Running head: CONSTRUCTION OF A RATED L2 SPEECH CORPUS

Construction of a Rated Speech Corpus of L2 Learners' Spontaneous Speech

Su-Youn Yoon, Lisa Pierce, Amanda Huensch, Eric Juul, Samantha Perkins,

Richard Sproat, Mark Hasegawa-Johnson

University of Illinois at Urbana-Champaign

abstract

This work reports on the construction of a rated database of spontaneous speech produced by second language (L2) learners of English. Spontaneous speech was collected from 28 L2 speakers representing six language backgrounds and five different proficiency levels. Speech was elicited using formats similar to that of the TOEFL iBT and the SPEAK (Speaking Proficiency English Assessment Kit) test. A total of 182 minutes of spontaneous speech were collected, segmented and assessed by two phonetically trained, experienced ESL instructors. The raters assigned a general fluency score and phone accuracy score with additional detailed comments on pronunciation errors. This database was designed with several applications in mind: the development of computer aided pronunciation and fluency training: automatic assessment of fluency and pronunciation; as a tool for researchers working in automatic speech recognition and for linguists, more generally. This database will be released to the public in the near future.

Key-Words: rated speech corpus, L2, automated scoring

Introduction

This study reports on the construction of a rated, spontaneous speech database of second language (L2) learners of English. The purpose of such a rated speech corpus is to aid in the development of automatic speech fluency assessment and computer aided pronunciation training (CAPT). The rated speech database will be used for the training and evaluation of such systems. It is generally acknowledged that a rated speech corpus is necessary for the development of such tools and many such efforts are reported in the relevant literature. For example, Witt (1998), Kim, Franco and Neumeyer (1997) and Bratt, Neumeyer, Shriberg and Franco (1998) collected non-native speakers' read speech and the accuracy of each phone was scored by trained raters. However, both databases were constructed from *read* speech. As such, it is impossible to analyze the nature of spontaneous speech. Spontaneous speech, for both L1 and L2 speakers, is complex in nature. It is characterized by pauses, filled pauses, hesitations, increased assimilation both within and across word boundaries, environmentally determined alternations as well as lenition and fortition phenomena predictable from higher level prosodic structures.

Recently, the Center for Spoken Language Understanding (CSLU) released the Foreign Accented English database. This database contains 4,925 spontaneous speech samples in English spoken by non-native speakers from 22 different native languages. Each speech file includes about 20 seconds of self-introduction. Three native speakers rated the accentedness of the sound files using a 4-point scale with 0 indicating "no accent" and 4 indicating a "very strong accent." Clearly such a database is a valuable resource for those researchers and scholars developing automated assessment systems for overall speech fluency. However, this database does not include an accuracy score for each phone, which would be useful for the research related to L2 learners' pronunciation in spontaneous speech such as acquisition of L2 phoneme and its actual use.

The database reported on here is constructed from spontaneous speech produced by L2-English learners. It was designed specifically for training and evaluating fluency and pronunciation in the context of spontaneous speech. The speech samples were recorded using an elicitation format similar to those used in the TOEFL iBT and the SPEAK test – both of which are fluency assessment tools. The database includes a general fluency score – again based on the TOEFL assessment rubric and a phone accuracy score. All scoring was done by raters who are both experienced ESL teachers and linguistically trained phoneticians. The database includes L2 speakers from five different language backgrounds and at different fluency levels (from beginner to advanced). It is annotated with raters' holistic fluency scores, scores for each phone, a transcription of both the target phone and any substituted phones, as well as detailed comments on the nature of any pronunciation errors. Given the level of annotation detail, it is anticipated that this corpus will be an excellent resource for researchers studying the spontaneous speech of L2 learners, for educators, for professionals in educational testing and assessment, and for researchers working in automatic speech recognition technology.

Construction of the annotated spontaneous speech database

Participants

28 non-native speakers of English were recorded in the phonetics lab at the University of Illinois Urbana-Champaign. Of the participating students, 22 were recruited from intermediate and advanced level pronunciation classes at the Intensive English Institute (IEI) at the University of Illinois. Six participants were graduate students in the Linguistics department at the University

of Illinois at Urbana-Champaign. The number of students from each language group and background information are provided in table 1 and 2 below. Details of the rating methods/procedures are provided in the section titled “Rating”.

Table 1.

Native Languages of Speakers

Language	Korean	Chinese	Spanish	Other
Number of Speakers	14	8	3	3

Table 2.

Background Information of Speakers

	Mean	Range
Age	27.7	18~34
length of residence in US	1.3 years	1 month ~ 6 years
Age at start of English instruction	13.6	10 ~ 31

Asian students represented about 80% of the speaker population; 50% were Korean and 28% were Chinese. Other represented groups included Arabic and Turkish (10%).

The two groups (students from the IEI and the graduate students from the linguistic department) were differently distributed in age and length of residence (LOR) in the US. The mean age of the IEI students was 26.4, while the mean age of graduate students was 31.6. The mean LOR of the IEI students was 6.4 months: the mean LOR of graduate students was 3.8 years. The age of onset of English instruction was similar across students, with an average of 13.6.

Material and procedures

The speech was recorded in a sound attenuated booth in the phonetics laboratory at the University of Illinois at Urbana-Champaign. The speech data were collected using prompts that were composed of 8 questions: two questions required the participants to describe a movie that they liked or a country they wished to visit. Two questions were picture description tasks and two questions required the learners to provide an opinion about a social issue (after reading a short passage). Finally, there were two prompts that required the participants to give directions (after reading a map).

The questions were presented in a PowerPoint presentation on a computer screen. Participants were given 30 to 60 seconds (depending on the prompt) to prepare and 30 – 60 seconds to respond. An electronic beep signaled when they were to begin and end speaking. The allocated response time was tracked on the computer screen and was automatically reset at the end of either the response or the preparation time. In total, each speaker provided a 6.5 minute speech sample.

The frequency of each phoneme is important in automatic pronunciation assessment. In order to detect segmental pronunciation errors reliably, each phoneme should occur with a reasonable minimum frequency. In assessments using read speech, this is less problematic since it is possible to use sentences balanced for the distribution and frequency of phonemes. Obviously, the frequency of individual phonemes is less controllable in spontaneous speech samples. In order to address this, pronunciation error patterns predictable from differences between the L1 and L2 phonological systems, were collected from Swan & Smith (2002). From their study, English phonemes that cause the greatest difficulty for L2 learners whose native languages are Korean, Chinese, and Spanish were identified, and these phonemes were included in the map task prompt.

Description of the database

Transcription and statistics

The speech data were transcribed at the word level by two linguistics students. Word fragments, filled pauses, and silent pauses longer than 0.2 second were included in the transcription. Unintelligible words were treated as unknown words.

From the transcription, the distribution of the words and phonemes were analyzed. The speakers spoke 98.13 words per minute on average, with the fastest speaker producing 947 word tokens – twice as many as the slowest speaker, who produced 474 word tokens. However, there were fewer differences in word types used among the speakers; the speaker with the greatest diversity in word types used 290 different word tokens, while the least diverse speaker used 197 word tokens.

Rating

General Score

All files (28 speakers * 8 responses) were rated based on the TOEFL iBT speaking rubrics by two experienced ESL teachers. The TOEFL iBT rubric provides a general description for 5 levels of fluency. The raters provided a general score for each sound file using a 0-4 point scale where 0 indicates no response or no attempt to respond and 4 indicates native like fluency. Table 3 provides the TOEFL iBT speaking rubrics.

Table 3.

iBT Fluency Scores of Speakers

Score	General Description
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse.
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas.
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech although problems with delivery and/or overall coherence occur, meaning may be obscured in places
1	The response is very limited in content and/or coherence is only minimally connected to the task or speech is largely unintelligible.
0	Speaker makes no attempt to respond or response is unrelated to topic

Before rating the speech files, the raters were trained on 27 sample speech files. Twelve of the speech files were taken from the Educational Testing Service's (ETS) "The Official guide to the New TOEFL iBT" (2006a) and fifteen were sample files from the database in this study. In the initial training, however, more than 50% of the samples were rated with a score of 2 even though fluency differences among them were noticeable. Therefore, in order to get a more refined picture of the variation, the 0-4 scale was modified to allow 0.5 increments and raters then underwent an additional training session. During the latter training, raters built consensus around each score and proto-typical files for each level (0, 0.5, 1, 1.5 etc) were selected and used as references during the actual rating.

A total of 224 files were divided into three sections; Section 1 was scored twice by both raters; Sections 2 and 3 were each rated by one rater. Section 1 consisted of 58 files, while Sections 2 and 3 each consisted of 83 files. For any given speaker, both raters rated at least one file in common and then four additional files individually. The scores of the 6 responses were averaged to get a final speaker fluency score (these are summarized in Table 4).

Table 4.

Fluency Scores of Speakers

Score range	2.0 ~ 2.5	2.6~3.0	3.1 ~ 3.5
Number of subjects	14	8	6

The inter-rater reliability was calculated based on the Pearson correlation and mean square errors. Table 5 provides the Pearson coefficient and mean square errors between the two raters' scores. The reliability of the response scores was calculated based on 58 files scored by both raters. For each speaker, each rater's scores were averaged separately and the reliability of the speaker scores was based on these two mean scores.

Table 5.

Inter-rater Reliability of General Scores

	Mean Square Error	Pearson Coefficient
Speaker	0.14	0.70**
Response	0.27	0.70**

Pearson coefficients were 0.7 for both levels and the two raters' scores showed statistically significant correlation.

The scores were classified into three groups; exact agreement, adjacent agreement (that is, the difference between two scores is equal to one level), and non-adjacent agreement (the difference between two scores are larger than one level). Table 6 provides the agreement ratio between the two raters.

Table 6.

Agreement of Response Score between Two Raters

	Exact	Adjacent	Non-adjacent
Response (%)	62	21	17

Because a 0.5 increment scoring system was used, the non-adjacent agreement ratio was noticeably higher than the non-adjacent agreement ratio reported in ETS's research report on the scoring of the TOEFL Academic Speaking Test (2006b). This latter study was based on the TOEFL iBT speaking rubric – which uses a whole number scoring system, that is without the .5 increments. However, the exact agreement ratio reaches levels similar to those in the ETS report.

Since the Non-adjacent agreement was rather high, it is important to investigate whether this is attributable to a particular speaker. Therefore, the average speaker scores were classified into three groups based on the difference between two raters' scores and analyzed. Table 7 shows the agreement ratios of the average speaker scores.

Table 7.

Agreement of Average Speaker Score between Two Raters

	Difference <0.5	0.5<=Difference<1.0	1.0 < Difference

Speaker (%)	86	14	0

The two raters' scores differed by less than 0.5 for most speakers (86%). For four speakers (14%), the two raters' scores differed by more than 0.5. The raters were brought together to listen again to the 8 sound files from the four speakers about whom they had disagreed in order to discuss possible reasons. The content of the discussion is summarized in the Discussion Section.

Phone rating

The same two raters rated each phone in the spontaneous speech data. The speech files were automatically segmented using a forced alignment algorithm and the target utterance was transcribed on tier 3 of the TextGrid. Since raters already had experience with the sound files, several steps were taken to guard against rater-bias based on the earlier overall fluency ratings. First, the speech files were segmented into sub-files of approximately 10 seconds each. They were then provided to the raters in random order with a minimum interval of approximately three weeks after the overall fluency rating. In the event that raters recognized the speaker and/or the segmented sound file extracted from the original speech and remembered the overall fluency rating, they were asked to disregard the latter when assessing individual phones.

Sound files were accompanied by a TextGrid file, created in Praat (Boersma, P. and Weenink, D., 2006), a software program for the analysis of speech. The TextGrid file was synchronized with the acoustic wave form and included a word tier, a phoneme tier, a score tier and a comment tier. The phoneme tiers were designed to contain speaker's target pronunciation of each word. The phoneme tier was filled automatically with pronunciation forms taken from the ISLE dictionary (Johnson and Fleck, 2007). The phoneme tier was modified to reflect actual forms (as they deviated from the ISLE forms). Phone scores were assigned by raters and recorded on the score tier.

The raters labeled each phone using a binary scores ("correct" or "error") with the latter further classified as "substitution", "insertion", "deletion" and "bad". For an error that involved substituting a target phoneme, the raters wrote the phoneme that was actually produced in the comment tier. The raters also wrote comments on vowel length and stress.

In order to calculate intra-rater reliability, several sound files were assigned twice without the rater’s knowledge. Similarly, several sound files were assigned to both raters for inter-rater reliability. If raters were to have different assessments of the inserted and deleted phone, the number of scores of the two raters might be different and the Pearson correlation or the Kappa score could not be used. Therefore, a phone accuracy measure of speech recognition was used to measure inter-rater and intra-rater reliability. The scores of the two raters were aligned using a minimum edit distance algorithm and the reliability was calculated using formula (1).

$$Accuracy = \frac{N - D - I - S}{N}$$

N=total number of phones in the transcription (1)

D= total number of deletions

I=total number of insertions

S=total number of substitutions

Intra-rater reliability scores were calculated using about 8 minutes of spontaneous speech for each rater. Inter-rater reliability scores were calculated using 72 minutes of spontaneous speech. Intra-rater reliability was 96% and 92%. Inter-rater reliability was 89%.

Discussion

General Score

After providing general scores for overall fluency, the raters listened to 4 speakers' sound files over which they had disagreed and discussed possible reasons for the disagreement. One speaker (speaker A) among these 4 speakers was chosen randomly and analyzed in detail. Speaker B, whom both raters assigned similar average scores, was selected and the characteristics of the speech sample were compared to that of Speaker A. Table 8 shows fluency scores and the characteristics of two speakers' speech.

Table 8.

Characteristic of Two Speakers' Speech

	Mean of fluency scores	Difference between raters' scores	Speaking rate	Number of disfluency	Number of errors
Speaker A	2.9	0.7	1.26	32	0
Speaker B	2.2	0.2	1.36	29	5

In a side-by-side comparison, the number of disfluencies (pauses, hesitations, filled pauses etc.) was similar across the two speakers, although they evinced different numbers of actual speech errors; 0 for speaker A and 5 for speaker B. Speaker A also had a higher mean fluency score than speaker B. These findings are indicative of features that influence perceptions of fluency.

Mizera (2006) found two important features which strongly correlated with a human rater's perception of fluency. He pointed out that "accuracy" and the narrow meaning of "fluency"[1] are the most important characteristics of "fluent speech". He demonstrated that there

[1] In the narrow meaning, fluency is considered as one of the component of language fluency – especially temporal aspect of speech. In this definition, fluent speech is continuous and smooth speech and characterized by few disfluencies.

is a correlation between a rater's fluency score and the number of disfluencies; this differs, however, from fluency scores vis-a-vis the number of grammatical errors. The number of disfluencies is a relevant feature of “temporal fluency” while the number of grammatical errors is a relevant feature of “accuracy”. In the above example, Speaker A demonstrated differences in temporal fluency and accuracy. This is supported by a low number of grammatical errors but a high number of disfluencies. Raters showed larger differences in scores when the speaker manifested differences between skill sets - in this case, the accuracy and temporal fluency of speaker A. Conversely, raters assigned similar scores when the speaker was similarly skilled in both accuracy and temporal fluency - in this case, speaker B. The differences between raters are related to perceptual models of fluency. A detailed analysis of the causes of disagreement will be an important research question for future work.

Phone Score

In phone rating section, reference was made to the rating system for the individual phones, that is, that each phone was rated using a binary score of "correct" or "error" with the latter further sub-categorized by error-type as follows;

Insertion: the speaker pronounces a word with an additional phone.

Deletion: the speaker deletes a phone.

Substitution: the speaker substitutes a different phone for target phone.

Bad: an error which cannot be classified into insertion, deletion or substitution

After the individual phone rating was completed, the (evidently) catch-all category of bad” errors were analyzed in detail. We found that most of those errors marked simply as “bad” were classifiable as one of two types; the sound substituted for the target phone was unclassifiable by the raters or the error was a combination of errors.

An error might occur in a phone that has a less categorical instantiation – for instance, a vowel that is neither target-like nor clearly a substitution would fall under the designation of "bad". Equally, differences in voice onset times (VOT) for voiceless stop consonants were designated as "bad" when the VOT values were too short or too long for the categorical placement of the targeted phoneme, but not enough to nudge the production into a different category.

Secondly, if the targeted lexical item was [bəkΛz] (*because*) but it was produced [bi:kΛz] where the first vowel was long and tense rather than reduced to schwa, the error was attributable to both a "substitution" error and an error in stress placement. The deviation from target was marked "bad" and an explanation was then noted in the comment line.

In order to investigate the most frequent error type, the errors were classified into sub-categories and the ratios of sub-categories were calculated.

Table 9.

Ratio of Sub-categories among Errors (%)

Category	Rater 1	Rater 2
Substitution	32.0	37.3
Insertion	10.8	13.6
Deletion	21.7	15.0
Bad	35.5	34.1

The most common errors excluding “bad” were substitution errors and, as might be expected, these tended to vary by L1 groupings, e.g [l]~[r] substitutions for Japanese and Korean speakers; [p]~[f] for Korean speakers. There were also less obvious or intuitive errors of substitution. For instance in the phrase *I am*, a fluent L1 speaker would not have any juncture between the two words. However, since there is a constraint in English against vowels occurring together (unless they are diphthongs), the fluent L1 English speaker inserts a glide, e.g. [aijæm]. The L2 speakers often inserted glottal stops between the two vowels, resulting in an error.

Insertion errors were most often cases of epenthetic or paragogic vowels – that is, vowels added by the speaker to "repair" syllable structures that would be illicit in the L1.

Deletion errors were also quite common, most often occurring in codas or occasionally in complex onsets. However, stop consonants deleted in codas of words in prosodically weak positions, that is, in places where L1 speakers are also likely to delete them, were not marked as errors (e.g. "You can't go" where the /t/ in *can't* would be deleted.). This brings up an important question: undershoot, (or lenition), assimilation, reduction and deletion of consonants and vowels in prosodically weak positions is common in L1 connected speech. The question the raters struggled with was whether the same variant in L2 speech was target-like, given the spontaneous speech environment, or if the same phenomena constituted errors in the L2 grammar.

In the phone rating section, we reported an inter-rater reliability rate of 89%. Although the speech data included the complex characteristics of spontaneous speech, the agreement ratio was similar to Witt (1999) which was based on read speech. In order to improve the inter-rater reliability in the future, the phonemes upon which raters disagreed were examined and analyzed.

In spontaneous speech, positionally determined variants seemed to be a significant reason for disagreement. In connected, fast or casual speech, an L1 English speaker may "undershoot" an articulatory target, resulting in a lenited form. Equally, assimilation – both within and across word boundaries – as well as reduction and deletion of consonants and vowels in prosodically weak positions is common in L1 English connected speech. For example, a fluent English speaker may produce *cupboard* with the medial [b] lenited to the point that it approaches a [β] (a voiced bilabial fricative) – which acoustically strongly resembles [v]. The utterance – devoid of topdown processing - results in minimal-pair counterpart, that is, *covered*. While this is common in connected speech of L1 English speakers, the same variant in L2 speech may be marked as 'non-native,' particularly if the speaker's L1 has [β] as an allophonic variant, e.g. Spanish.

While raters reported being sensitive to environmentally determined variants, in discussions with the researchers, they indicated that they had judged the accuracy of a variant based on overall patterns of speech – that is, generalizations made over the whole sound file of each speaker. After the discussion, the raters decided to consider consistency in rating variants found in connected speech. For example, a speaker who demonstrated a general substitution pattern of [d] for [ð], but in an instance where [d] could be expected (e.g. *add the*) the use of [d] was not considered a variant that was determined by articulatory assimilation across a word boundary.

Conclusion

This paper reported on the construction of a rated database of spontaneous speech produced by L2 learners of English. It is annotated with both general fluency scores and individual phone accuracy scores. The construction of such a database highlighted several difficulties that should be considered in future or related work. As mentioned above, speech fluency is a subjective judgment that can be influenced by both temporal features and accuracy.

In order to achieve reliably high agreement ratios – both within and between raters – a significant amount of training is required for raters or it is necessary to identify and work with raters who already have experience. Clearly, a rated database requires large amounts of labor: recruiting a balanced pool of participants, segmenting and annotating sound files, recruiting and training raters, and of course the actual rating. Phone rating took the longest amount of time in the course of the database construction; one minute of phone rating required an average of 25 minutes.

The database will be released to the public in the near future. The database is still relatively small, comprising just 182 minutes of spontaneous speech from 28 L2 speakers, but it is still useful in developing automated scoring algorithms. However, it would certainly be desirable to develop additional databases of this kind.

References

P. Boersma and D. Weenink, (2006). Praat: doing phonetics by computer (Version 4.5.02) [Computer program]. Retrieved November 16, 2006, from http://www.praat.org/

H. Bratt, L. Neumeyer, E. Shriberg, and H. Franco (1998). Collection and Detailed Transcription of a Speech Database for Development of Language Learning Technologies. *Proceedings of international Conference on Spoken Language Processing*, pp. 1539-1542

Educational Testing Service (2006a). *The Official Guide to the New TOEFL iBT*, New York: McGraw-Hill

Educational Testing Service (2006b). *Investigating the Utility of Analytic Scoring for the TOEFL Academic Speaking Test (TAST)*, Princeton, NJ. : X. Xi, P. Mollaun

M. Johnson and M. Fleck (2007). International Speech Lexicon (Version 0.2.0) [online dictionary]. Retrieved January 3, 2008, from http://www.isle.uiuc.edu/dict

Y. Kim, H. Franco and L. Neumeyer, (1997). Automatic Pronunciation Scoring of Specific Phone Segments for Language Instruction, *Proceedings of European Conference on Speech Communication and Technology*. pp. 649-652.

G. Mizera, (2006). *Working memory and L2 Oral Fluency*. Unpublished Ph.D.dissertation, University of Pittsburgh, Pittsburgh, US.

M. Swan and B. Smith. (2002). *Learner English*. Cambridge: Cambridge University Press.

S. Witt, and S. Young, (1998). Performance Measures for Phone-Level Pronunciation Teaching in CALL, *Proceedings of the Workshop on Speech Technology in Language Learning*, pp 99-102.