Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners

Murray J. Munro
Simon Fraser University

Tracey M. Derwing
University of Alberta

One of the chief goals of most second language learners is to be understood in their second language by a wide range of interlocutors in a variety of contexts. Although a nonnative accent can sometimes interfere with this goal, prior to the publication of this study, second language researchers and teachers alike were aware that an accent itself does not necessarily act as a communicative barrier. Nonetheless, there had been very little empirical investigation of how the presence of a nonnative accent affects intelligibility, and the notions of "heavy accent" and "low intelligibility" had often been confounded. Some of the key findings of the study—that even heavily accented speech is sometimes perfectly intelligible and that prosodic errors appear to be a more potent force in the loss of intelligibility than phonetic errors—added support to some common, but weakly substantiated beliefs. The study also provided a framework for a program of research to evaluate the ways

This research was supported in part by a grant from the University of Alberta to the second author and a SSHRC postdoctoral fellowship to the first author. Thanks to J. Flege, B. Derwing, A. Schmidt, H. Southwood, N. Takagi and three anonymous reviewers for comments on earlier drafts.

Correspondence concerning this chapter should be addressed to Murray J. Munro, Department of Linguistics, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6. Telephone: (604) 291–3654. Fax: (604) 291–5659. Internet: mjmunro@sfu.ca

in which such factors as intelligibility and comprehensibility are related to a number of other dimensions. The authors have extended and replicated the work begun in this study to include learners representing other L1 backgrounds (Cantonese, Japanese, Polish, Spanish) and different levels of learner proficiency, as well as other discourse types (Derwing & Munro, 1997; Munro & Derwing, 1995). Further support for the notion that accent itself should be regarded as a secondary concern was obtained in a study of processing difficulty (Munro & Derwing, 1995), which revealed that nonnative utterances tend to require more time to process than native-produced speech, but failed to indicate a relationship between strength of accent and processing time.

The approach to L2 speech evaluation used in this study has also proved useful in investigations of the benefits of different methods of teaching of pronunciation to ESL learners. In particular, it is now clear that learner assessments are best carried out with attention to the multidimensional nature of L2 speech, rather than with a simple focus on global accentedness. It has been shown, for instance, that some pedagogical methods may be effective in improving intelligibility while others may have an effect only on accentedness (Derwing, Munro, & Wiebe, 1998).

For several decades, pronunciation experts have stressed improved intelligibility as the most important goal of pronunciation teaching. As early as 1949, Abercrombie argued that most "language learners need no more than a comfortably intelligible pronunciation" (p. 120). Although this view has been echoed by Gilbert (1980), Pennington and Richards (1986), Crawford (1987), and Morley (1991), there is intolerance for foreign accents in some circles, particularly employers (Sato, 1991). This discrimination appears to have acted as a catalyst for the rise of accent reduction programs, which aim to reduce or eliminate foreign accents altogether. These programs inherently suggest that an accent is, in itself, a bad thing, and is subject to *treatment*, *intervention*, or even *eradication* in much the same way as a language pathology. This

attitude is given credence by pronunciation experts such as Griffen (1980/1991), who stated that "the goal of instruction in pronunciation is that the student (or patient) should learn to speak the language as naturally as possible, free of any indication that the speaker is not a clinically normal native" (p. 182).

The popularity of accent reduction programs may be supported by a general bias against foreign accentedness in speech. Numerous studies have shown that native-speaker (NS) listeners tend to downgrade nonnative speakers (NNSs) simply because of foreign accent (e.g., Anisfeld, Bogo, & Lambert, 1962; Brennan & Brennan, 1981a, 1981b; Kalin & Rayko, 1978; Lambert, Hodgson, Gardner, & Fillenbaum, 1960; Ryan & Carranza, 1975). Thus, second language instructors, curriculum designers, and writers of textbooks may feel obliged to focus attention on accent reduction, without regard to specific features that may interfere with intelligibility, because *any* accentedness is seen as a problem. This assumption is manifested in tests of spoken performance that equate accentedness with a lack of intelligibility (e.g., the Test of Spoken English). However, there is as yet no indication that reduction of accent *necessarily* entails increased intelligibility. The effects of nonnative-like pronunciations on intelligibility are far from clear. In the present study, we have attempted to gain a better understanding of the interrelationships among accentedness, intelligibility, and listeners' perceptions of accent and of comprehensibility.

Error Gravity Hierarchies

Several researchers have attempted to isolate the role of pronunciation, as compared to other linguistic features, in the interpretation of meaning. Gynan (1985), for instance, found that listeners judged that the phonology of Spanish NNSs of English interfered with comprehension to a greater extent than grammatical errors did. Ensz (1982), on the other hand, found grammar was more important than pronunciation for comprehensibility when American NNSs were judged by NSs of French. In a study of

English-accented German, Politzer (1978) found that vocabulary errors affected listening comprehension most significantly, followed by grammar and then by pronunciation. Albrechtsen, Henriksen, & Færch (1980) found little correlation between measures of pronunciation and intonation accuracy and the overall comprehension of taped passages of Danish-accented English. Nonetheless, Fayer and Krasinski (1987) observed that nonnative patterns in pronunciation and hesitation were very strong contributors to listener distraction and annoyance. Albrechtsen et al. (1980), however, have argued that it is pointless to search for hierarchies in error gravity; they claim that the frequency of errors, irrespective of type, is a determining factor. They also suggested that other studies' findings are limited in terms of generalizability because of their narrow focus at the word or sentence level. In their own study, Albrechtsen et al. found evidence of serious adverse effects on comprehension that were related to discourse factors. The apparent contradictions in all of these studies may be at least partially explained by the differences in the target languages under study, as well as by differences in methodology (cf. Schairer, 1992). The effects of second language accent on intelligibility remain unresolved.

Accent Gravity Hierarchies

Not only is there little empirical evidence regarding the role of pronunciation in determining intelligibility, but also there is no clear indication as to which specific aspects of pronunciation are most crucial for intelligibility. Several studies have attempted to establish hierarchies of pronunciation errors (Albrechtsen et al., 1980; Gimson, 1970; Johansson, 1978; Schairer, 1992). In these cases, too, the differences in target languages and research methodologies make firm conclusions impossible. For example, Gimson has argued that accurate production of consonants is more essential to comprehension in English than native-like production of vowels, whereas Schairer came to exactly the opposite conclusion for English-speaking learners of Spanish. Several researchers

have found evidence that prosodic errors are more serious than segmental errors (Anderson-Hsieh, Johnson, & Koehler, 1992; Johansson, 1978; Palmer, 1976). On the other hand, Koster and Koet (1993) and Fayer and Krasinski (1987) argued that segmental errors have the more detrimental effects on comprehension.

Intelligibility, Comprehensibility and Pronunciation

To gain a better understanding of these issues, we need to examine carefully the relationship between foreign accent and speech intelligibility. Intelligibility may be broadly defined as the extent to which a speaker's message is actually understood by a listener, but there is no universally accepted way of assessing it. Many researchers have employed listeners' orthographic transcriptions in their attempts. For instance, Lane (1963) measured intelligibility by counting the total number of words listeners transcribed correctly. Barefoot, Bochner, Johnson, and von Eigen (1993), however, counted percentages of *key* words recognized. Brodkey (1972) considered that accurate paraphrases reflected good intelligibility. In addition, some researchers have asked listeners to directly rate intelligibility on a Likert scale (Fayer & Krasinski, 1987; Palmer, 1976).

In this study, we chose to obtain two types of assessments of listener comprehension in addition to foreign accent ratings. First, we adopted a measurement of intelligibility using a technique similar to that used by Gass and Varonis (1984).[1] In that study, listeners wrote out sentences produced by nonnative speakers. Gass and Varonis assigned scores on the basis of deviations between the transcripts and the intended utterances (e.g., missing words, wrong words). Second, we asked listeners to assign perceived comprehensibility judgments using a 9-point Likert scale. We then examined the relationships between these scores and their relationship with global foreign accent scores.

On the basis of previous work, we anticipated that intelligibility, perceived comprehensibility, and accentedness would be correlated. Varonis and Gass (1982), for instance, argued that the

"main factor involved in judgments of pronunciation was overall comprehensibility or ease of interpretation" (p. 127). However, it cannot be concluded, even when content is controlled, that accent and intelligibility are identical dimensions; that is, the focus of listeners' perception of accent may be somewhat different from the focus of a judgment of comprehensibility.

Method

Speech Materials

Speakers. The speech samples used in this experiment were elicited from 10 native speakers of Mandarin (5 male and 5 female), who had learned English after puberty. All were proficient speakers of English who had scored no less than 550 on the TOEFL, and all had spent a minimum of one year in Canada as graduate students at the University of Alberta. Assessments by the authors, both of whom have had many years of experience with English as a Second Language (ESL) students, indicated that their English pronunciation ranged from moderately to heavily foreign-accented. Recordings were also made of 2 native speakers of Canadian English (1 male and 1 female).

Recording. Individual recording sessions were held in a sound-treated room with high fidelity audio equipment. We gave the speakers a page of cartoons that illustrated an amusing story and asked each person to describe the events depicted. No preparation was allowed; nor were there any verbal exchanges between the experimenter and the speaker during the narration. The entire task took two to three minutes for each participant. To simplify the stimulus preparation procedure, we digitally rerecorded the speech samples at 10 kHz using a Kay Computerized Speech Lab (CSL). We used the waveform editing feature of the CSL to divide the speech samples into shorter excerpts that were of sufficiently short duration to be transcribed by listeners after a single listening.[2] We selected three excerpts from the initial 30 seconds of the

narrative from each speaker, for a total of 36 samples. It was not practical to attempt to break the original recordings down into new samples of exactly identical durations, because this would have resulted in utterances that did not necessarily begin or end at phrasal or clausal boundaries. Instead, the excerpts ended at locations of natural pauses in the utterances, as identified by us. As a result, the final stimulus set of 36 samples varied somewhat in length: the mean length was 10.7 words, with a range of 4 to 17 words. We rerecorded the stimuli in random order onto a cassette tape.

Listeners. The listeners were 18 native speakers of English who were enrolled in either an introductory linguistics course or an ESL teaching methodology course at the University of Alberta. All reported normal hearing, and all had a basic knowledge of articulatory phonetics. We paid each person an honorarium of $10 upon completion of the experiment.

Procedure

We held two listening sessions. During Session 1, we handed the listeners booklets with numbered spaces for transcriptions of each of the 36 utterances. Each space in the booklet also included a Likert scale numbered from 1 to 9. In previous work (Munro & Derwing, 1994, 1995) we found a scale of this size to be effective for eliciting judgments of nonnative speech. We instructed the listeners to listen carefully to each utterance and then write out in standard orthography exactly what they had heard; in other words, to write the utterances word for word. (This was the intelligibility task.) Upon completion of each orthographic transcription, they assigned a perceived comprehensibility rating by circling a number from 1 to 9, where 1=*extremely easy to understand* and 9=*impossible to understand*.

We presented the stimuli through a high fidelity playback system in a quiet room. Before beginning the task, we provided the listeners with two practice stimuli for orthographic transcription and rating. During the experiment, one of the experimenters

controlled the tape by pressing a pause button at the end of each utterance. A new stimulus was not presented until all listeners had finished transcribing the previous one. The entire session lasted approximately 20 minutes.

Session 2 was held four days later. This time we presented the listeners with the same 36 stimuli, but we asked them to rate the degree of foreign (non-English) accent in each sample. We again used a 9-point scale, where 1=*no foreign accent* and 9=*very strong foreign accent*. The same two example stimuli were provided for practice at the beginning of the session. The session lasted approximately 10 minutes.

Results

Coding

We transcribed the complete set of utterances in broad phonetic transcription, playing the stimulus items as many times as necessary, so that the total numbers of phonemic, phonetic, and grammatical errors could be tallied. We made the transcriptions without any knowledge of the scores assigned by the listeners.

We defined phonemic errors as either the deletion or insertion of a segment, or the substitution of a segment that was clearly interpretable as an English phoneme different from the correct one. The total phonemic errors for the Mandarin speakers' productions ranged from 0 to 3 per utterance, with a mean of 0.9. Phonetic errors involved the production of a segment in such a way that the intended category could be recognized but the segment sounded noticeably nonnative. The number of phonetic errors per utterance ranged from 0 to 4, with a mean of 1.6.

We found only 19 morphosyntactic errors in the entire set: 1 utterance contained 3 errors, 3 utterances contained 2 errors and 10 utterances contained 1 error. Of the 30 speech samples produced by the nonnative speakers, 15 were error free. Of the grammatical errors identified, 6 involved inappropriate use of

prepositions, 3 involved errors in subject-verb agreement, 3 were errors in verb tense and 2 were errors in verb form. We also noted one instance of each of the following types of errors: inappropriate article, incorrect number, missing subject, missing object, and missing relative pronoun.

We rated the intonation of each speech sample independently on a scale where 1=*native-like* and 9=*not at all native-like*. We then compared the ratings. In cases where we had assigned ratings more than one scalar unit apart, we played the stimuli again and reevaluated those stimuli independently. In four cases the final ratings were two scalar units apart; all others were identical or only one scalar unit apart. We averaged the ratings for the final analyses. The scores ranged from 1.0 to 8.5, with a mean of 3.8.

We coded the transcriptions provided by the listeners for exact word matches, substitutions (defined as the substitution of one word for a phonetically and semantically similar word, e.g., *who* for *he*), novel words (defined as the insertion of a word bearing no phonological resemblance to a word in the stimulus utterance), and regularizations (e.g., *he walks* for *he walk*). We also identified word omissions and categorized them as either content (nouns, verbs, adjectives, adverbs) or function words (particles, determiners).

Analyses

Judgment Tasks

We tabulated the comprehensibility and accentedness judgments for each stimulus. The mean perceived comprehensibility ratings (pooled across listeners) ranged from 1.0 to 7.6. As expected, the six samples produced by the native speakers of English received the six lowest mean accent scores (i.e., the most native-like ratings). In addition, five of the native English samples received the best mean perceived comprehensibility scores (ranging from 1.0 to 1.4). However, one of the native English samples was rated worse (2.4) than 11 of the nonnative samples. Although

this stimulus received a rating indicating that it was largely heard as unaccented (viz. 1.6), for some reason it was rated as less comprehensible than many of the other samples. This finding does not seem surprising, given that nonpathological native speech may vary in comprehensibility because of such factors as rate of speech, speech clarity, voice quality, and word choice.

Figures 1, 2 and 3 illustrate the distributions of the accent, perceived comprehensibility, and intelligibility scores for the stimuli produced by the native Mandarin speakers. The accent ratings (Figure 1) are fairly evenly distributed across Categories 2 to 8. Only a very small number of the judgments (4%) were ratings of 1, indicating no foreign accent. The listeners were apparently quite successful at recognizing which speech samples were produced by the nonnative speakers. The comprehensibility judgments (Figure 2) show a strikingly different pattern. Twenty-two percent of the samples were rated as *extremely easy to understand* (Category 1) and 64% of the ratings were in Categories 1, 2, or 3. The skewed distribution indicates that the perceived comprehensibility ratings were, on the whole, less harsh than the accent ratings.

Orthographic Transcription Task

The frequencies of the various types of transcription errors appear in Table 1. The orthographic transcriptions of the native English speakers' productions were not completely free of errors; in fact, 44 errors were noted, most of which we classified as substitutions. The number of errors in the transcriptions of the Mandarin speakers was much higher (636), but it must be remembered that there were 10 Mandarin speakers and only 2 native English speakers. When this difference is taken into account, it can be seen that the mean number of errors per speaker was nearly three times greater in the transcriptions of the Mandarin speakers than in those of the native English speakers (63.6 vs. 22.0).

We assigned each of the 648 orthographic transcriptions an intelligibility score on the basis of the number of words that

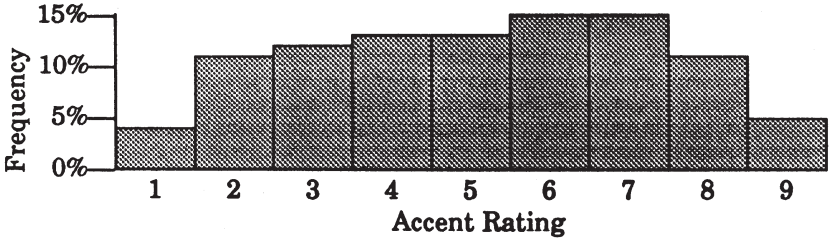

Figure 1. Distribution of listener ratings of the strength of foreign accent (1=*no foreign accent*; 9=*very strong foreign accent*).

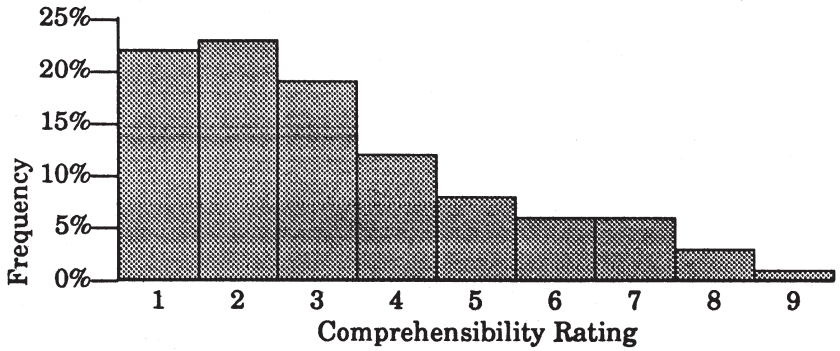

Figure 2. Distribution of listener ratings of perceived comprehensibility (1=*extremely easy to understand*; 9=*impossible to understand*).

exactly matched our corresponding transcription. We also computed an overall intelligibility score for each of the 36 utterances by taking the mean of the 18 listeners' scores for the utterance. The scores for the Mandarin speakers' productions ranged from 39% to 100%; the native English speakers' production scores ranged from 94% to 99%. Five productions were 100% intelligible to all listeners. Surprisingly, these utterances were all produced by Mandarin speakers. The listeners' success in transcribing these stimuli was probably not due to utterance length, because the lengths varied from 7 to 13 words. These items were therefore representative of stimuli in the middle of the length range. An additional seven stimuli from the Mandarin speakers were transcribed with intelligibility scores equal to or above that of the

Table 1

Frequencies of Transcription Error Types

	Mandarin Speakers		English Speakers	
	No.	%	No.	%
Omission (Function Word)	135	21	11	25
Omission (Content Word)	154	24	9	20
Novel Word	76	12	1	2
Substitution	183	29	22	50
Regularization	88	14	1	2
Total	636		44	
Errors per speaker	63.6		22.0	

native English stimulus with the lowest intelligibility score. Finally, five nonnative stimuli were transcribed with at least one error by every listener.

Figure 3 illustrates the distribution of intelligibility scores for the stimuli produced by the Mandarin speakers. Again, the distribution is highly skewed; the largest category by far (64%) is the one including scores from 91% to 100%. In fact, 53% (275) of the transcriptions of the nonnative stimuli received accuracy scores of 100%. Moreover, of the orthographic transcription errors reported in Table 1, more than one third were trivial errors: either omissions of function words or regularizations. On the whole, then, it appears that the nonnative speech samples used in this study were highly intelligible. The distribution of these scores resembles the distribution of the perceived comprehensibility scores more closely than that of the foreign accent scores, though it differs in some respects from both.

We excluded one nonnative stimulus item from further analyses on a number of grounds. First, it received an overall intelligibility score that was considerably lower (39%) than that of the next worst stimulus (68%). Second, it was the first item heard by the listeners (after the practice items) and third, it was relatively long (15 words). Possibly the listeners were not prepared for a stimulus

of this level of difficulty at the outset of the task. Thus, their poor performance on this item may reflect something other than poor comprehension.

Cross-Task Comparisons

One issue here was whether there were significant interlistener differences in the patterns of ratings under the two rating conditions (accent and perceived comprehensibility). First, we assessed interrater reliability on the two ratings tasks by computing intraclass correlations (Shrout & Fleiss, 1979). The correlations

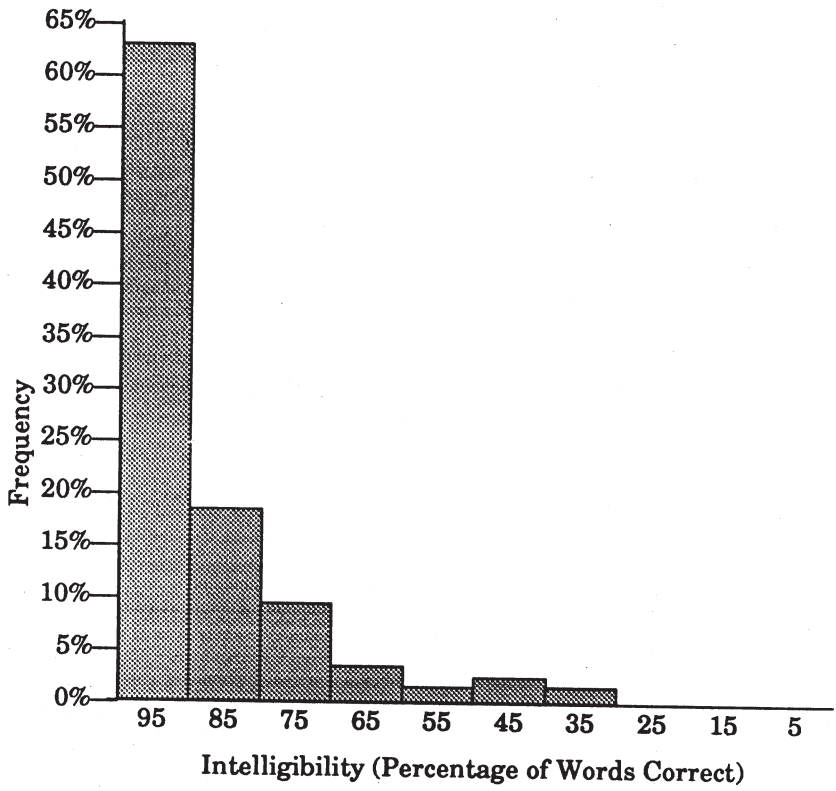

Figure 3. Distribution of intelligibility scores (percentages of words transcribed correctly per utterance).

were very high for both the comprehensibility ratings (0.96, $p<0.05$) and the accent ratings (0.98, $p<.05$), indicating that the raters tended to agree with one another on both.

However, when we examined the correlations between the perceived comprehensibility and accentedness scores for the individual listeners, we observed sizable individual differences in the strength of correlations (see below). For this reason, we decided that it would be misleading to use mean rating data pooled across all listeners in the subsequent comparative statistical analyses. Instead, we chose to examine the data from the individual listeners to gain a better understanding of the results.

We calculated Pearson correlation coefficients (r) for each listener for the accent and intelligibility judgments of the 29 nonnative speech samples[3] and the total numbers of phonemic, phonetic, and grammar errors, intonation ratings and utterance length (in words). We adopted a significance level of 0.05 for all correlational analyses described below.

We first assessed the relationships among the three data sets obtained from the listeners. For all but 1 of the 18 listeners there was a significant positive correlation between the perceived comprehensibility ratings and the accent ratings: evidence that perceived comprehensibility and accent were nonorthogonal dimensions for most listeners. However, the significant correlations ranged from 0.41 to 0.82, indicating that the strength of the relationship between perceived comprehensibility and accent varied a great deal from listener to listener. For 15 of the listeners (83%) there was a significant negative correlation between the perceived comprehensibility (high to low) and transcription intelligibility scores (low to high). The relationship between these two variables suggests that the listeners' perceived comprehensibility ratings tended to reflect their actual understanding of the utterances, measured by their ability to write down exactly what they had heard (not entirely surprising, given that the judgments were made immediately after the transcription task). Again, however, the significant correlations showed a wide range (–0.44 to –.0.90). Finally, for only 5 listeners (28%) was there a significant correlation

between the accent scores and the orthographic transcription (intelligibility) scores. These values ranged from –0.37 to –0.48.

Next, we examined two subsets of stimuli. First, we considered only the five stimuli that were transcribed perfectly by all 18 listeners. Figures 4 and 5 show the distributions of comprehensibility and accent ratings for these stimuli. As expected, the perceived comprehensibility scores tended to be quite low (indicating that the stimuli were easy to understand): as a result, the distribution in Figure 4 is highly skewed. In contrast, Figure 5 illustrates that the accentedness judgments are much more evenly distributed across the range of possible scores.[4] That none of the utterances ever received a foreign accent rating of 1 (perfectly native-like) indicates that all listeners believed that they were produced by nonnative speakers of English. Furthermore, the listeners apparently perceived a wide range of accentedness in stimuli that were nonetheless perfectly transcribed. Highly intelligible stimuli were not necessarily assigned low accent scores.

Table 2 gives the numbers and percentages of the significant correlations between the various stimulus assessments and the three sets of listener scores. The majority of listeners (over 70% in all cases) showed significant correlations between the phonemic, phonetic, intonation, and grammar scores and the accent scores. This finding suggests that our assessments do indeed reflect stimulus properties that the listeners took into account when making their accent judgments. The numbers of listeners showing correlations between these properties and the perceived comprehensibility scores were somewhat lower, however. This tendency was particularly true for the two categories of segmental errors; only 44% and 11% of the listeners showed correlations with the phonemic and the phonetic scores, respectively. This finding suggests that these stimulus properties have more relevance to perceptions of accent than to perceptions of comprehensibility. Further support for this hypothesis surfaced when we considered their relationships with the orthographic transcription intelligibility scores. Only a handful of listeners showed relationships between any of the stimulus properties and the intelligibility

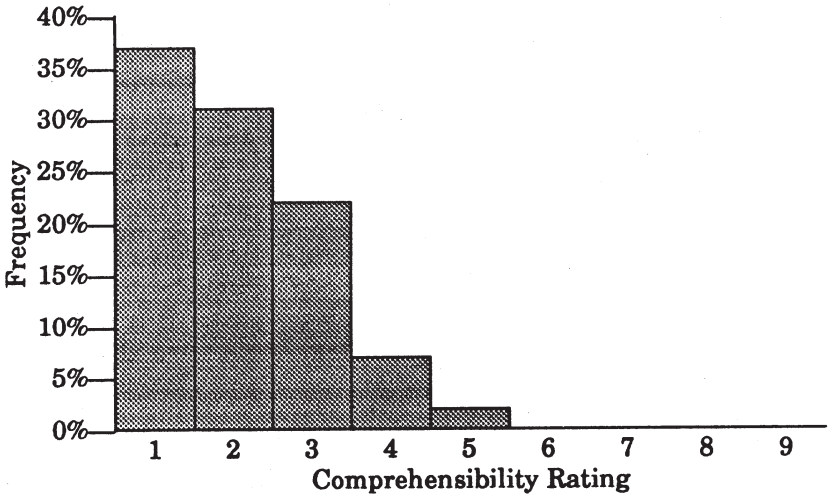

Figure 4. Distribution of perceived comprehensibility scores for the five stimuli transcribed orthographically without error by all listeners.

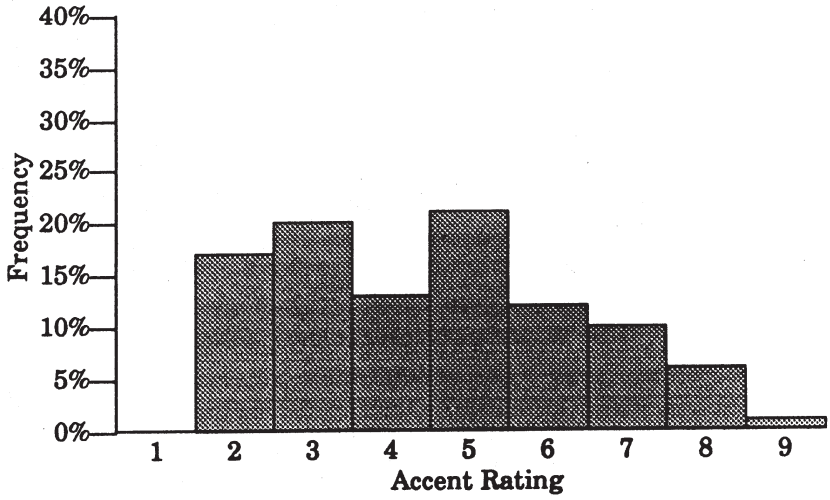

Figure 5. Distribution of foreign accent scores for the five stimuli transcribed orthographically without error by all listeners.

Table 2

Number of Significant Correlations Between Perceived Comprehensibility, Accent, and Intelligibility Scores and a Stimulus Measure ($p<.05$)

Stimulus Measure	Perceived Comprehensibility		Accent		Intelligibility (Words Correct)	
	No.	%	No.	%	No.	%
Phonemic Errors	8	44	14	78	5	28
Phonetic Errors	2	11	13	72	0	0
Intonation	15	83	16	89	4	22
Grammatical Errors	10	56	14	78	3	17
Utterance Length	0	0	0	0	0	0

scores. In fact, none showed such a relationship for the phonetic scores. Finally, utterance length did not correlate with any of the scores. Apparently, the stimuli were of suitable length for the listeners to make the required judgments and perform the orthographic transcription task. Had some of the utterances been too long, we would have expected some significant correlations with utterance length.

We also examined intercorrelations among the stimulus assessments, as shown in Table 3. Correlations significant at $p<.05$ are marked with an asterisk. The number of grammatical errors was significantly correlated with both the number of phonemic errors and the number of phonetic errors. In addition, the intonation ratings were correlated with phonemic error scores. In general, speakers who made grammatical errors also tended to make pronunciation errors. The surprising lack of correlations between phonemic and phonetic errors and phonetic errors and intonation suggests that errors in each of these categories were independent of one another.

Table 3

Intercorrelations (Pearson r) of Stimulus Characteristics

	Phonetic	Intonation	Grammar
Phonemic	.22	.39*	.48*
Phonetic		.23	.39*
Intonation			.28

*$p<.05$

Discussion

A group of native English listeners transcribed and rated for comprehensibility and foreign accent a set of speech samples produced by 10 proficient ESL learners. Overall, they found the nonnative stimuli to be highly intelligible. In fact, more than half of the transcriptions received intelligibility scores of 100%, and many others contained only minor errors. Although the utterances also tended to be highly rated in terms of perceived comprehensibility, the range of scores on the accent rating task was quite wide, with a noteworthy proportion in the "heavily accented" range.

There are a number of reasons to suppose that the three types of scores under consideration here correspond to related but partially independent dimensions. Evidence for a relationship among the three dimensions comes from the fact that most listeners showed a correlation between intelligibility and perceived comprehensibility and between perceived comprehensibility and accent. The latter observation parallels Varonis and Gass' (1982) findings. Varonis and Gass had observed high correlations (r=.81 to .90) between judgments of comprehensibility on a 5-point scale and the total number of times a particular utterance was judged to reflect good pronunciation in a binary (good/bad) judgment task. Our results confirm this finding for extemporaneous speech. However, we found a number of important differences as well. First, the distributions of perceived comprehensibility and accent scores were noticeably different; the listeners tended to assign harsher scores when rating accent. Second, the strength of the correlation

among any of the three possible pairings of dimensions tended to be in the moderate range for most listeners; there was not one perfect correlation. Third, far fewer listeners showed a significant correlation between intelligibility and accent than between intelligibility and perceived comprehensibility. The accent scores were a much poorer reflection of the listeners' actual comprehension of an utterance than were the perceived comprehensibility scores.

We found a fourth important difference when we examined a subset of the data. The listeners sometimes rated utterances as moderately or heavily accented even when able to transcribe them perfectly. This finding demonstrates empirically that the presence of a strong foreign accent does not necessarily result in reduced intelligibility or comprehensibility.

The intelligibility scores were the most direct test of what the listeners actually understood, because they indicated which words in each utterance the listeners had correctly identified. We assessed comprehensibility and accent, however, by asking listeners to make judgments on a 9-point scale. The lack of complete congruence between intelligibility and perceived comprehensibility was probably due to factors that the listeners took into account when making comprehensibility judgments but that did not necessarily determine whether an utterance was fully understood. It seems reasonable to speculate that processing difficulty may have played a role. For instance, two foreign-accented utterances may both be fully understood (and therefore be perfectly intelligible), but one may require more processing time than another.[5] Alternatively, special top-down processing may be required if an initially unintelligible word or phrase becomes transparent when the meaning of the rest of the utterance is clear. The need to allocate extra processing resources to an utterance might cause a listener to assign a lower comprehensibility score.

How the listeners made the accentedness judgments is less clear. Presumably they assessed the extent to which the pronunciation of each utterance deviated from some notion of what a native-like version would be. The foreign accent scores did not predict intelligibility very well. Perhaps, when judging accentedness,

listeners were primarily influenced by variables that caused the speech samples to sound deviant but that ultimately had little impact on whether the message was understood. For instance, the majority of the listeners showed significant correlations between accentedness and our assessments of phonemic errors, phonetic errors, and goodness of intonation. However, fewer showed correlations between perceived comprehensibility and these measures, and very few showed correlations with intelligibility. Varonis and Gass (1982) showed that pronunciation judgments can also be influenced by nonphonological properties such as grammatical errors. Our correlation between the grammatical error counts and the accent scores seems to confirm their finding, though it may simply indicate that speakers who make pronunciation errors also tend to make grammatical errors.

Varonis and Gass (1982) hypothesized that the main factor involved in judgments of pronunciation is "overall comprehensibility or ease of interpretation" (p. 127). Although our data do not provide grounds for rejecting this hypothesis, we do believe that it needs some qualification. Their results were based on binary (good/bad) judgments of accent that were pooled over listeners. We believe that our method of assessing both perceived comprehensibility and accent with a 9-point rating scale permits a more appropriate comparison between the two data sets. We also compared the rating data with actual transcription scores. In this way we obtained listeners' perceptions of both comprehensibility and degree of accent, as well as a measure of their actual level of understanding (intelligibility). In addition, we made quantitative measures of some of the factors that may contribute to accent. Unlike many other studies, ours used extemporaneous utterances rather than excerpts from reading passages or sentence stimuli. As a result, we examined accent and intelligibility under circumstances that better reflect naturally occurring speech. Thus, this study addresses the relationship between accent and intelligibility more directly.

Nonetheless, we plan to assess perceived comprehensibility, intelligibility, and accentedness in controlled utterances, using

identical scales, in future research. Our findings suggest that the role of comprehensibility in accent judgments varies from listener to listener and that accent scores cannot be relied upon as a means of assessing comprehensibility. Moreover, accent scores are poorer indicators of intelligibility than are perceived comprehensibility scores.

Implications for Second Language Teaching and Research

These findings have important implications for pronunciation assessment and instruction for adult second language learners. As far as we know, these are the first experimental data demonstrating what pronunciation experts have long believed: Although strength of foreign accent is indeed correlated with comprehensibility and intelligibility, a strong foreign accent does not necessarily cause L2 speech to be low in comprehensibility or intelligibility. Given this finding, it makes little sense to assess pronunciation on scales of the type that range from *not accented, perfectly comprehensible* at one endpoint to *accented and difficult to understand* at the other. Rather, scales of accent, perceived comprehensibility, and intelligibility ought not to be confused with one another. The nature of the scale to be used in assessment should be determined according to the goals of the instructor and the learner. If comprehensibility and intelligibility are accepted as the most important goals of instruction in pronunciation, then the degree to which a particular speaker's speech is accented should be of minor concern, and instruction should not focus on global accent reduction, but only on those aspects of the learner's speech that appear to interfere with listeners' understanding.

This raises two problems for those who teach pronunciation to second language learners. First, at present little empirical evidence indicates which particular aspects of foreign-accented speech are most detrimental to comprehensibility and intelligibility. As a result, instructors are left without much guidance as to what to teach (or how to teach it, cf. Macdonald, Yule, & Powers, 1994). Second, there are individual differences in the perception

of nonnative speech. Although our listeners tended to agree among themselves in their judgments, there were also important individual differences in the relationships among accentedness and comprehensibility ratings and intelligibility scores. It follows that opinions of a particular speaker's most serious pronunciation problems may vary from listener to listener. There are a number of possible explanations for the variability in this study. First, individual listeners may have interpreted the instructions differently. Some, for instance, may have focused more on the syntactic properties of the stimuli than others.[6] Second, familiarity with accented speech may have influenced some listeners' results (cf. Gass & Varonis, 1984; Wingstedt & Schulman, 1987). Only one listener reported having any regular contact with Mandarin speakers (that person's orthographic transcription score was well below the mean); however, of the six people who reported having fairly frequent contact with more than one other accent, five had orthographic transcription scores above the mean. As pointed out earlier with respect to NSs, individuals may vary in terms of rate of speech, speech clarity, voice quality, word choice, and so forth. All of these variables affect the comprehensibility of NNSs' speech as well. Finally, irrespective of differences in experience with L2 speech, there are probably individual differences in the ability to comprehend it.

Clearly, we need further studies of those aspects of L2 pronunciation that have the greatest impact on intelligibility. This study dealt only with one variety of accent (Mandarin), and the samples were elicited from individuals who were all proficient in English. Studies that include a variety of accents produced by speakers with differing levels of proficiency, and that give attention to differences among raters, should help to elucidate the relative contributions to intelligibility of specific elements (subsegmental, segmental, prosodic) of pronunciation. For instance, our study shows that intonation figures importantly in listener judgments of comprehension and accent, at least for Mandarin speakers of English. In addition, a recent study by Anderson-Hsieh et al. (1992) has provided promising empirical evidence in favor of

prosody as a factor in the intelligibility of L2 speech, but this work must still be regarded as preliminary, given that a clear distinction between accent and intelligibility was not made. Theoretical analysis by Catford (1987) on functional load in English may provide a direction for future studies at the segmental level. We ourselves plan to explore this issue in more detail by examining how accent and intelligibility are related to other variables, such as processing time and subjective listener reactions to nonnative pronunciation.

Notes

[1]Gass and Varonis, however, did not use the term "intelligibility scores" to refer to their assessments.
[2]One of the authors (MJM) verified this by transcribing (in standard orthography) the complete set of utterances.
[3]Inclusion of the ratings of the native speaker samples might have led to spuriously high correlations, because all but one of these samples received very good ratings on both scales.
[4]It should be noted that an examination of the data from a slightly different perspective led to very similar results. When we considered all the transcriptions which received intelligibility scores of 100% and examined the comprehensibility and accent ratings from the relevant listeners, distributions very similar to the ones in Figures 4 and 5 were observed. Here we report the results of only the more stringent analysis.
[5]In a detailed follow-up to the present study, we demonstrate (Munro & Derwing, 1995) that foreign-accented speech requires greater processing time than native speech.
[6]In debriefing, some listeners asked whether or not they should have taken grammatical errors into account.

References

Abercrombie, D. (1949). Teaching pronunciation. *English Language Teaching*, *3*, 113–122.

Albrechtsen, D., Henriksen, B., & Færch, C. (1980). Native speaker reactions to learners' spoken interlanguage. *Language Learning*, *30*, 365–396.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, *42*, 529–555.

Anisfeld, M., Bogo, N., & Lambert, W. (1962). Evaluational reactions to accented English speech. *Journal of Abnormal Social Psychology*, *69*, 89–97.

Barefoot, S., Bochner, J., Johnson, B., & von Eigen, B. (1993). Rating deaf speakers' comprehensibility: An exploratory investigation. *American Journal of Speech-Language Pathology, 2,* 31–35.

Brennan, E. M., & Brennan, J. S. (1981a). Accent scaling and language attitudes: Reactions to Mexican American English speech. *Language & Speech*, *24*, 207–221.

Brennan, E. M., & Brennan, J. S. (1981b). Measurements of accent and attitude toward Mexican-American speech. *Journal of Psycholinguistic Research*, *10*, 487–501.

Brodkey, D. (1972). Dictation as a measure of mutual intelligibility: A pilot study. *Language Learning*, *22*, 203–220.

Catford, J. C. (1987). Phonetics and the teaching of pronunciation: A systemic description of English phonology. In J. Morley (Ed.), *Current perspectives on pronunciation: Practices anchored in theory* (pp. 83–100). Washington, DC: TESOL.

Crawford, W. W. (1987). The pronunciation monitor: L2 acquisition considerations and pedagogical priorities. In J. Morley (Ed.), *Current perspectives on pronunciation: Practices anchored in theory* (pp. 101–121). Washington, DC: TESOL.

Derwing T., & Munro, M. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19,* 1–16.

Derwing T., Munro, M., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48,* 393–410.

Ensz, K. Y. (1982). French attitudes toward typical speech errors of American speakers of French. *The Modern Language Journal, 66,* 133–139.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, *37*, 313–326.

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

Gilbert, J. (1980). Prosodic development: Some pilot studies. In R. C. Scarcella & S. D. Krashen (Eds.), *Research in second language acquisition: Selected papers of the Los Angeles Second Language Acquisition Research Forum* (pp. 110–117). Rowley, MA: Newbury House.

Gimson, A. C. (1970). *An introduction to the pronunciation of English* (2nd ed.). London: E. Arnold.

Griffen, T. (1991). A non-segmental approach to the teaching of pronunciation. In A. Brown (Ed.), *Teaching English pronunciation: A book of readings*

(pp. 178–190). London: Routledge. (Reprinted from *Revue de Phonétique Appliquée*, *54*, 81–94, 1980).

Gynan, S. N. (1985). Comprehension, irritation, and error hierarchies. *Hispania*, *68*, 160–165.

Johansson, S. (1978). Studies in error gravity: Native reactions to errors produced by Swedish learners of English (*Gothenburg Studies in English, 44*). Gothenburg, Sweden: University of Gothenburg, Department of English.

Kalin, R., & Rayko, D. S. (1978). Discrimination in evaluative judgments against foreign-accented job candidates. *Psychological Reports*, *43*, 1203–1209.

Koster, C. J., & Koet, T. (1993). The evaluation of accent in the English of Dutchmen. *Language Learning*, *43*, 69–92.

Lambert, W. E., Hodgson, R., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *Journal of Abnormal Psychology*, *60*, 44–51.

Lane, H. (1963). Foreign accent and speech distortion. *Journal of the Acoustical Society of America*, *35*, 451–453.

Macdonald, D., Yule, G., & Powers, M. (1994). Attempts to improve English L2 pronunciation: The variable effects of different types of instruction. *Language Learning*, *44*, 75–100.

Morley, J. (1991). The pronunciation component of teaching English to speakers of other languages. *TESOL Quarterly*, *25*, 481–520.

Munro, M. J., & Derwing, T. M. (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing*, *11*, 253–266.

Munro, M., & Derwing, T. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*, 289–306.

Palmer J. (1976). Linguistic accuracy and intelligibility. In *Proceedings of the 4th International Congress of Applied Linguistics* (pp. 505–513). Stuttgart, Germany: Hockschul Verlag.

Pennington, M. C., & Richards, J. C. (1986). Pronunciation revisited. *TESOL Quarterly*, *20*, 207–225.

Politzer, R. L. (1978). Errors of English speakers of German as perceived and evaluated by German natives. *Modern Language Journal*, *62*, 253–261.

Ryan, E. B., & Carranza, M. A. (1975). Evaluative reactions of adolescents toward speakers of standard English and Mexican American accented English. *Journal of Personality & Social Psychology*, *31*, 855–863.

Sato, C. J. (1991). Sociolinguistic variation and language attitudes in Hawaii. In J. Cheshire (Ed.), *English around the world: Sociolinguistic perspectives* (pp. 647–663). Cambridge: Cambridge University Press.

Schairer, K. (1992). Native speaker reaction to non-native speech. *Modern Language Journal*, *76*, 309–319.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Varonis, E. M., & Gass, S. M. (1982). The comprehensibility of nonnative speech. *Studies in Second Language Acquisition*, *4*, 114–136.

Wingstedt, M., & Schulman, R. (1987). Comprehension of foreign accents. In W. Dressler (Ed.), *Phonologica 1984: Proceedings of the Fifth International Phonology Meeting, Eisenstadt, 25–28 June 1984* (pp. 339–344). Cambridge: Cambridge University Press.