COLLECTION AND DETAILED TRANSCRIPTION OF A SPEECH DATABASE FOR DEVELOPMENT OF LANGUAGE LEARNING TECHNOLOGIES

Harry Bratt *Leo Neumeyer* *Elizabeth Shriberg* *Horacio Franco*

SRI International, Menlo Park, CA, USA

ABSTRACT

We describe the methodologies for collecting and annotating a Latin-American Spanish speech database. The database includes recordings by native and nonnative speakers. The nonnative recordings are annotated with ratings of pronunciation quality and detailed phonetic transcriptions. We use the annotated database to investigate rater reliability, the effect of each phone on overall perceived nonnativeness, and the frequency of specific pronunciation errors.

1. INTRODUCTION

In this paper we describe the methodologies for collecting and annotating a Latin-American Spanish speech database. The database includes recordings by native and nonnative speakers. A panel of listeners rated the pronunciation quality of the nonnative data, and a group of expert phoneticians phonetically transcribed a subset of the nonnative data. The database was intended for use in the development of hidden-Markov model (HMM) based speech technologies for language learning [4], including robust speech recognition of nonnatives, automatic pronunciation scoring [4, 2, 1], and detection of mispronunciations [3, 5, 7].

To develop these technologies, reliable human ratings at the utterance level, as well as more detailed phone-level pronunciation information, are needed to calibrate and validate the system. The biggest challenge involved collecting detailed phone-level information for approximately 200,000 phones. We also provide an analysis for the reliability and usability of the phone-level data.

2. SPEECH DATABASE

We collected a total of 38,254 utterances from 127 native speakers, and 43,460 utterances from 206 nonnative speakers.

Subjects All native subjects were chosen from the same dialect background to as much an extent as possible. We targeted Latin American dialects only, focusing on Mexico, Colombia, Venezuela, Ecuador, Peru, and Bolivia, and avoiding, in particular, Argentina, Chile, Uruguay, and all Caribbean countries, which typically have more marked regional dialects. Other specifications such as education level and time in the United States were used to help standardize dialect.

All nonnatives were native American English speakers who had studied some Spanish locally or abroad. The levels of proficiency varied a great deal, and an attempt was made to balance the speakers for proficiency. This was done by assigning an initial "nativeness" rating of 1 through 5 to each nonnative who answered our recruitment ad, and collecting an equal number of each category for each gender. These initial ratings were assigned by native Spanish speakers based on the reading of three Spanish sentences, usually over the phone.

Data Collection The speech data was digitally recorded on a Sparc 5 workstation, using native audio, at 16 kHz and 16-bit linear PCM on two channels. The Sennheiser HMD-410 was used for all of the primary channel recordings. Eleven different secondary microphones were used for recording the secondary channel. Recordings were done in a relatively quiet office environment.

All speech data collected is read speech. Most of the prompts were taken from Spanish newspaper data, available through the Linguistic Data Consortium (LDC) and further filtered for various criteria such as length, spelling, and nonstandard characters. The resulting prompt set contained 94,000 sentences and had a vocabulary size of 36,567 unique words.

Training Data (Natives) For training, we collected speech from 102 native speakers: 51 male and 51 female. Most of the prompts, 300 sentences each, were drawn randomly without replacement from the newspaper pool described above. In addition, the speakers were asked to read digit strings, the words "si" and "no," and prompted to produce, in isolation, various "mouth noises," such as breaths and coughs, which we expected a future system would encounter in spontaneous speech. The amounts of each type of prompt are presented in the "Train" column in Table 1.

Development Data Twenty native speakers were collected for development. The "Development Native" column in Table 1 shows the amount of each type of prompt presented to these speakers. To help test our ability to detect mouth noises in the middle of speech, subjects were prompted to insert the same mouth noises collected for training into specific places within newspaper sentences ("News with Disruptions").

Another type of prompt introduced for the development test set is "Common Sentences." These were a set of 40

Table 1: Prompts per Speaker

Prompt Type	Train	Development	
		Native	Non-nat
Mouth Noises	6		
Ten-Digit Strings	10	10	10
Newspaper Sentences	300	150	75-150
Si/No Words	6	6	6
Isolated Words		100	50
Common Sentences		40	40
News w/Disruptions		20	
Short Common Sents			39[1]

sentences, drawn from the full pool of 94,000 newspaper sentences so as to maximize the number of occurrences of each pronunciation problem in a list. The list of potential pronunciation problems was made by a linguist and a Spanish language instructor, and was intended to include phones in different contexts that are known to be difficult for native American English speakers to pronounce. Examples are the diphthong "[eu]", /r/ after [l] [n] and [s] (which should be trilled), and [p] [t] and [k] in any context (which nonnatives may aspirate). Because the algorithm for finding sentences maximizes the number of problem areas, these common sentences are particularly long and have some uncommon vocabulary items.

The last column in Table 1 summarizes the prompts read by the 206 nonnative speakers. The new type of prompt added here is "Short Common Sentences." These are grammatically and lexically simple sentences created by a linguist and Spanish language instructor to have the same high number of problem phone target areas as Common Sentences. They therefore are neither as long nor nearly as difficult to read as Common Sentences.

3. UTTERANCE-LEVEL RATINGS

The utterance-level ratings we collected were judgments by native Spanish speakers (from the same dialect regions described in Section 2) on an ordinal scale of 1 to 5, corresponding to perception of nativeness. The raters were also asked to reject utterances that were truncated, did not match what was written, or had other problems not due to pronunciation. The entire set of nonnative speech data was rated by at least one rater.

Previous utterance-level ratings for French [4] had been collected from expert language teachers trained on grading overall pronunciation quality. Our approach for the current effort was to find native speakers with no necessary language-related expertise and select the best-correlated, through a pilot study. Of the eleven raters in the pilot, we chose the five best-correlated. The raters were calibrated to each other by presenting them with a small set of data to rate. The ratings were then discussed within the group, trying to get all to agree with a majority vote. This process was iterated several times, with different sets of data for the raters to converge on their ratings.

A common pool of data, 4,116 utterances, was set aside to be rated by all raters. This pool was balanced by sentence type and speaker. A subset of that pool, 820 utterances, was presented twice to each rater, to allow us to determine intra-rater reliability. The order of the stimuli was randomized separately for each rater.

Table 2: Intra- and Inter-Rater Correlation

Language	Expert Raters?	Intra-rater		Inter-rater	
		r	N	r	N
Spanish	no	0.79	554	0.78	2,787
French	yes	0.76	$\simeq 350$	0.76	$\simeq 350$

The remaining set of the 43,460 nonnative utterances was randomly divided among transcribers, balancing for sentence type and speaker. The rating task took about 50 hours per rater, spread over two to three weeks. To maintain the raters' calibration, they were allowed to play from a labeled list of example utterances (whose labels had been agreed upon by all raters) before each rating session.

The final data showed that the nonexperts in the current study had similar levels of both intra- and inter-rater correlation to those of the experts of the previous study. The sentence-level correlations are shown in Table 2. The intra-rater correlation assesses the consistency of repeated judgments of the same utterance by the same rater. The inter-rater correlation assesses the consistency of judgments across raters by correlating a rater's scores with the average of the other raters' scores.

4. PHONE TRANSCRIPTION

Gathering phone-level data is one of the most challenging problems for pronunciation scoring, yet such data is crucial for training a system that can give detailed feedback on specific phone-level pronunciation problems. Our goal was to phonetically transcribe a total of 3,573 utterances from the 206 nonnative speakers (all of the nonrejected common pool from the utterance-level raters), representing a total of approximately 200,000 phones.

The first step in acquiring phonetic transcriptions is to define the transcription conventions, which crucially involves defining the appropriate level of phonetic detail to be transcribed. Choosing too narrow a phonetic transcription would take far too much time for the amount of data we need to transcribe, but too broad a transcription would not give enough necessary detail to pinpoint pronunciation problems. One factor that made this task more tractable was that the native language of all the nonnative speakers was the same (American English). Therefore, we could expect to observe a relatively small set of common pronunciation problems. Also, we were interested only in nonnative phones; phones that the transcribers perceived as natively produced did not need to be described in any detail.

Given these issues, our approach was to define two sets of phones plus a set of diacritics. The first set of phones consists of all the native phones[2] in the targeted dialect of Spanish. The second set of consists of phones of American English, such as some reduced vowels and the labiodental fricative [v], which we expected to see carry over into nonnative pronunciations of Spanish. The diacritics were allowed to modify appropriate native phones. The transcribers were instructed that using a diacritic on a

[2]Essentially all the phonemes, with the addition of only five allophonic variants: [β], [ð], [ɣ], [z], and [ŋ].

phone implied that the phone was not perceived as native, and the diacritic explained in the way in which it was nonnative. Diacritics included aspiration for the voiceless stops, gliding for the nonlow vowels, and length (i.e., nonnatively long). A catch-all diacritic, "*", was included to represent a sound that was perceived as a nonnative rendition of a phone but for which no more specific method of indicating its nonnativeness was available.

In this way, we reduced the transcription problem to a simpler one in terms of cognitive effort for the transcriber, and ease of information entry, while still encoding the most important piece of information in all the transcriptions—the judgment of the nativeness of any given phone.

4.1. Transcribers / Transcription Tool

We recruited four native Spanish-speaking phoneticians to provide the detailed phonetic transcriptions. They used a Java-based transcription tool, enabling them to work offsite.

One dilemma in designing the transcription task involves whether or not the transcriber will see the "canonical" transcription (i.e., the dictionary or "correct" transcription) of the utterance. A transcriber who is shown the canonical transcription may sometimes be influenced toward using the canonical phones and thus fail to transcribe a nonnative phone. This would result in some nonnative phones being transcribed as native. On the other hand, a transcriber who is not shown the canonical transcription may make more simple mistakes, resulting in native phones being transcribed as nonnative, deleted, and so forth.

We attempted to overcome this dilemma by not showing the canonical transcription, yet trying to reduce the introduction of mistakes. To accomplish this we added a "double-check" feature to the transcription tool. We make the canonical transcription known to the tool but never show it directly to the transcriber. Instead, when the transcriber finishes each utterance, the tool compares the entered phone string with the canonical string, and highlights the locations where they differ. This displays visually where the transcriber is claiming nonnativeness occurred. The transcribers are instructed to double-check those regions to verify that they were intentional, and not the result of an accidental transcription error.

5. DATA AND ANALYSIS

From the 3,573 utterances to be phonetically transcribed, a common pool was constructed by choosing one Newspaper Sentence from each of the 206 speakers. The remaining 3,367 utterances were randomly divided among transcribers, balancing for sentence type and speaker. The common pool was mixed into each transcriber's data and each batch was randomly shuffled.

One of the main uses of the phone-level transcriptions is to train automatic systems to detect mispronunciations by nonnatives [3, 5, 7]. Some of the algorithms we planned to develop would be phone-specific—focusing on one phone, or a set of related phones, at a time. That is, given an "expected" (i.e., canonical) phone, known from the prompt text, we need to see what the speaker actually uttered. The speaker may have produced that phone natively, produced a nonnative version of that phone, or a native or nonnative version of a different phone, or deleted that phone altogether. We determine which of these possibilities happens for each expected phone by applying a dynamic programming alignment of the canonical phone string with the transcriber's phone string. The resulting information is used as the basis for the following analyses.

5.1. "Best" Phones: Three Criteria

The analysis done so far on the phone transcriptions is focused on providing information to aid in our goal of phone-specific automatic mispronunciation detection. In particular, we want to know which phones are most worth developing specific algorithms for. We lay out and examine three criteria that together can answer this question.

Reliably Transcribed (by humans) The first criterion measures whether any given phone is reliably transcribed by the human raters. If the four raters were unable to agree on whether a certain phone, say [p], was native- or nonnative-sounding, for example, then it would not be worth trying to have a machine match the nativeness judgments for [p].

We used the 206 common sentences to make this judgment, and used the kappa coefficient statistic [6] to determine how reliably the transcribers agree on the transcription for each of the 28 native phones. On twelve of the phones, all four transcribers showed at least a moderate level of agreement (using $K \geq 0.40$ to mean "moderate" agreement).

Many phones in which we were interested, however, such as the voiceless stops, [l], and the trilled [r], did not show moderate agreement among the transcribers. We therefore looked at two ways of getting stronger agreement by sacrificing some information. The first method is to throw away the data from one rater, for a given phone (one could throw away different raters' data for different phones). For five phones, including the three voiceless stops, there existed some set of three raters who did agree sufficiently with each other. This method is an option as long as three quarters of the originally collected data is enough for the phone of interest.

The second method we used for achieving a higher level of rater agreement is to collapse all of the different categories that the raters assigned to a given nonnative phone into one. This raised kappa significantly for four phones: [ð̃], [ɣ], [m], and [ɾ].

"Shibboleth" phone A shibboleth sound is one that "gives away" the nonnativeness of a speaker. This second criterion, then, measures how well a speaker's ability to pronounce a certain phone predicts the overall nativeness of that speaker. We use this to tell us, in effect, how important it is for a speaker to fix the pronunciation of a given phone. For example, if [p]'s tend to be reliably transcribable (criterion 1), yet speakers who produce [p] consistently nonnatively don't tend to be rated as poor speakers, and those who produce [p] well don't tend to be rated as good speakers, then it may not be worth trying to have a speaker improve the pronunciation of [p].

To assess this, we correlate the utterance-level scores (by averaging the ratings of the five nonphonetician raters explained in Section 3) with a score derived from the phone-level transcriptions of the phoneticians. We derive this latter score by taking a ratio of the number of nonnative versions of a phone in an utterance to the total number of times that phone should have occurred. Thus, if a sentence has five [p]'s and the speaker deleted one, nonnatively aspirated another, but pronounced the remaining three natively, that utterance will receive a score of 2/5, or 0.4. Because there are often few occurrences of any given phone in a single utterance, we obtain a more robust score by looking at the speaker level. That is, we correlate the average of all utterance-level scores for a speaker with the average of all transcription-derived scores for that speaker, over all 206 speakers. These correlations are not expected to be extremely high, since the transcription-derived scores are based on only one phone, but we can look at the relative correlations to see which phones correlate more than others to the overall nativeness of a speaker. The correlations and the percentile rankings for each phone are presented under the "Shibboleth" columns of Table 3.

Frequently Nonnative The final criterion considers how common it is to mispronounce a given phone. The motivation for including this criterion is that for phones passing the first two criteria it should be more useful to focus on those most commonly mispronounced.

The final column in Table 3 shows the percent of time that a phone gets labeled as other than a native version of itself by the transcribers.

5.2. Results

The approximants ([β], [ð], and [ɣ]) appear to be the most reliable class of phones to transcribe, tend to be the best predictors of overall nonnativeness, and tend to be among the most frequently mispronounced. The tap ([ɾ]) is the best shibboleth, and ranks high on the other two criteria as well. Some phones that we had expected to be useful, though—such as the voiceless stops, most of the vowels, [l] and [r], turned out not to have consistent enough transcriptions across all four transcribers, although eliminating one transcriber does help in most of these cases.

6. SUMMARY

We introduced a Latin-American Spanish database intended for the development of language learning technologies. We evaluated the inter- and intra-rater reliability of pronunciation ratings and the consistency of detailed phonetic transcriptions. We also introduced a measure for evaluating the effect of phone specific errors on the overall perceived quality of pronunciation. Based on this study, we learned what phones are good candidates for use in an automatic mispronunciation detection system.

7. ACKNOWLEDGMENTS

We gratefully acknowledge support from the U.S. Government under the Technology Reinvestment Program (TRP). The views expressed here do not necessarily reflect those of the Government.

Table 3: Results of Three Criteria

phone	no. cats	Reliable? K	Reliable? σ	Shibboleth? r_{spkr}	Shibboleth? %ile	Non-nat?
p	5	0.36	0.03	-0.36	61	38%
t	12	0.34	0.03	-0.44	89	34%
k	9	0.32	0.02	-0.34	50	47%
b	4	0.90	0.08	-0.05	7	42%
β	9	0.70	0.02	-0.43	86	73%
ð	10	0.55	0.03	-0.56	93	68%
ɣ	6	0.51	0.08	-0.39	75	73%
s	9	0.57	0.07	-0.11	18	6%
z	5	0.35	0.10	-0.02	4	84%
m	8	0.76	0.05	-0.09	14	17%
n	8	0.15	0.07	-0.12	21	6%
ŋ	4	0.46	0.06	-0.18	36	33%
l	8	0.22	0.04	-0.38	68	28%
ɾ	11	0.36	0.02	-0.59	96	42%
r	8	0.29	0.03	-0.39	71	77%
w	11	0.43	0.11	-0.21	39	40%
y	14	0.39	0.05	-0.32	46	18%
a	20	0.26	0.03	-0.35	54	17%
e	19	0.18	0.02	-0.41	79	24%
i	14	0.41	0.05	-0.41	82	20%
o	13	0.23	0.03	-0.35	54	20%
u	13	0.14	0.06	-0.28	43	18%

8. REFERENCES

1. C. Cucchiarini and L. Boves. Automatic assessment of foreign speakers' pronunciation of dutch. In *Proc. of EUROSPEECH 97*, pages 713–716, Rhodes, 1997.
2. H. Franco, L. Neumeyer, Y. Kim, and O. Ronen. Automatic pronunciation scoring for language instruction. In *Proc. Intl. Conf. on Acoust., Speech and Signal Processing*, pages 1471–1474, Munich, 1997.
3. Y. Kim, H. Franco, and L. Neumeyer. Automatic pronunciation scoring of specific phone segments for language instruction. In *Proc. of EUROSPEECH 97*, pages 649–652, Rhodes, 1997.
4. L. Neumeyer, H. Franco, M. Weintraub, and P. Price. Automatic text-independent pronunciation scoring of foreign language student speech. In *Proceedings of ICSLP '96*, pages 1457–1460, Philadelphia, Pennsylvania, 1996.
5. O. Ronen, L. Neumeyer, and H. Franco. Automatic detection of mispronunciation for language instruction. In *Proc. of EUROSPEECH 97*, pages 645–648, Rhodes, 1997.
6. Sidney Siegel and N. John Castellan, Jr. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, second edition, 1988.
7. S. Witt and S. Young. Language learning based on nonnative speech recognition. In *Proc. of EUROSPEECH 97*, pages 633–636, Rhodes, 1997.