who were trained by the same staff member who developed the transcription conventions. Transcribers worked from a copy of the transcription conventions and an instruction sheet that suggested a breakdown of the task into separate steps to encourage systematic check ing for errors. On average, labeling the data required five passes through a given file.

1. Identifying turns and dropping/editing time stamps
2. First pass at transcription, wide-band channel only
3. Review of wide-band transcription, insertion of special symbols, resolution of ambiguities
4. Checking of label file against narrow-band channel, flagging of channel -specific acoustic events and turns affected by cross-channel interference
5. Final check for spelling errors, running of a script to catch formatting errors such as mismatched parentheses, etc.

Attempts to collapse steps 3 and 4 were unsuccessful, because the signal quality of the telephone channel generally was not sufficient to support fine-tuning the transcriptions. Finally, all transcriptions where checked by one reviewer to ensure quality control and consistency across the database.

## 5.0 Summary and Conclusion

The Hispanic-English Data Collection effort has shown that it is possible to create a high quality, fully transcribed, four channel conversational speech database by non-native speakers of English. However, producing this data turned out to be more costly in time, labor expenditures, and resources than initially anticipated. Recording, editing, and transcribing one 2h40m recording session takes on average 160 hours of skilled and semi-skilled labor (this figure does not include the time spent on material design and hardware setup). Two remaining speaker pairs (about 5h) and about 60% of the third speaker pair remains untranscribed. We anticipate completion of the database by August, 1998.

The Center for Language and Speech Processing at Johns Hopkins University is currently conducting preliminary recognition experiments to assess the difficulty of the recognition task that might be encountered in language learning exercises and thus gauge how effectively current ASR technology might be used in computer-aided language instruction (Byrne et al., 1998). As such, the Hispanic-English database should be of interest to researchers in large vocabulary conversational speech recognition as well as to researchers in CALL.

## 6.0 References


Byrne, W., Knodt, E., Khudanpur, S., Bernstein, J. (1998) Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English, S*TiLL - ESCA-Workshop on Speech Technology in Language Learning, Proceedings.*

LDC (1997) CallHome American-English Speech Corpus, Available from the Linguistic Data Consortium, Philadelphia, PA. LDC Catalog No.: LLDC97S41.

Ehsani, F. and E. Knodt. (1998) "Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New CALL Paradigm," *Advancing Technology Options in Language Learning, Proceedings.*

Entropic Research Laboratory, Inc. (1995) LATINO-40 Hispanic Speech Corpus, LDC Catalog No.: LDC95S28.

Godfrey, J., and E. Holliman, and J. McDaniel, (1992) "SWITCHBOARD: Telephone Speech corpus for research and development." *ICASSP.*

Mackey, Alison. (1994) *Using Communicative Tasks to Target Grammatical Structures: A Handbook of Tasks and Instructions for their Use.* Sydney Australia, 1994.

Ordinate Corporation. (1998) "The PhonePass Test." Menlo Park, CA.

LDC (1990) TIMIT. Acoustic-Phonetic Continuous Speech Corpus, NTIS PB91-505065.

## Appendix B: Material Design

### a. Sentence Prompts for Read Speech, English

es001 This was easy for us.
es002 It's illegal to postdate a check.
es003 Cut a small corner off each edge.
es004 We are open every Monday evening.
es005 Swing your arm as high as you can.
es006 Guess the question from the answer.
es007 He ate four extra eggs for breakfast.
es008 Coconut cream pie makes a nice dessert.
es009 Serve the coleslaw after I add the oil.
es010 Keep the thermometer under your tongue!
es011 Basketball can be an entertaining sport.
es012 Straw hats are out of fashion this year.
es013 Jane may earn more money by working hard.
es014 An official deadline cannot be postponed.
es015 Call an ambulance for medical assistance.
es016 Mosquitoes exist in warm, humid climates.
es017 Even I occasionally get the Monday blues!
es018 The news agency hired a great journalist.
es019 My instructions desperately need updating.
es020 A large household needs lots of appliances.
es021 Orange juice tastes funny after toothpaste.
es022 She didn't even give me a chance to refuse.
es023 Don't ask me to carry an oily rag like that.
es024 Critical equipment needs proper maintenance.
es025 Good service should be rewarded by big tips.
es026 Combine all the ingredients in a large bowl.
es027 Too much curiosity can get you into trouble.
es028 Are your grades higher or lower than Nancy's?
es029 Eat your raisins outdoors on the porch steps.
es030 It's hard to tell an original from a forgery.
es031 Who authorized the unlimited expense account?
es032 Drop five forms in the box before you go out.
es033 No, he didn't know of any handyman-carpenter.
es034 Barb's gold bracelet was a graduation present.
es035 Any contributions will be greatly appreciated.
es036 They remained lifelong friends and companions.
es037 Michael colored the bedroom wall with crayons.
es038 A doctor was in the ambulance with the patient.
es039 The best way to learn is to solve extra problems.
es040 In the long run, it pays to buy quality clothing.
es041 Chocolate and roses never fail as a romantic gift.
es042 She had your dark suit in greasy wash water all year.
es043 Last year's gas shortage caused steep price increases.
es044 December and January are nice months to spend in Miami.
es045 Draw every outer line first, then fill in the interior.
es046 Gregory and Tom chose to watch cartoons in the afternoon.
es047 Military personnel are expected to obey government orders.
es048 Kindergarten children decorate their classrooms for all holidays.
es049 Scientific progress comes from the development of new techniques.
es050 She encouraged her children to make their own Halloween costumes.

## b. Sentence Prompts for Read Speech, Spanish

ss001 Sus problemas empeoraron cada vez más.
ss002 La gente no sabe lo que está manejando.
ss003 Pero no parece ser de causas naturales.
ss004 Estoy satisfecho por el momento, agregó.
ss005 Ahora sólo soy un actor, como los demás.
ss006 La violencia, sin embargo, ha continuado.
ss007 Es asombroso, dijo, sacudiendo la cabeza.
ss008 Estoy muy preocupado acerca de los niños.
ss009 Las investigaciones llegaron a la verdad.
ss010 Espero que mi optimismo no sea exagerado.
ss011 Esta idea tiene un atractivo superficial.
ss012 Estoy listo para cualquier cosa, explicó.
ss013 La gente en realidad no sabe qué va pasar.
ss014 Esa es la forma en que he crecido, agregó.
ss015 El mundo ha cambiado dramáticamente, dijo.
ss016 Es la historia de un siglo, nuestro siglo.
ss017 El número exacto es imposible de calcular.
ss018 Las esperanzas de paz no mueren fácilmente.
ss019 No creo que haya nada especial en la rueda.
ss020 Nosotros no podemos ser tan ingenuos, dijo.
ss021 No estoy convencido que se sientan cómodos.
ss022 Nada es imposible en nuestro mundo, indicó.
ss023 Ese es esencialmente el problema hasta hoy.
ss024 Por cierto, deploramos todas esas acciones.
ss025 Todo lo que instalaron funcionó debidamente.
ss026 Nosotros vemos las cosas en forma diferente.
ss027 No hay intención de celebrar reunión alguna.
ss028 No me parece que la situación haya cambiado.
ss029 La respuesta al proyecto ha sido abrumadora.
ss030 Este acuerdo no cumple todo lo que queríamos.
ss031 Es evidente que esa situación es inaceptable.
ss032 No podemos evitarlo. Lo llevamos en la sangre.
ss033 El presente informe está basado en los hechos.
ss034 Hemos concluído así nuestro programa para hoy.
ss035 No les adjudica un papel positivo ni negativo.
ss036 No hay que considerar sino los propios hechos.
ss037 El punto de principio me parece ser importante.
ss038 Su estilo de vida también recibe gran atención.
ss039 Hasta el momento no tomó ninguna determinación.
ss040 La defensa de los derechos humanos fue constante.
ss041 Tratemos de ver lo anterior un poco más de cerca.
ss042 Hay que incorporar a las mujeres a esos programas.
ss043 Pienso que el intercambio de ideas fue provechoso.
ss044 No es un problema personal, sino que de principios.
ss045 Además, se están preparando las siguientes medidas.
ss046 El informe correspondiente será publicado en breve.
ss047 Sin embargo, esa propuesta no recibió apoyo alguno.
ss048 Sin embargo, hay que crear las condiciones para ello.
ss049 No lo he pensado, ni permito que me lo sugieran.
ss050 Pero no considero que sea una posición definitiva.

## c. Conversational Materials: Subject Instructions

### Picture Sequencing

TELL A STORY

You will be given an envelope with 3 or more cartoon pictures. These pictures are part of a story. You don't know what the story is because you have only half of the pictures. Your partner has the other half. Work with your partner to put together the whole story. Here is how you do it:

1. Take the first envelope. Check whether you both have the right envelope. If your envelope has A1 written on it, your partner's envelope should say B1.
2. Empty the envelope and look at your pictures.
3. Tell your partner what you have.
4. Ask your partner what he or she has. Work together to find the beginning and end of the story.
5. Fill in the remaining pictures.
6. When you are done with one story, put the pictures back into the envelope. Take the next envelope.

### Story Completion

TELL A STORY

INSTRUCTIONS:

On each page you will see one or more pictures. Look at each page and talk about it with your partner. Look at each page separately. The individual pages do NOT hang together to form a story.

1. Ask your partner to describe the picture.
2. Ask your partner what happened before.
3. Ask your partner what will happen next.
4. Switch roles--tell your version of the story
5. Come to an agreement.
5. Go to the next page.

1. What is going on here?
2. What happened before?
3. What will happen next?

## Conversational Games: Scruples

### HOW TO PLAY SCRUPLES

Scruples is a game about difficult moral situations. You and your partner will be given a stack of playing cards. Each card displays a question. You are asked to make a decision in a difficult situation. The goal of the game is to discuss the situation with your partner and to come to an agreement.

- Pick a card and ask your partner the question written on the card.
- Listen to what your partner would do in that situation.
- Ask your partner WHY s/he would act in this way.
- Would you do the same thing? What would you do differently? Explain why.
- Try to come to come to an agreement.
- Go to the next cared and switch roles.

Note: Some questions contain English words you may not understand. Ask your partner what the word means. If both of you do not know the word or if the question is not clear, pick another card.

## Going to Mars

### GOING TO MARS

INSTRUCTIONS:

You will be sent to Mars to start a colony. You can only take 5 other people with you. Pick five people from the list below. Choose carefully--you will be with these people for the rest of your life.

Discuss your choice with your partner. Come to an agreement.

doctor
teacher
police officer
priest
carpenter
scientist
farmer
musician
lawyer
artist
cook
politician
who else?

## Appendix C: Transcription Conventions

Instructions for Transcribing the Hispanic-English Data

Step 1

Dropping the time stamps: Go through wide-band recording with the Annotator. Mark segments identifying turns.(For a definition of a "turn" see the transcription conventions.) Drop time stamps into waveform. Leave about 300 ms of silence at the beginning and end of turn. (Note: visible and audible breaths at the beginning of speech are part of the waveform). Import time stamps into emacs buffer. If file contains automated time stamps, edit if necessary.

Step 2

Do a first pass at transcription, drop special symbols where appropriate. Mark ambiguities you want to come back to with your own special symbol. This is the time to check for potential errors in the "turn" segmentation. Correct if necessary.

Step 3

Review the transcription, listening to the "good" wide-band recordings, Correct errors, resolve all ambiguities. Zoom out twice to visualize larger conversational units. Use C-p or menu item to replay marked paragraphs. If turns are very long, use mouse to play shorter segments. Now is the time to listen more closely to non-linguistic acoustic events or unusual productions. Mark accordingly. Also, resolve your problem cases. If you can't decide, mark as ((unintelligible)).

Step 4

Mark telephone channel specific acoustic events: Copy the <file.wb.lab> to <file.nb.lab> so you can review the telephone channel together with the wide-band transcription. Mark turns distorted by simultaneous speech with symbol [xta] or [xtb].If you detect any special acoustic events not audible on wide band, mark turn with [nbn]. (note: all microphone fumbling noises are wide-band only!)

Step 5

Run the spell checker through the label file. Use query-replace for possible [symbol} errors (= mismatching parens). After finishing this section do "unmark all paragraphs" followed by "mark waveform with whole buffer." This updates an finalized all waveform marks to reflect the time stamps of the transcription.

Step 6

Copy file.nb.lab back to file.wb.lab, so both are identical.

**Transcription conventions used in Hispanic-English Data Collection**

I. General principles

1. Transcriptions follow speech verbatim without attempting to correct errors.

2. Transcription follow standard usage in capitalization, hyphenation and punctuation.

3. All abbreviations are written out. EXAMPLE: <Mr.> => <Mister>

4. Only standard forms are used to transcribe contractions.
EXAMPLE: <gonna> => <going to> but <they aren't> or <she's>

3. Acronyms pronounced like a word are written in all caps
EXAMPLE: AIDS NASA

Acronyms pronounced like the individual letters are written in all caps with spaces be between the letters:
EXAMPLE: C I A H I V C E O

4. All numbers are written out:
EXAMPLE: twenty-two nineteen-ninety-five

5. Standard spelling is used for interjections and hesitations:
EXAMPLE: uh-huh mm-hm uh uhm hm oh okay jeez

6. Proper name and place names are marked by an ampersand:
EXAMPLE: &Mary &Jones &Arizona &Harper's

II. Special Conventions used in close transcriptions

1. Speakers are identified by A: and B:

2. Transcriptions preserve the integrity of a "turn." The beginning and the end of a turn is marked by a time stamp. (+beg. seconds - end seconds) that corresponds to the data file.

Separate turns are defined by two criteria: a) the thrust of the conversation switches from A to B, and b) a clearly marked end of a meaningful segment.

EXAMPLE:

+ 3.737625 4.777250 A: Tell me what you have.
+ 4.529687 4.843250 B: Let me see, I have three pictures.
+ 4.923977 5.483931 A: Okay. I think I only have two.

3. Turns will not be split up, simply because the other speaker responds with short interjections and or approval. When A is interrupted by B but continues to speak, the content of B's interruption will be transcribed following the end of A's turn. Simultaneous speaking of A and B is not marked in the text. It can be inferred from the time stamps that indicate the beginning and end of interruptions or simultaneous speech.

The exception is when simultaneous speech creates echo effects or distortions in the telephone channel. In this case, the interrupted turn is marked by [xta] and [xtb] respectively.

If A's turn is very long, and B only comments sporadically, B's utterances may be treated not as one, but as separate turns.

EXAMPLE:

+ 25.747375 35.268812 A: Then I have, well, I'll tell you all (())
that I have, then I have &Snoopy
that's about to open the
umbrella, uh, and there is a
big cloud with rain coming out.

+ 28.823842 29.24567 B: Okay.

+ 32.892345 33.92363 B: yeah.

Note: do not mark or transcribe isolated occurrences of non-lexical items, interjections, or noises.

Special symbols:

Special symbols mark non-speech events, mispronunciations, foreign words, etc. In the listing below, "text" represents speech data to be marked.

| | |
|---|---|
| {text} | Noise made by talker: {laugh} {cough} {sneeze} |
| [text] | Noise not made by the talker: [background noise] [telephone] |
| [wbn] | Noise occurring only in the wide band channel |
| [nbn] | Noise occurring only in the narrow band channel. |
| [xta], [xtb] | Channel interference resulting from simultaneous speech, on telephone channel only. |
| [[text]] | Comment used to describe unusual characteristics of immediately preceding or following speech: [[previous word lengthened]] |
| ((text)) | Unintelligible; text is best guess at transcription: ((what's that?)) |
| (()) | Unintelligible; can't even guess text |
| <text> | Speaker uses a language other than English |
| >>text<< | Word fragment, write out complete word if you can hear it |
| @text@ | Unusual production; e.g., word turns into laughter |
| +text+ | Mispronounced word (spell it in usual orthography) |

Note: this symbol is *not* used for systematic mispronunciation due to foreign accent.

//text// Text spoken aside, e.g., to the operator.

$word$ Nonexistent word invention, for example: shaked for shook

How to handle stutters: If stutters consists of clearly distinguishable word fragments, use the >><< fragment sign for each part of the stutter: e.g:

Wa-wa-wa what? is transcribed as >>what<< >>what<< >>what<< what?

If stutter cannot be transcribed in this way (because it is too murky) use {stutter} or (( )).