*Presented at the 1998 ESCA Conference on Speech Technology in Language Learning. Marholmen, Sweden*

# Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English

*William Byrne* †, *Eva Knodt* ‡, *Sanjeev Khudanpur* †, *Jared Bernstein* *

Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD USA †

Entropic Research Laboratory, Inc. Menlo Park, CA USA ‡

Ordinate Corporation, Menlo Park, CA USA *

byrne@jhu.edu, knodt@entropic.com, khudanpur@jhu.edu, jared@ordinate.com

## Abstract

We describe the protocol used for collecting a corpus of conversational English speech from non-native speakers at several levels of proficiency, and report the results of preliminary automatic speech recognition (ASR) experiments on this corpus using HTK-based ASR systems. The speech corpus contains both read and conversational speech recorded simultaneously on wide-band and telephone channels, and has detailed time aligned transcriptions. The immediate goal of the ASR experiments is to assess the difficulty of the ASR problem in language learning exercises and thus to gauge how current ASR technology may be used in conversational computer assisted language learning (CALL) systems. The long-term goal of this research, of which the data collection and experiments are a first step, is to incorporate ASR into computer-based conversational language instruction systems.

## 1 Introduction

While automatic speech recognition (ASR) has matured so that large vocabulary speaker dependent dictation is commercially feasible, non-native accents, disfluent speech, and conversational dialogue pose substantial difficulties for ASR systems. To support speech recognition research on conversational and non-native speech, we implemented a protocol for collecting spontaneous conversations by Hispanic speakers of English. The resulting corpus possesses a set of unique features that make it valuable for advanced speech recognition research on the linguistic characteristics of language learners:

· Conversations are spontaneous and goal-oriented, covering a broad range of grammatical structures and pragmatic tasks.

· Detailed, time-aligned transcriptions identify mispronunciation, hesitations, and other characteristics of non-native English.

· Recordings are made simultaneously on four channels (wide-band and telephone speech).

· English and Spanish read text is available for all subjects.

Initial experiments suggest that the speech in this database is significantly more difficult to recognize than conversations between native English speakers. We expected that the constrained, task-directed nature of the conversational topics would simplify the language modeling task and compensate for the poor acoustic modeling of non-native speech. However, this appears not to be the case. The vocabulary coverage of the material by existing native English conversational corpora is good except for occurrences of proper names, task specific terms and lapses into Spanish; however, the effectiveness of language models built on these native English corpora as measured by perplexity is poor. This appears to be due in part to the language learners' difficulties with English, and also to the presence of a fair amount of free conversation unrelated to the given tasks.

## 2 Database Overview

The Hispanic-English database covers two different types of speech: wide-band recordings of read speech and four channel, simultaneous, wide-band and telephone channel recordings of spontaneous conversational speech.

### 2.1 Speaker Demographics

The Hispanic-English speech corpus comprises approximately 20 hours of closely transcribed, spontaneous, conversational speech data from 11 speaker pairs, plus an additional 100 Spanish and English sentences read by each speaker. Participating subjects were paid and were recruited from the Hispanic Community local to Palo Alto, California. All were adult native speakers of Spanish as spoken in South and Central America. The criteria for selection was a minimum of one year of residence in the US, and a basic ability to understand, speak, and read English.

As part of the recruiting process, the subjects' proficiency in English was tested. We used a telephone-based, automated English proficiency test developed by Ordinate Corporation [1]. The test measures the ability of the test taker to comprehend and produce (or reproduce) spoken US English at a normal conversational speaking rate. Table 1 provides a breakdown of subjects according to gender, geographical origin, and test scores. Figure 2.1 plots

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

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

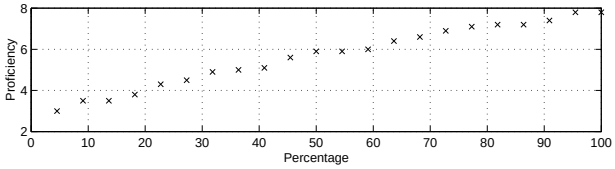



*Figure 1: Cumulative distribution of speaker proficiency scores. Median score is 5.9.*

the cumulative distribution of speaker scores. Scores in the range of 6.5-8.0 indicate native-like skills in speaking and understanding English. Scores between 3.5 and 6.5 indicate increasing skills, with the lower range indicating that speakers are able to understand slow, simple conversational material with occasional clarification. The subject population is fairly evenly distributed over this range of skills, with a slight bias towards higher scoring speakers.

## 2.2 Read Speech Corpus

The read speech corpus comprises a total of approximately 2200 read utterances (50 English and 50 Spanish utterances per speaker). English sentence prompts were selected from the TIMIT database [2]. The Spanish sentence prompts represent a subset of the materials used in the LATINO-40 database [3]. Recordings were made with a high-quality noise-canceling, head-mounted microphone (Shure SM10A) in a quiet laboratory environment. The data was digitized at 16 bits per sample and a sampling rate of 16kHz. Two slightly different verification procedures were used for the English and Spanish data. The Spanish data was verified by a reviewer who flagged utterances containing mispronunciations, omissions, or insertions of words not found in the written prompt. In verifying the non-native English utterances, an attempt was made to distinguish between systematic mispronunciation due to accent, and genuine reading errors. Utterances containing errors of the second type were flagged but retained in the database.

## 2.3 Conversational Speech

### *2.3.1 Recording Infrastructure*

The conversational data was simultaneously recorded on four channels. The telephone speech was digitized using a Dialogic D/240SC-T1 card running on a DEC Alpha 1000 4/233 interfaced to a T1 line carrying an ISDN-PRI signal providing 23 B channels. Two of these channels were used to place phone calls to each subject in two separate offices and to record the incoming speech of the two T1 channels into separate files. The wide-band recording of each subject was sampled locally by an SGI workstation. A single program controlled the recording processes on the local SGI workstations and the telephony software on the DEC Alpha. The telephony software itself was programmed using the Dialogic API.

The application placed phone calls to two separate telephone numbers and internally cross-routed the incoming signals so that both speakers were able to converse with one another. Since wide-band recordings were made simultaneously on the workstations, the subjects had to be physically present in the office during the recordings. They were, however, in separate rooms and were not able to talk to each other except over the telephone. While the recording apparatus was occasionally intrusive, having subjects carry out their tasks by telephone is a fairly natural way to force them to communicate by speaking and to prevent them from using visual, or other non-verbal cues.

### *2.3.2 Material Design*

In designing the conversational materials, we 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 so as to accommodate the proficiency of any given speaker. The selected topics draw on tasks and exercises commonly used in foreign language instruction and TESOL [4] and fall into three categories:

*Task Type 1: Picture Sequencing*

*Description:* Subjects received half of a randomly shuffled set of cartoon drawings. They were asked 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.

*Task Type 2: Story Completion*

*Description:* Subjects were given two identical copies of a set of drawings depicting unrelated scenes from a larger narrative context. Subjects were asked 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 card game called "Scruples." Players are asked to negotiate an agreement on how to resolve a moral dilemma or conflict. The problems displayed on the cards are posed terms of an hypothetical situation: "Suppose that such and such is the case, how would you resolve the situation?" In the second game, subjects were given a list of ten professions (e.g., a teacher, a police man, a priest etc.) and were asked to agree on five professionals to take along on a space colonization mission.

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

*Pragmatic attitudes:* Argumentative.

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

The "Scruples" task turned out to be most popular one among subjects, due, most likely, to the engaging and provocative nature of the issues. Even speakers with relatively low proficiency scores were remarkably creative in handling this task.

#### *2.3.3 Transcription Procedures*

The conventions for transcribing the data draw on existing sources (SWITCHBOARD and Call Home [5]) that were tailored to the specific nature of the material at hand. Transcription procedures were designed to meet the following objectives:

· Provide an accurate, time-aligned labeling of acoustic events.

· 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.

Transcriptions were generated with the Entropic 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. The wide-band data was transcribed by a group of specially trained transcribers who reviewed each other's work to ensure consistency across transcriptions. The telephone bandwidth speech was then reviewed to ensure that it was consistent with the transcriptions derived from the broad-band speech. All transcriptions were spot-checked by one reviewer for accuracy.

Time alignments between the conversations and transcriptions were found by marking the start-time and end-time of conversation turns. Since complete sentences are somewhat rare in conversational speech, identification of these turns is somewhat subjective. The transcribers were instructed to mark the turns so that they were "linguistically meaningful" [6]. This differs from the frequently chosen approach that identifies a turn boundary as any significant region of silence.

## 3 Recognition Experiments

The immediate goal of this data collection is to determine how well current ASR systems transcribe conversational exercises of the sort carried out in second language instruction. A more ambitious goal is the detailed modeling of non-native conversational speech. At the time of this writing the data collection and transcription is still in progress. We report here on pilot experiments intended to gauge the difficulty of the problem.

Transcribed conversations are currently available between the speaker pairs in Table 2. This data is fairly small, so 5 test/training partitions were used. In each partition, utterances from one speaker pair served as the test set and utterances from the remaining conversations were used for training.

The transcriptions of the five speaker-pairs contain 4137 utterances with 107,162 words. Not counting word fragments, there are 3,335 unique words. By comparison, a collection of SWITCHBOARD conversation transcriptions used for language modeling contains approximately 2.6 million words in roughly 240,000 utterances. All but 355 of the words in these new conversations were among the 22,000 most frequent words in SWITCHBOARD. Many of the new words were proper names, unusual words specific to the exercises, and novel words invented by the transcribers to describe unusual pronunciations.

To form a 5K word, closed, test set vocabulary, the 3,335 words in the test set were augmented by the remaining most frequent words in the SWITCHBOARD vocabulary. Bigrams were built for each of the speaker pairs by building a bigram on the held-out speaker-pair transcriptions and interpolating it with a SWITCHBOARD bigram built on the 5K vocabulary. Characteristics of the speaker pair language models are given in Table 2. These bigram models are intended only as a serviceable first attempt at language modeling for this task; developing more sophisticated models is a focus of current effort.

For the initial experiments with this data we used an HTK-based(see [7] and the references within) telephone bandwidth, SWITCHBOARD conversational recognizer. The system has cross-word state-clustered triphones with 7461 states and 12-mixture Gaussian observation distributions based on PLP-Cepstral acoustic features. On a

*Table 2: Interpolated speaker-pair language models: training and test set sizes and test set perplexities.*

| Speaker Pairs | Utterances / Words | | Test Perp. |
|---|---|---|---|
| | Test | Train | |
| ero+hfr | 1240 / 27185 | 2897 / 79977 | 108.4 |
| gba+aes | 683 / 10737 | 3454 / 96425 | 113.7 |
| lbl+elo | 914 / 28529 | 3223 / 78633 | 123.5 |
| mro+ahe | 580 / 10909 | 3557 / 96253 | 140.7 |
| pra+mgo | 720 / 29802 | 3417 / 77360 | 105.4 |

*Table 3: Recognition results: number of test set words and sentences and the Word Error Rate for each speaker.*

| Spkr | #Snt | # Wrd | WER | Spkr | #Snt | # Wrd | WER |
|---|---|---|---|---|---|---|---|
| ero | 143 | 1379 | 66.0 | hfr | 129 | 1075 | 64.4 |
| gba | 45 | 226 | 70.8 | aes | 37 | 128 | 60.2 |
| lbl | 65 | 488 | 96.1 | elo | 74 | 549 | 96.4 |
| mro | 91 | 838 | 66.6 | ahe | 73 | 326 | 71.5 |
| pra | 30 | 265 | 90.2 | mgo | 22 | 101 | 89.1 |
| Sum/Avg | 709 | 5375 | 73.6 | | | | |

SWITCHBOARD test set these models yield approximately 42% Word Error Rate (WER); with improved language models and speaker adaptation, 36% WER is possible [8].

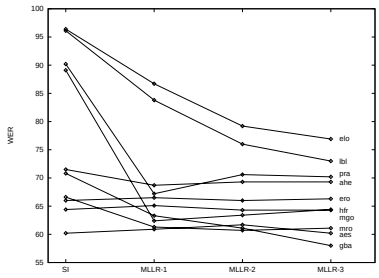
Another possibility would have been to decode the wide-band recordings using a read-speech recognizer. Since we do not have a large corpus of wide-band conversational speech, we have begun by evaluating the telephone speech first. Application of wide-band, read-text acoustic models to this task needs to be addressed.

Pilot recognition results of "speech-only" utterances (no laughter or significant channel noise) are given in Table 3. The first conversation between each speaker-pair is held out for supervised adaptatation.

While these results are discouraging, particularly given the small test set vocabulary size, we expect speaker adaptation to give significant improvement. The effect of supervised adaptation using Maximum Likelihood Linear Regression is given in Figure 2. For the worst speakers, several iterations of MLLR yields substantial improvement. We note anecdotally that using speech from one speaker pair did not help in adaptation for other speaker pairs. Speaker adaptation rather than speaker independent task adapted modeling may prove most effective. Collection of data for supervised adaptation is problematic for these subjects, however. Apart from the difficulty of using read text to adapt conversational acoustic models, many of the subjects are not skilled enough in English to read aloud fluently.

## 4 Conclusion

A new corpus of conversational English intended to capture the speech of native Spanish speakers for use in acoustic and language modeling for CALL has been described and some preliminary recognition results reported. While this speech appears significantly more difficult to recognize than native conversational English, we expect performance on this task to benefit from progress in speaker adaptation in general and in the modeling of non-native conversational speech in particular.




| Spkr | # Utts |
|---|---|
| ero | 19 |
| hfr | 18 |
| gba | 56 |
| aes | 62 |
| lbl | 10 |
| elo | 10 |
| mro | 38 |
| ahe | 31 |
| pra | 7 |
| mgo | 6 |

Overall Results

| System | SI | MLLR-1 | MLLR-2 | MLLR-3 |
|---|---|---|---|---|
| WER | 73.6 | 68.9 | 67.2 | 66.6 |

*Figure 2: Supervised speaker adaptation: overall results; per-speaker performance vs. number of MLLR iterations; and the number of adaptation utterances.*

**Acknowledgments**

We thank Entropic Cambridge Research Laboratory, Cambridge, UK, for software used at the 1997 LVCSR workshop and A. Stolcke for use of the SRI LM Toolkit.

## References

[1] Ordinate Corporation. "The PhonePass Test", January 1998, Menlo Park, CA.

[2] W. Fisher, V. Zue, J. Bernstein, and D. Pallet, "An Acoustic-Phonetic Data Base," *J. Acoust. Soc. Am.* **81**, Suppl. 1,1987.

[3] Entropic Research Laboratory, Inc. LATINO-40 Hispanic Speech Corpus, available from the Linguistic Data Consortium (LDC), Philadelphia PA.

[4] A. Mackey. Using Communicative Tasks to Target Grammatical Structures: A Handbook of Tasks and Instructions for their Use. Sydney Australia, 1994.

[5] J. Godfrey and E. Holliman and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," ICASSP 1992. Both SWITCHBOARD and the CallHome American-English Speech Corpus are available from the LDC.

[6] D. Jurafsky, *et.al.*, "Automatic Detection of Discourse Structure for Speech Recognition and Understanding," 1997 IEEE Workshop on Speech Recognition and Understanding, Santa Barbara CA.

[7] S. Young, "A Review of Large-Vocabulary Continuous Speech Recognition," IEEE Signal Processing Magazine, Sept. 1996.

[8] F. Jelinek, "1997 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Reports," CLSP, Johns Hopkins University, Baltimore MD.