Factors in Raters' Perceptions of Comprehensibility and Accentedness

Heesung Grace Jun

Jinrong Li

Iowa State University

The purpose of this study is to examine native-speaker (NS) and non-native speaker (NNS) raters' thought processes while rating L2 speech samples for their comprehensibility and accentedness and identify the factors that may cause listeners to rate speech in certain ways. Think-aloud or vocalization of the thought processes of each speech rater was used to understand what aspects of speech and pronunciation six raters noticed while rating seven ESL learners' speech for comprehensibility and accentedness. We found that there were both similarities and differences between the factors noticed while rating for accentedness and comprehensibility. In addition, the NS and NNS raters showed some major differences in the aspects mentioned during think-aloud.

INTRODUCTION

The relationship between the comprehensibility and accentedness of second language (L2) speech has been shown to be complex and significant for the teaching and assessment of L2 pronunciation. Arguably, comprehensible speech is a more reasonable and realistic goal than accent-free speech. However, it might not be easy to draw a clear-cut line between the two aspects, as research has indicated that the ease of understanding of an L2 speech sample might be affected by accented features of the speech sample (e.g., Munro & Derwing, 1995).

Several studies have revealed the complex interrelatedness of ratings of comprehensibility and accentedness of L2 speech. Generally, accentedness refers to the degree to which "the pronunciation of an utterance sounds different from an expected production pattern" (Munro, Derwing, & Morton, 2006, p. 112). On the other hand, understanding of L2 speech has been further differentiated into intelligibility and comprehensibility. While intelligibility emphasizes actual understanding, which is usually assessed by transcription tasks, comprehensibility focuses on listeners' estimation of difficulty in understanding (Munro, Derwing, and Morton, 2006, p. 112). We have chosen to focus on comprehensibility rather than intelligibility in the current study because comprehensibility has been shown to more affected by factors related to accentedness.

Munro and Derwing (1995) and Derwing and Munro (1997) examine the relationship between comprehensibility, intelligibility, and accentedness. The findings of these two studies show that there is a "quasi-independent" relationship between comprehensibility and accentedness

(Derwing & Munro, 1997, p. 1; Munro & Derwing, 1995, p. 73). More specifically, they found that comprehensibility and accentedness were related, but that their correlations were not very strong for most of the raters. Moreover, the strength of correlation varied among different raters. Thus, the findings seemed to provide more support for the idea that some accent features do not interfere with comprehensibility.

Derwing and Munro (1997) added open-ended questions to the rating sheet to explore the factors that might have affected raters' judgment. The study identified eight common categories of contributing factors including segmental features, grammar, speech rate, prosodic features, fluency, enunciation, speaking volume, and vocabulary. The researchers concluded that the factors were weighted differently in rating for comprehensibility and accentedness. However, it is not clear whether there might be other underlying factors affecting comprehensibility and accentedness. Furthermore, it is unclear if these findings would hold true for both native-speaker (NS) raters and non-native speaker (NNS) raters. Therefore, the purpose of this study is to further examine NS and NNS raters' thought processes while rating L2 speech samples for their comprehensibility and accentedness and identify the factors that cause listeners to rate speech in certain ways.

We chose to address this issue by using think-aloud protocols. Most recent research on raters' thought processes when they are evaluating oral skills is based on either stimulated recall (e.g., Winke, 2008), stimulated recall and post-task interviews (e.g., Isaacs & Thomson, 2009), or verbal reports, also called think-aloud protocols (e.g., Brown, Iwashita, & McNamara, 2005). Think-aloud protocols are commonly used to explore people's thought processes while completing a task, but few studies besides Zielinski (2008) have used it to explore raters' thought processes while rating for comprehensibility and accentedness. Therefore, we believe that the technique might help to notice the details of the raters' internal thought processes while rating and provide evidence for why raters rated as they did.

Specifically, the study seeks to address the following three research questions:
1. What do raters notice while rating ESL learners' speech for comprehensibility and accentedness?
2. Are there differences between aspects of pronunciation that raters notice when rating for comprehensibility and when rating for accentedness?
3. Are there differences between native-speaker (NS) and non-native speaker (NNS) raters in the aspects they notice?

METHOD

Participants

Speech samples were collected from seven ESL learners of different proficiency levels. Table 1 below summarizes the characteristics of the speakers. Since all of them are students at the same university which requires international students to take an English placement test upon their arrival, we classified them into different proficiency levels based on their test results. Specifically, 101C is a course designed to help undergraduate students with their academic writing skills, and 101D aims to help graduate students with their skills in writing research

articles. When undergraduate students are waived from 101C, they can take regular first-year composition classes with native-speakers. On the other hand, when graduate students are waived from 101D, they are not required to take any extra ESL classes.

Table 1. Characteristics of Speakers Rated in the Study

Speaker	Gender	L1 background	English proficiency	Level of education
1	Female	Korean/English	101C waived	Sophomore
2	Female	Chinese	101D	Graduate
3	Male	Chinese	101C	Freshman
4	Female	Chinese	101C	Freshman
5	Female	Chinese/Japanese	Intensive English	Graduate
6	Male	Indian	101D waived	Graduate
7	Female	Chinese	101C waived	Junior

Among the seven participants, Speakers 1, 6, and 7 had the highest proficiency level since their English placement test results indicated that they did not need to take extra ESL writing classes at the undergraduate or graduate level. Speakers 2, 3, and 4, on the other hand, were placed into ESL writing classes for undergraduates or graduates, suggesting that their proficiency level was lower. Speaker 5 is a graduate student; however, English is her third language, and at the beginning of the study, she was still in the intensive English program which is usually for students who do not meet the university's foreign language test requirement. Thus, Speaker 5 is at the lowest proficiency level in this group. Additionally, as shown in Table 1, five of the seven speakers are native Chinese speakers.

The raters were three NS raters and three NNS raters. Table 2 summarizes their characteristics. All have had some foreign language learning and ESL teaching experience. Four were doctoral students in applied linguistics (AL), while the other two were instructors in an intensive English program.

Table 2. Characteristics of Raters in the Study

Rater	Gender	L1 background	Occupation
1	Male	NS of English	ESL instructor
2	Male	NS of English	Doctoral student in AL
3	Male	NS of English	ESL instructor
4	Female	Korean	Doctoral student in AL
5	Male	Russian/Ukrainian	Doctoral student in AL
6	Male	Chinese	Doctoral student in AL

None of the raters were acquainted with the speakers, and thus, they were not familiar with the speakers' voices and/or accents. The Chinese rater might have been affected by the fact that the speakers were mostly Chinese, as studies have shown that non-native listeners' listening comprehension is affected differently depending on the native language of the speakers (Gass & Varonis, 1984; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002). However, since the rater is a trained ESL teacher, we believe that the possible effect of one rater's familiarity with the

speech characteristics of the majority of the speakers on the study's results could be regarded as minimal in these circumstances.

Materials

Speech samples were elicited by using a short passage from Bailey and Nunan (2005) for reading-aloud and three questions for spontaneous speech on a familiar topic (traveling, hobbies, and health). See Appendix A for the passage and questions. The 9-point rating scales for comprehensibility and accentedness were adopted from Munro and Derwing (1995) and Derwing and Munro (1997) since they have been widely used in studies on listeners' perceptions of L2 comprehensibility and accentedness. See Appendices B and C for the rating scales.

Procedures

Participants were recruited on a voluntary basis. Each speaker met with one of the researchers to read the passage aloud and talk about a topic of his or her own choice. No time limit was set for the spontaneous speech. The researchers recorded the speakers' performances individually using a digital voice recorder and then imported the data into Audacity, a voice recording and editing software program. In Audacity, the researchers extracted the first thirty-second segment of each response as speech samples and prepared a total of fourteen speech samples, with two from each of the seven speakers. The speech samples were then transformed into MP3 files and arranged in random order.

The rating sessions were arranged with each rater individually. For each session, a rater came to a computer lab to listen to the speech samples, using a headset to perform the think-aloud task which was recorded by a digital voice recorder. The raters were asked to listen to the speech samples and rate them for comprehensibility. They were also asked to think aloud about factors that might have affected their comprehensibility ratings at the same time. Then, the raters listened to the recordings again and gave an accentedness rating to each recording while thinking aloud about factors that might have affected their accentedness ratings. The think-aloud files were then imported to Audacity and converted into MP3 format for transcribing and coding.

We did not provide any training to the raters ahead of time because the purpose of the study is to explore raters' intuitive perception of L2 speech. Prior training might influence the raters in their analysis of the factors underlying their rating of comprehensibility and accentedness of the L2 speech samples. Since all the raters have had exposure to ESL students, we assumed that their understanding of comprehensibility and accentedness and different degrees of the two speech aspects would be comparable.

Analysis

In a pilot study, the researchers listened to raters' think-aloud data and came up with a coding scheme of tentative categories of underlying factors that might have affected raters' judgment of the comprehensibility and accentedness of speech samples. Using this coding scheme, the researchers coded the new think-aloud data after transcribing the key words in the raters' comments. The think-aloud files were separated into twenty-eight segments, with each focusing

on either the comprehensibility or the accentedness of the fourteen speech samples. For the first two segments, the authors listened, transcribed, and coded together, while discussing ways to handle controversial coding problems and developing rules of coding. Based on the data from this first stage, the authors made modifications to the existing coding scheme by adding new categories and deleting old ones. The final scheme can be found in Table 3. The authors then listened to and transcribed the rest of the segments (twenty-six segments) individually and later coded the data together while comparing and discussing the transcripts to have 100% agreement.

RESULTS AND DISCUSSION

Table 3 displays the final coding scheme used for coding the think-aloud data together with the total counts for each aspect. The full coding results of the raters' think-aloud data can be found in Appendix D. Twenty-four aspects of speech were identified as having been mentioned in the think-aloud data, and these aspects were categorized into four groups—segmental features, suprasegmental features, global impression, and others. Counts were obtained for each aspect, with total and separate counts for focus of rating (i.e., whether the rater was rating for accentedness or comprehensibility) and the native vs. non-native rater distinction.

RQ1: Aspects noticed while rating for comprehensibility and accentedness. The first research question asked what raters noticed while rating ESL learners' speech for comprehensibility and accentedness. In very broad terms, we can say that the raters noticed all of the twenty-four aspects in the final coding scheme, but to go into more detail, the total counts from Table 3 will be considered. The first main finding is that raters frequently mentioned ease of understanding (67) while rating for comprehensibility and accentedness (90) while rating for accentedness. However, this is somewhat of an expected outcome because we can expect raters to mention and describe the speech feature for which they are currently trying to assign a rating. The more interesting finding seems to be that while raters were rating for accentedness, they mentioned comprehensibility (33), and in the same way, when raters were rating for comprehensibility, they mentioned accentedness (22). This finding could point to the possible interrelationship or "quasi-dependence" between comprehensibility and accentedness that was suggested by Derwing and Munro (1997). Also under the global impression category, identification of L1 (31) and speech rate (20) were often noted by raters as factors influencing comprehensibility and accentedness ratings.

Other general findings were obtained from the total counts. Firstly, segmentals, particularly consonants (43) and vowels (32), had very high total counts. This could be because segmentals are more salient features in speech. Another possible reason is that these features have been studied extensively, and thus raters may already have accessible language and terms to discuss errors associated with individual sounds. Raters were often able to point out specific segmental errors such as /r/ insertion after a vowel or specific vowels and diphthongs that were mispronounced. They were even able to suggest reasons for incorrect sounds such as "…the tongue is too far forward for s's" (Rater 1).

Of the suprasegmental features, linking was the most commonly mentioned aspect (31). In contrast to when they were commenting on segmental features, when raters were commenting on suprasegmental features, their discussion tended to be more general. In most cases, raters related their general judgments from intuition rather than pointing out specific reasons for their decision

as they had done while talking about segmental features. For example, a typical comment on intonation was "…the intonation doesn't sound right…" (Rater 4). This may point to the raters' lack of precise language to discuss suprasegmental features.

Table 3. Final Coding Scheme for Think-Aloud Data with Totals for Each Aspect

Categories	Aspects	Total
Segmental	Consonants	**43**
	Vowels	**32**
	Syllables (schwa insertion)	4
Suprasegmental	Intonation	23
	Linking	**31**
	Stress	12
	Rhythm	15
Global impression	Enunciation	8
	"Word pronunciation"	9
	"Pronunciation"	16
	Identifying L1	**31**
	Ease of understanding (C)	67
	Comprehensibility (A)	**33**
	Accentedness (C)	**22**
	Accentedness (A)	90
	Speech rate	**20**
	Pauses/fillers	12
	Good sense of language	3
	Fluency	4
Others	Grammar	**6**
	Speech impediment/lisp	11
	Type of speech (read vs. spon.)	6
	Quality of recording	6
	Recognition of words	12

An interesting finding from the Others category was related to grammar features. When Raters 1 and 3 rated the speech of Speaker 1, they noticed grammatical mistakes. Speaker 1 speaks very fluently, almost native-speaker-like, and this may be why the raters could shift their attention away from pronunciation and focus on grammatical aspects. This finding suggests that with high comprehensibility and low accentedness, the rater's attention can shift to grammar and the actual content of the speech samples. However, when we considered the average ratings together with the types of think-aloud comments for each of the speech samples, no general patterns could be found regarding the relationship between the ratings and the aspects noticed for each speech

sample. We did not find that students with lower comprehensibility and/or greater accentedness ratings had certain types of think-aloud comments.

RQ2: Differences in aspects noticed for comprehensibility rating and accentedness rating. The second research question asked whether there were differences between what is noticed when raters are rating for comprehensibility and when rating for accentedness. Table 4 shows the top aspects noticed under the two rating conditions. A few major overlaps can be found from comparisons to think-aloud results.

Table 4. Most Noticed Features for Comprehensibility and Accentedness Ratings

Comprehensibility rating	Accentedness rating
Consonants (21)	Consonants (22)
Pronunciation (14)	Identification of L1 (22)
Vowels (12)	Vowels (20)
Linking (12)	Linking (19)
Speech rate (12)	Intonation (16)
Recognition of words (10)	

First, consonants, vowels, and linking seemed to be noticed for both comprehensibility and accentedness. These were the overlapping factors. The remaining three aspects under the comprehensibility rating condition—pronunciation, speech rate, and recognition of words—affect the ease of understanding, and the raters could have mentioned them more while rating for comprehensibility. An example of a rater's think-aloud comment that makes concurrent reference to speech rate and comprehensibility is "Although this person talks slowly, it's easy to understand" (Rater 1). On the other hand, when raters were listening for accentedness, they often tried to identify the L1 or the source of the accent, e.g., "Chinese speaker who learned to pronounce /r/ like Americans" (Rater 1) or "it's a strong South Asian, Indian, Pakistan, Bangladesh, Sri Lankan accent" (Rater 2). Furthermore, the suprasegmental features of linking and intonation seemed to have been noticed in the rating of accentedness more than in the rating of comprehensibility.

RQ3: Differences between NS and NNS raters. The third research question asked whether there were differences between NS raters and NNS raters in the aspects that they noticed in the learner speech. Table 5 shows the major differences between NS raters and NNS raters.

Table 5. Differences Between NS and NNS Raters

	NS raters	NNS raters
Consonants	13	30
Vowels	6	26
Intonation	2	21
Linking	2	29
Comprehensibility (in accentedness rating)	26	7

Speech impediment/lisp	10	1

NNS raters noticed segmental features more often than NS raters did. In addition, NNS raters were much more aware of other pronunciation features such as intonation and linking. On the other hand, NS raters commented on comprehensibility more often than NNS raters while rating for accentedness. The NS raters often made comments on ease of understanding, such as the amount of "attention" or "extra effort" needed for comprehension and "being frustrated" because of the incomprehensibility of a speech sample. It appears that accented speech may have had a larger effect on the NS raters' ease of understanding or comprehensibility of speech samples compared to NNS raters. Another finding was that only NS raters noticed and commented on the lisp that one of the speakers had. All ten comments were about the same speaker. The one comment coded as speech impediment/lisp from a NNS rater was "stuttering" for another speaker. Overall, the NNS raters focused more on specific pronunciation features, while NS raters were more global in their assessments, focusing on the overall impression of the speech sample or paying attention to whether they understood the intended message. These findings are based on total counts for the NS and NNS groups. Despite the fact that there are only three members per group, analysis of the raw counts for individual raters showed that no one rater highly swayed the results.

CONCLUSION

In this study, we explored aspects of pronunciation and speech that raters attend to while rating comprehensibility and accentedness of ESL learners' speech samples. Through the analysis of raters' think-aloud data, it was found that there were both similarities and differences between the factors noticed while rating for comprehensibility and accentedness. In addition, the NS and NNS raters showed some major differences in the aspects mentioned during think-aloud.

There are several limitations in our study that need to be addressed. One limitation concerns the rating procedure. In our study, the selection of read-aloud speech samples from the participants was not entirely random. We selected the first 30 seconds of the recordings from everyone, and raters were already familiar with the content when they gave ratings to the later speech samples. Thus, later ratings in the randomized sequence might have been affected by previous exposure to the same content or voice. The raters also had a single sheet of paper to record all the students' scores. This might have encouraged them to rate students against each other since they could compare their scores back and forth. Furthermore, the same segments were used for the two ratings (comprehensibility and accentedness) and the ratings were done one right after the other, so raters could remember the speakers, and practice effects and/or carryover effects were probably present. Also, the number of times a rater listened to the speech samples was not controlled for, and as a result, a couple of raters listened twice to some speech samples, although most listened only once.

Other limitations lie in data analysis and coding. Aggregate counts were used in the results, so we did not take into consideration the issue of verbose vs. reticent raters and the possible bias this may have created. In addition, in the coding process, we identified features that were difficult to categorize, for example, "sounds natural" or "speaks clearly," and these were not included in the total counts in the final coding scheme.

Directly following from the limitations of our study, we would like to make some suggestions for future studies in the same topic area. First, regarding the rating procedure, the order of rating should be counter-balanced (accentedness → comprehensibility for one half of the raters and comprehensibility → accentedness for the other half of raters). Alternatively, to increase the reliability of the ratings and think-alouds, it might be better for a future study to select different segments for the comprehensibility ratings and the accentedness ratings or conduct separate rating sessions for comprehensibility and accentedness with at least a few days between the two rating sessions to reduce the practice effect coming from memory. Also, we would suggest using different segments from different speakers if the pronunciation is going to be evaluated for global impression. To lessen the possibility of raters comparing speakers against each other, we suggest having each score be written on a separate slip of paper so that the raters would be less likely to remember the scores they had recorded for previous speakers.

In the data analysis and coding stage, the counts for each aspect could be standardized for each rater to reflect the amount of think-aloud that each rater produced. To solve the problem of think-aloud comments that are difficult to categorize, we suggest conducting follow-up interviews to clarify what the raters meant. In addition, a comparison of the think-aloud data with the actual speech samples by the researchers or a third rater might help shed light on what went unnoticed or what was not noted by the raters, particularly aspects such as sugrasegmentals that raters might not have had the language to discuss during the rating and think-aloud sessions.

Other suggestions regarding study design are for researchers to make sure that they convey the meaning of the terms "comprehensibility" and "accentedness" to the raters a priori so that they can be clear about what they are going to be rating. Rater training and calibration of standards should be provided so that raters are not comparing speakers against each other but against some standard. Furthermore, the raters may need to be provided think-aloud training beforehand. Using a few speech samples for practice would help clarify what is expected and may help to control for the different number of think-aloud comments made across raters.

ABOUT THE AUTHORS

Heesung Grace Jun is a PhD student in applied linguistics and technology at Iowa State University where she teaches academic writing to international graduate students. Her research interests are second language writing and language assessment. Email: gracejun@iastate.edu

Jinrong Li is a PhD student in the program of Applied Linguistics and Technology at Iowa State University. Her research interests are computer-mediated communication and second language acquisition, L2 academic writing, and assessment of speaking and writing. Email: jinrong@iastate.edu

REFERENCES

Bailey, K. & Nunan, D. (2005). *Practical English language teaching: Speaking*. New York: McGraw-Hill.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks* (ETS RR-05-05). Princeton, NJ: Educational Testing Service.

Derwing, T. M. & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 20,* 1-16.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on nonnative speech. *Language Learning*, *34,* 65-89.

Isaacs, T. & Thomson, R. (2009, March). Judgments of L2 comprehensibility, accentedness, and fluency: The listeners' perspective. Paper presented at the 31st Language Testing Research Colloquium, Colorado, Denver.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, *36*(2), 173-190.

Munro, M. J. & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73-97.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*, 111-131.

Winke, P. (2008, September). Beyond the rubric: Raters' thought processes while rating speech samples. Paper presented at the Midwest Association of Language Testers conference, Iowa City, Iowa.

Zielinski, B. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*, 69-84.

APPENDIX A

1. Read-aloud

Instructions: Go through the passage and see if there is any new vocabulary. You can check with me if there is any. When you are ready, read aloud into the microphone.

When a student from another country comes to study in the United States, he has to find the answers to many questions, and he has many problems to think about. Where should he live? Would it be better if he looked for a private room off campus or if he stayed in a dormitory? Should he spend all of his time just studying? Shouldn't he try to take advantage of the many social and cultural activities which are offered? At first it is not easy for him to be casual in dress, informal in manner, and confident in speech. Little by little he learns what kind of clothing is usually worn here to be casually dressed for classes. He also learns to choose the language and customs, which are appropriate for informal situations. Finally he begins to feel sure of himself. But let me tell you, my friend, this long-awaited feeling doesn't develop suddenly – does it? All of this takes will power.

(Paragraph taken from Bailey & Nunan, 2005)

2. Free response questions

Instructions: Choose and answer one of the following free response questions.

1) If you could visit any place in the world for a month, where would you go and what would you do there?

2) Nowadays, there are a lot of activities, hobbies, or forms of amusement and entertainment to choose from. What do you enjoy doing more than anything else?

3) What are some things that people can do to take care of their health?

APPENDIX B

1. Please rate the speech samples for their comprehensibility on a scale of 9 (1= extremely easy to understand and 9 = extremely difficult or impossible to understand). While you are rating, think aloud about the factors that influenced your rating of the speakers' comprehensibility.

Speech samples	1	2	3	4	5	6	7	8	9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									

APPENDIX C

2. Please rate the speech samples for their accentedness on a scale of 9 (1 = no accent and 9 = extremely strong accent). While you are rating, think aloud about the factors that influenced your rating of the speakers' accentedness.

Speech samples	1	2	3	4	5	6	7	8	9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									

APPENDIX D

Aspects mentioned in raters' think-alouds

Categories	Aspects	Comp	Acct	Total	NS	NNS
Segmental	Consonants	21	22	43	13	30
	Vowels	12	20	32	6	26
	Syllables (schwa insertion)	3	1	4	2	2
Suprasegmental	Intonation	7	16	23	2	21
	Linking	12	19	31	2	29
	Stress	5	7	12	3	9
	Rhythm	6	9	15	8	7
Global impression	Enunciation	6	2	8	3	5
	"Word pronunciation"	5	4	9	6	3
	"Pronunciation"	14	2	16	8	8
	Identifying L1	9	22	31	14	17
	Ease of understanding (C)	67	0	67	38	29
	Comprehensibility (A)	0	33	33	26	7
	Accentedness (C)	22	0	22	13	9
	Accentedness (A)	0	90	90	54	36
	Speech rate	12	8	20	6	14
	Pauses/fillers	8	4	12	5	7
	Good sense of language	0	3	3	3	0
	Fluency	3	1	4	1	3
Others	Grammar	4	2	6	5	1
	Speech impediment/lisp	4	7	11	10	1
	Type of speech (read vs. spon.)	6	0	6	4	2
	Quality of recording	6	0	6	5	1
	Recognition of words	10	2	12	2	10