The Modern Language Journal

Suprasegmental Measures of Accentedness and Judgments of Language Learner Proficiency in Oral English

OKIM KANG
Northern Arizona University
Department of English
Liberal Arts Building 18
Room 140
Flagstaff, AZ 86011-6032
Email: okim.kang@nau.edu

DON RUBIN
University of Georgia
Language and Literacy Education
125 Aderhold Hall
Athens, GA 30602
Email: drubin@uga.edu

LUCY PICKERING
Georgia State University
Applied Linguistics & ESL
P.O. Box 4099
Atlanta, GA 30302-4099
Email: esllup@langate.gsu.edu

In high-stakes oral proficiency testing as well as in everyday encounters, accent is the most salient aspect of nonnative speech. Prior studies of English language learners' (ELLs') pronunciation have focused on single parameters of English, such as vowel duration, fundamental frequency as related to intonation, or temporal measures of speech production. The present study addresses a constellation of suprasegmental characteristics of nonnative speakers of accented English, combining indices of speech rate, pause, and intonation. It examines relations between those acoustic measures of accentedness and listeners' impressions of second-language oral proficiency. Twenty-six speech samples elicited from iBT TOEFL® examinees were analyzed using a KayPENTAX Computerized Speech Laboratory. Monolingual U.S. undergraduates ($n = 188$) judged the speakers' oral proficiency and comprehensibility. A multiple regression analysis revealed the individual and joint predictiveness of each of the suprasegmental measures. The innovative aspect of this study lies in the fact that the multiple features of accentedness were measured via instrumentation rather than being rated by judges who may, themselves, be subject to rating biases. The suprasegmental measures collectively accounted for 50% of the variance in oral proficiency and comprehensibility ratings, even without taking into consideration other aspects of oral performance or of rater predilections.

THE CONSTRUCTS OF COMPREHENSIBILITY and accentedness relate in complex ways to native speaker (NS) judgments of English language learners' (ELLs') oral proficiency. No clear isomorphism has been established between degree of accentedness and comprehensibility. Speakers who succeed in reducing the degree of "foreignness" in their accents (based on expert observer judgments) may still be heard as incomprehensible by lay listeners (Munro & Derwing, 1995).

The Modern Language Journal, 94, 4, (2010)
DOI: 10.1111/j.1540-4781.2010.01091.x
0026-7902/10/554–566 $1.50/0

Comprehensibility is generally regarded as the listener's ability to understand the meaning of an utterance in its context (J. Jenkins, 2002; Smith & Nelson, 1985). However, there is no clear consensus on how the construct should be construed, especially in contrast to the related construct of intelligibility (Field, 2005; Isaacs, 2008). In most cases, comprehensibility measurement relies on "expert" NS listener ratings (Piske, MacKay, & Flege, 2001), whereas *intelligibility* is measured by the listener's ability to accurately transcribe the speaker's utterance (J. Jenkins, 2000; Smith & Nelson, 1985). In the present study, we are primarily concerned with broadly construed comprehensibility.

Several studies of comprehensibility have investigated its relation to NS listener judgments of foreign accent, but several methodological weaknesses in those studies demand remediation. For example, when comprehensibility is measured with only a single 7- or 9-point item (e.g., Derwing & Munro, 1997; Munro & Derwing, 1995), it is difficult to estimate the reliability of that measurement. Likewise, studies that sought to establish relations between accentedness and speaker proficiency level are vulnerable because judgments of speaker proficiency are so susceptible to extraneous biases. In fact, rater judgments of nonnative-speaker (NNS) oral proficiency may be affected by factors as diverse as familiarity with particular accents (Bent & Bradlow, 2003; Derwing & Munro, 2005; Rajadurai, 2007), listeners' attitudes to speakers' cultural group (Lippi-Green, 1997; Rubin, 1992; Rubin & Smith, 1990), and listener expectation based on generalized negative stereotypes (Lindemann, 2003). Yet another extraneous rater factor that can distort ratings of oral proficiency is embodied in the "interlanguage speech intelligibility benefit" (Bent & Bradlow). This benefit accrues to listeners from particular language backgrounds who manifest particular tolerance for certain NNS accents closest to their own. In other words, an NNS rater may privilege a second-language (L2) speaker from a similar first-language (L1) background and have a more difficult time understanding a speaker from a very different L1 background (cf. Major, Fitzmaurice, Bunta, & Balasubramanian, 2002). Perceptions of low oral proficiency may reflect a variety of rater biases and may have a limited basis in the linguistic characteristics of the speaker's vocal production. Therefore, a measurement strategy that could avoid the need for rater judgments of oral proficiency is much to be desired.

However, what objective measure could stand in as a more reliable and valid proxy than rater judgments of oral proficiency? The current consensus is moving toward an appreciation of the key role that suprasegmental differences in speaking rates, pitch patterns (i.e., intonation), and pausing phenomena play in comprehensibility and listeners' assessments—as opposed to segmental phonetic phenomena (Anderson-Hsieh, Johnson, & Koehler, 1992; Derwing & Munro, 1997, 2005). In other words, listeners can tolerate a great deal of inaccuracy in pronouncing consonants and vowels, so long as pitch and pausing are used appropriately.

Increasingly, researchers are using computer-assisted analysis to more objectively measure these suprasegmental elements of accent (e.g., Levis & Pickering, 2004; Pickering, 2001, 2004; Schuetze-Coburn, Shapley, & Weber, 1991; Wennerstrom, 1998) rather than relying on subjective (and possibly biased) judgments of accentedness by NS listeners. However, empirical relations between computer-derived measures of accentedness and listeners' impressionistic assessments of proficiency and comprehensibility of NNSs by NS listeners have not been widely investigated. Accordingly, the present study focused on relations between measurable acoustical features of speech and listeners' ratings of speaker oral proficiency and comprehensibility.

REVIEW OF LITERATURE

Nonnative accentedness may derive from several sources, including differences in producing individual phonetic segments as well as in sentence prosody. Derwing and Munro (1997) concluded that improvement in NNS comprehensibility (vis-à-vis NS listeners) "is more likely to occur with improvement in grammatical and prosodic proficiency than with a sole focus on correction of phonemic errors" (p. 15), and their position is mainly supported by subsequent studies (e.g., Derwing & Rossiter, 2003; Field, 2005).

Prosody in comprehensibility research usually includes speech rate, pausing, stress, and pitch patterns, or intonation. A relatively slow speech rate is commonly cited as a facilitating characteristic of foreign language discourse (Derwing, 1990). In fact, Derwing and Munro (2001) noted that the "ideal" rate of English production for English NS listeners of NNS speech is 4.1 syllables per second, compared to 4.7 for NS production. A slow rate of production enables increased time for listener processing and clearer grammatical boundary markers—relative to a rapid rate of speech (Chaudron, 1988). A relatively slow speaking rate appears to be especially critical for the comprehension of speech that NSs find highly accented (e.g., Chinese-accented English; Anderson-Hsieh & Koehler, 1988). Recently, Kormos and Denes (2004) observed that speech rate—along with mean length of utterance and phonation time ratio (i.e., the percentage of time within the speech sample actually spent producing speech sounds)—appeared to be the best predictors of NSs' perceptions of NNSs' fluency.

Pause length and pause placement are also related to comprehensibility. In studies of pause structure in Dutch and in English, pause length correlated with the production and perception of grammatical unit boundary strength (Nakajima & Allen, 1993; Swerts & Geluykens, 1994).

Analyses of NNS speech, moreover, have shown a qualitative difference in both placement and length of pauses, which can materially affect the overall prosodic structure of the discourse (Anderson-Hsieh & Venkatagiri, 1994; Pickering, 1999; Riggenbach, 1991). Pickering (1999), in a study of lectures given by NS teaching assistants and international teaching assistants (ITAs), found that silent pauses in the NNS data were both longer and more irregular than those in the NS data and tended to regularly break up conceptual units. This prevalence of *empty pauses* (irregular moments of silence unrelated to board work or to dramatic effect) may be linked to negative perceptions of ITAs on the part of U.S. undergraduate students (Rounds, 1987).

Nonstandard word stress can likewise undermine comprehensibility (Gallego, 1990). Lexical stress plays a central role in determining the profiles of words and phrases, and misplaced word stress appears to be more perceptually salient to NS listeners than are instances of mispronounced phonemes (Bond, 1999). Field (2005) tested both NS and NNS listeners in a psycholinguistic study in which he manipulated lexical stress as well as vowel quality on sets of disyllabic words. Intelligibility, as part of the broad comprehensibility construct, for both groups of listeners was significantly handicapped by the modified stress patterns, particularly when the lexical stress shifted to the right (i.e., to the second syllable).

Nonnative speakers' intonation likewise appears to be a key factor in NS listeners' misinterpretation of information structure or misunderstanding NNS speaker intent. Intonation is typically defined as the linguistically meaningful use of vocal pitch level and pitch movement in phrases. Current discourse-based, componential models of the intonation structure of English (Brazil, 1997; Tench, 1996) identify significant pitch choices on the tonic syllable or focus word of each thought group. Native English speakers choose a rising tone on key syllables to reflect new (or unrelated) information. They choose falling or level tones to reflect given (or related) information, or a suspension of the current informational context. Each tone is also assigned a pragmatic meaning within the context of the discourse. Falling tones indicate a speaker "telling" something to a hearer, and rising tones indicate that the speaker is "reminding" the hearer of something. The choice of tone on the focus word can affect both perceived information structure and social cues in L2 discourse.

For example, Wennerstrom (1994) determined that Japanese, Thai, and Chinese speakers of English displayed low, falling tones at boundaries between related propositions, where NS listeners would anticipate rising or mid-level tones (see also Pirt, 1990, for similar findings in a study of Italian learners). A preference for falling tones is common in the speech of ELLs from Korea, Greece, and Indonesia (Hewings, 1995). Uniformly falling tones can be problematic in certain contexts, however, because NSs commonly use rising tones to avoid impressions of rudeness or animosity, for instance, to reduce face threats when expressing disagreement. Pickering (2001) also uncovered critical differences in tone choices when comparing ITAs with U.S. instructors. Whereas NS teachers oriented their tonal structure toward a state of informational and social convergence with their student listeners, ITAs failed to exploit the English tonal system to increase comprehensibility or to show involvement.

In addition to those violations of norms for English pitch *movement*, L2 learners manifest problems with pitch *level*. One unit of discourse defined by pitch level is analogous to a written paragraph and is variously labeled as a speech paragraph, intonational paragraph, or paratone. Paratones are delineated by an extrahigh pitch at the beginning of the unit (sometimes called the *reset*) and a gradual fall in pitch level to a low termination. Paratone structure is a cue for the hearer as to the informational structure of the discourse (Pickering, 2004; Wennerstrom, 2001). Typically, NNSs tend to manifest a narrow and compressed overall pitch range in comparison with NSs (Mennen, 1998; Pickering, 2004). This constricted pitch range limits NNSs' ability to indicate paratone boundaries, especially their onset (Cutler, Dahan, & Donselaar, 1997; Swerts & Geluykens, 1994; Yule, 1980). Accordingly, NS listeners may fail to discern rhetorically significant boundaries in NNS utterances. Indeed, NNSs' ability to differentiate rhetorical units by expanding their pitch range is predictive of their rated oral proficiency in English (Wennerstrom, 1998).

In sum, much previous research involving acoustic correlations of intelligibility or comprehensibility of NNS pronunciation has relied on potentially subjective raters rather than on computer-assisted instrumentation for measuring degree of accentedness. Candidates for acoustic variables that predict oral proficiency include such suprasegmental factors as rate, pause length and location, lexical stress, and various aspects of pitch. However, there have been few attempts to investigate the conjoint impact of these suprasegmental features on comprehensibility and

proficiency judgments. The current study was designed to fill that gap.

METHODOLOGY

Speech Samples

Twenty-six NNS speech samples were collected under high-stakes examination conditions. They were responses to an iBT TOEFL® integrated task that required examinees to respond for 60 seconds to a question that asked them to summarize and demonstrate understanding of a passage they had just read. To control for extraneous factors, only male voices were used in this study. To allow for generalization across L1s, four groups of speakers were sampled: Chinese (6), Spanish (6), Korean (8), and Arabic (8).

NS Listeners

The participants were 188 undergraduates at a large university in the southern United States. The total sample size of listeners yielded .80 statistical power for medium effect size, based on Gatsonis and Sampson's (1989) calcuation. Participation in this study fulfilled a research requirement in an introductory course in speech communication.

Procedures

Ratings of the NNS speech samples were conducted entirely online. Raters listened to the randomly ordered speech samples as streaming audio files. A brief online rating tutorial was provided, but raters received no other training regarding interpretation of the scales. To reduce possible rater error attributable to task novelty, raters first listened to one trial speech sample that was not included in the analysis (Derwing, Rossiter, Munro, & Thompson, 2004).

Ratings of English oral proficiency utilized criteria developed for the iBT TOEFL® oral tasks. Raters assessed each sample in terms of pronunciation/accent, grammatical accuracy, vocabulary, rate of speech, organization, and how well the requirements of the test prompt (task) were met. These analytic ratings were recorded on 7-point scales in the Likert format, with 1 representing low proficiency, 4 representing moderate proficiency, and 7 representing high proficiency. The reliability (Cronbach's alpha) for the proficiency scale was .96 and therefore the sum of these analytic items was used as a composite measure in the final analysis.

The measure of comprehensibility was developed for this study and was comprised of five 7-point bipolar scales (e.g., "hard to understand : : : : : : : easy to understand"). The five items were the following: Easy/hard to understand, incomprehensible/highly comprehensible, needed little effort/lots of effort to understand, unclear/clear, and simple/difficult to grasp the meaning.

Reliability for the five-item comprehensibility measure was .94. Accordingly, the sum of these five items was utilized as a composite measure for subsequent analysis. The composite comprehensibility and proficiency measures correlated at $r = .83$ ($p < .01$).

Suprasegmental Variables

A comprehensive suprasegmental profile comprising 29 rate, pause, stress, and pitch measures was generated for each of the 26 speech samples. These 29 acoustical variables (see Table 1) were selected on the basis of precedent in prior research (Brazil, 1997; Derwing, 1990; Derwing & Munro, 2001; Hincks, 2005; Kormos & Denes, 2004; Levis & Pickering, 2004; Pickering, 2004; Wennerstrom, 2001; Wichmann, 2000).

As a first step in the acoustical analysis, the 26 speech samples were transcribed using the model of intonation structure in discourse proposed by Brazil (1997). The framework has been used extensively to transcribe a wide variety of NS and NNS English (e.g., Brazil, Coutlthard, & Johns, 1980; Cheng, 2004; Hewings, 1995; Pickering, 2001, 2004; Warren, 2006).

Next, speech samples were converted to digital .wav files and transferred to a KayPENTAX Model 5400 Computerized Speech Laboratory (CSL) for computer-assisted analysis of acoustical features. Three acoustic indicators were generated: (a) spectrograms (frequency and location of vocalization), (b) frequency or pitch of fundamental formant (F_0), and (c) intensity (volume of vocalization). From these three indicators, plotted against the transcripts of the speech samples, the 29 suprasegmental variables of interest were derived. As shown in Table 1, these variables are classified into measures of rate, pause, stress, pitch, and paratone.

Rate and pause measures required an examination of the number of unfilled and filled pauses, the number of syllables per second, and syllables per run. (A *run* of discourse was operationalized as a stretch of speech bounded by pauses of 100 milliseconds or longer.) The remaining suprasegmental measures shown in Table 1 were established following the division of the spoken

TABLE 1
Summary of Suprasegmental Measures

Measures	Submeasures	Descriptions
Rate Measures	Syllable per second	Calculated by counting the number of syllables produced in the sample and dividing by the 60-second sample.
	Articulation rate	Calculated similarly to syllables per second but excluding silent pause time.
	Mean length of run	Runs are identified as stretches of speech bounded by pauses of 100 milliseconds or longer. The length of a run is expressed in syllables, and the number of syllables was calculated and divided by the number of runs.
	Phonation time ratio	Calculated as the percentage of time within the 60-second sample spent speaking, including filled pauses.
Pause Measures	Number of silent pauses	Calculated by counting the number of silent pauses of 100 milliseconds and longer in the 60-second sample.
	Mean length of pauses	Calculated by dividing the total length of silent pause time by the number of silent pauses of 100 milliseconds and longer in the 60-second sample.
	Number of filled pauses	Calculated by counting the number of filled pauses in the 60-second sample. Filled pauses were defined narrowly as nonlexical fillers such as *um*, *uh*, *er*, and so on. Repetitions, restarts, and repairs were not included in this measure (Kormos & Denes, 2004).
	Mean length of filled pauses	Calculated by dividing the total length of filled pauses by the number of filled pauses in the 60-second sample.
Stress Measures	Number of prominent syllables per run (pace)	Calculated by counting the total number of prominent syllables and dividing them by the total number of runs.
	Proportion of prominent words (space)	Calculated as the percentage of the prominent words (i.e., those containing prominent syllables) out of the total number of words.
	Prominence characteristics	Calculated as the percentage of tone units out of the total number of tone units containing a final prominence or termination.
Pitch Measures	Overall pitch range	Calculated by measuring the F_0 maxima and minima and producing range in Hertz for each task.
	High-rising tone choice	Calculated by identifying tone (rising, falling, or level) and termination (high, mid, or low) on tonic syllables.
	High-level tone choice	
	High-falling tone choice	
	Mid-rising tone choice	
	Mid-level tone choice	
	Mid-falling tone choice	
	Low-rising tone choice	
	Low-level tone choice	
	Low-falling tone choice	
	Pitch prominent syllable	Calculated by measuring the F_0 of five prominent and five nonprominent syllables and calculating the average F_0 value for each category.
	Pitch nonprominent syllable	
	Pitch new information lexical item	Calculated by measuring the F_0 of the same lexical item presented initially as new information and thus appearing in following instances as given information. Where possible, five lexical items were used to calculate the average F_0 for each category.
	Pitch given information lexical item	
Paratone Measures	Number of low termination tones	Calculated by counting the total number of low terminations followed by high-key resets.
	Avg. height of onset pitch	Calculated by averaging the pitch of high-key onsets.
	Avg. height of terminating pitch	Calculated by averaging the pitch of low terminations.
	Avg. paratone pause length	Calculated by averaging the length of pauses at paratone boundaries.

samples into tone units (Brazil, 1997). For each tone unit, one tonic syllable was identified. The tonic syllable is the prominent or stressed syllable in each phrase. The proportion of prominent syllables relative to total number of syllables was one tone variable. Other pitch measures included pitch levels of prominent syllables (high, medium, or low) and pitch movement within tone units (rising, level, or falling). Finally, paratone measures included the average F_0 level of paratone-initial pitch choices (usually high pitch) and the average F_0 of paratone termination choices (usually low pitch).

Although the computer-assisted acoustic analysis eliminates the problem of rampant rater bias, the process is by no means algorithmic. A certain amount of flexibility must remain within any system that attempts to describe suprasegmentals because they are conditioned by both time and position in the discourse (Beckman, 1997; Levelt, 1989; Vaissiere, 1995). Assigning pitch characteristics in this study required the analyst to fit the physical realization of the acoustic parameters with theoretically motivated categories of stress and pitch structure (Schuetze-Coburn et al., 1991). These parameters are gradient in nature, and for purposes of comparison, measurements were also taken of the range of baseline realizations of these significant features from three male NSs.

ANALYSIS

The data were analyzed primarily through a general multiple regression model. The general multiple regression model is a powerful statistical tool that enables us to ascertain both the conjoint and the unique contributions of several "predictor" variables on a "criterion," or outcome variable. The conjoint predictiveness of the entire model is reflected in the "total R^2" statistic. The unique relation of each individual predictor (i.e., purged of any overlap with other predictor variables) to the criterion variable is reflected in "partial correlation" statistics. Comprehensibility ratings and general oral proficiency ratings served as the two dependent or criterion variables in these regressions. The suprasegmental variables were the predictor variables.

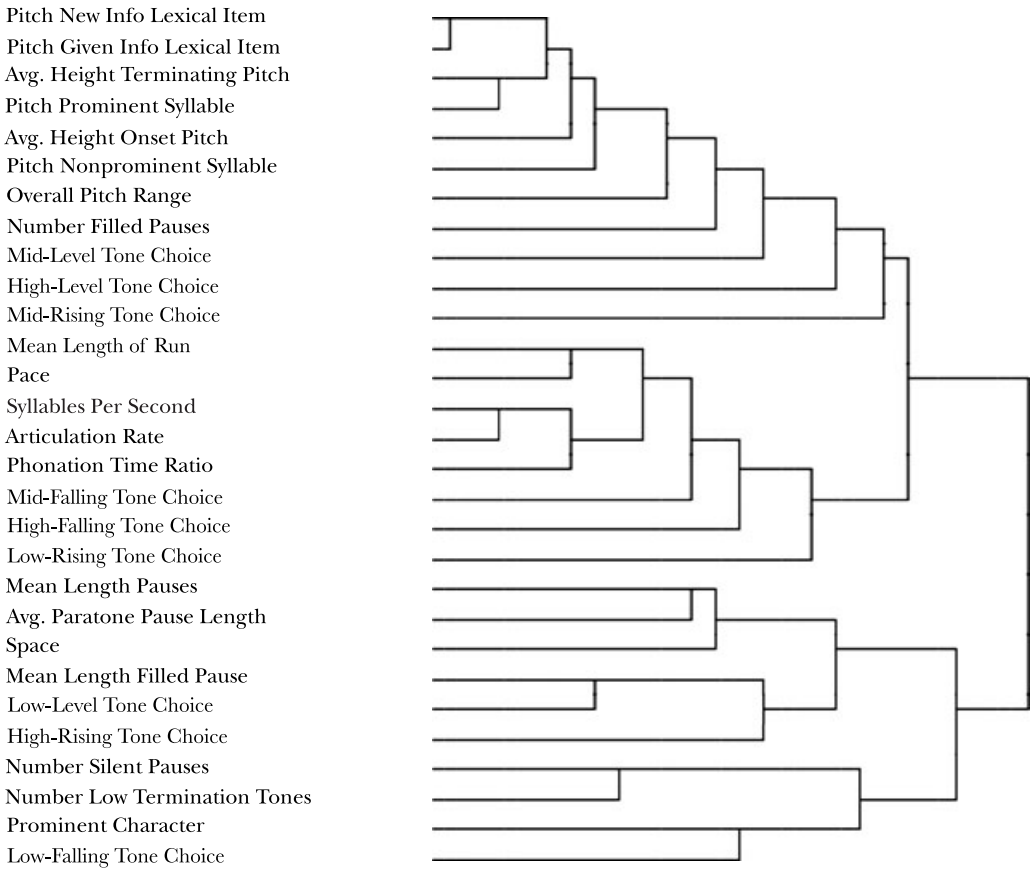
Because regression models involving 29 suprasegmental predictor variables would prove cumbersome and perhaps uninterpretable, it was desirable to conduct a preliminary statistical procedure to inductively group those acoustical indices into a smaller number of predictor variables. Hierarchical cluster analysis (HCA) was used for just that purpose. HCA is a method for finding relatively homogeneous clusters of variables. It starts with each variable in a separate cluster and then combines the cluster variables sequentially on the basis of proximity scores. Combining clusters reduces the number of clusters at each step. Proximity scores correspond to similar distributions of the variables. The function of HCA is thus analogous to factor analysis. Precedents for HCA in applied linguistics research include applications of cluster analysis to student activities and behaviors in classroom observation sessions (e.g., Gayle, 1980; Ross, 2001), where the ungainly matrix of observation data is reduced into a smaller subset of interpretable summaries of the similarities/dissimilarities among the observed classes. The method adopted for this study employed the *average linkage within groups* technique (i.e., SPSS "Within Groups" option), so as to optimize homogeneity within clusters of variables. All of the acoustic measures were first scaled to standardized scores. Thus, the scaled scores for each of the divergent acoustical measures all ranged from −1 to 1 and could be readily compared.

RESULTS

Hierarchical cluster analysis results are conventionally represented as dendrograms. Variables in a dendrogram are sequenced in terms of "distances," such that contiguous variables are relatively homogenous, whereas variables that are remote from each other are relatively heterogeneous. Each step in the clustering process appears as a node of the dendrogram tree. The dendrogram in Figure 1 depicts hierarchical clusters of acoustic markers along with nine nonclustered markers.

The goal of HCA is to arrive at the smallest set of clusters that are still meaningful. In this case, we proceeded five steps down from the top of the hierarchy (i.e., six hierarchical steps from just a single cluster of markers). Fewer than six steps resulted in uninterpretable supercategories. A greater number of steps down from the top of the hierarchy resulted in too many categories and defeated the goal of data reduction. The resulting clusters of acoustic features were as follows: (a) an "*um*" *factor* (low-level tone choice and mean length of filled pause); (b) *unit completeness* (proportion of prominent words or "space," average paratone pause length, and mean length of silent pauses); (c) *boundary marking* (number of silent pauses and low temination tones); (d) *pitch height* (average height of prominent and nonprominent

FIGURE 1
Hierarchical Cluster Analysis: Dendrogram Using Average Linkage (Within Group)



Note. The 29 distance scores may be obtained from the first author.

syllables, pitch levels of given and new lexical items and of paratone onsets and terminations, and overall pitch range); and (e) *suprasegmental fluency* (mid-falling tone choices, number of prominient syllables per run or "pace," mean length of runs, phonation time ratio, articulation rate, and syllables per second). Nine suprasegmental features failed to cluster in the HCA and therefore each had to be treated as a predictor in its own right. These nonclustered predictors were as follows: (f) high-, (g) mid-, and (h) low-rising tones, (i) high- and (j) mid-level tones, (k) high- and (l) low-falling tones, (m) number of filled pauses, and (n) prominent characteristics in stress measures. This clustering, based on distance scores, indicated that most of the tone movement choices were distributed rather independently (high distances) of other acoustic features.

Suprasegmental Predictors of Oral Proficiency Ratings

A multiple linear regression analysis was conducted in which the five clustered and nine nonclustered suprasegmental variables were regressed on composite oral proficiency ratings. Fifty-two percent of the variance in the oral proficiency ratings could be accounted for by the 14 suprasegmental features in concert ($F_{13,3609} = 269.18$; $p < .001$). As Table 2 shows, several variables emerged as significant individual predictors.

The suprasegmental fluency cluster, high-rising pitch, and mid-rising pitch were the best individual predictors of rated English language proficiency. Each of their partial correlations was positive, indicating directly proportional relations with oral proficiency. It is interesting to note

TABLE 2
Linear Regression of 14 Acoustical Variables on Oral Proficiency Ratings

Total $R^2 = .51$	Standardized Coefficient Beta	t	Significance	Zero-Order Corr.	Partial Corr.
"Um" Factor	0.01	0.23	.82	−.34	.00
Unit Completeness	0.27	3.59	.00	−.57	.06
Boundary Marker	0.32	6.77	.00	.38	.11
Pitch Height Factor	−0.10	−3.18	.00	.23	−.05
Suprasegmental Fluency	1.27	14.23	.00	.62	.23
Number of Filled Pauses	−0.07	−4.18	.00	.07	−.07
Mid-Rising Tone Choice	0.45	10.85	.00	−.21	.18
High-Level Tone Choice	0.12	2.83	.01	.04	.05
Mid-Level Tone Choice	0.04	2.36	.02	.05	.04
Low-Rising Tone Choice	−0.16	−2.43	.02	.23	−.04
High-Falling Tone Choice	0.24	3.89	.00	.37	.07
High-Rising Tone Choice	0.24	12.56	.00	−.22	.21
Low-Falling Tone Choice	−0.01	−.12	.90	−.17	.00
Prominent Character	0.33	7.58	.00	−.33	.13

that if one simply examined zero-order correlations (i.e., correlations that include shared as well as unique variances), it would appear that high-rising pitch and mid-rising pitch were negatively correlated with oral proficiency, but the partial correlations show us that their unique contributions were positive. The same pattern emerged for prominent word stress. Unit completeness and high-falling tone choice both exhibited minimal positive unique contributions to oral proficiency ratings, and pitch height exerted a slightly (but significantly) inverse effect.

Suprasegmental Predictors of Comprehensibility

A multiple regression model was likewise computed to determine the effects on composite comprehensibilty ratings of the five clustered and nine nonclustered acoustic features. Table 3 reveals results of this analysis. The 14 acoustical variables acting in concert ($F_{14,3609} = 252.97$, $p < .001$) accounted for 50% of the variance in the measure of comprehensibility.

Not surprisingly, given the high correlation between proficiency ratings and comprehensibility, the overall pattern of significant predictors was similar in terms of partial correlations (each predictor's unique contribution to comprehensibility). Suprasegmental fluency and high-rising tones were the strongest predictors, and mid-rising tones and boundary markers followed. Prominent word stress was directly proportional to comprehensibility, but pitch height was inversely related; that is, the higher pitch speakers evinced, the less comprehensible they were perceived by raters.

DISCUSSION

The purpose of this study was to determine relations between suprasegmental features of accentendness and subjective assessments of NNS oral proficiency and comprehensibility. The study explored the conjoint impact on NSs' judgment of NNSs' speech of a wide range of acoustically measured suprasegmental features of accent. The features investigated included indices of speech rate, pause patterns, and discourse-contingent stress, and intonational marking. Speech samples were from a high-stakes oral proficiency test.

Because this study was unprecedented in the range of acoustical analyses conducted, the hierarchical cluster analysis of the 29 suprasegmental variables was itself of interest. It revealed that the features clustered into five sets of suprasegmental markers ("*um*" *factor, unit completeness, boundary markers, pitch height factor,* and *suprasegmental fluency*) and nine nonclustered variables (high-, mid-, and low-rising tones; high- and mid-level tones; high- and low-falling tones; number of filled pauses; and prominence characteristics).

In prior studies of listeners' perceptions of prosodic cues in NS speech, listeners were found to locate major discourse boundaries using prosodic features such as pause length and pitch variation to predict when an utterance was likely to end (e.g., Cutler et al., 1997). In NNS speech, however, prosodic cues about discourse

TABLE 3
Linear Regression of 14 Acoustical Variables on Comprehensibility

Total $R^2 = .50$	Standardized Coefficients Beta	t	Significance	Zero-Order Corr.	Partial Corr.
"Um" Factor	−0.09	−1.92	.05	−.34	−.03
Unit Completeness	0.38	5.00	.00	−.52	.08
Boundary Marker	0.41	8.71	.00	.41	.14
Pitch Height Factor	−0.22	−6.75	.00	.15	−.11
Suprasegmental Fluency	1.26	13.91	.00	.59	.23
Number of Filled Pauses	−0.05	−2.90	.00	.02	−.05
Mid-Rising Tone Choice	0.37	8.74	.00	−.22	.14
High-Level Tone Choice	0.16	3.58	.00	−.01	.06
Mid-Level Tone Choice	0.01	.77	.44	.01	.01
Low-Rising Tone Choice	−0.23	−3.45	.00	.23	−.06
High-Falling Tone Choice	0.14	2.31	.03	.35	.04
High-Rising Tone Choice	0.29	10.65	.00	−.20	.18
Low-Falling Tone Choice	−0.15	−2.20	.03	−.12	−.04
Prominent Character	0.29	6.76	.00	−.30	.11

boundaries may be less distinct, due in part to a narrower overall range of pitch (Mennen, 1998; Pickering, 1999). Consequently, ineffective prosodic structuring in NNS discourse may contribute to confusion and misinterpretation of speaker intent. Pickering (1999) also reported that NNSs were unable to consistently use tone choices to create the units of organization found in the NS discourse. In the present study, as well, the HCA revealed that most (but not all) tone choices were nonclustered with any of the other suprasegmentals; that is, tonal features were simply not integrated into broader suprasegmental patterns, for the most part.

The most potent variable in both regression analyses was the suprasegmental fluency cluster. This cluster included all of the rate measures (syllables per second, articulation time, phonation time, and mean length of runs), one stress measure (pace, or average number of prominent syllables per run), and one intonation measure (mid-falling tones). The inclusion here of the rate measures concurs with previous studies investigating fluency (Anderson-Hsieh & Koehler, 1988; Freed, 2000; Riggenbach 1991, 2000). Similarly, the stress measure, pace, was previously shown to be a reliable predictor of fluency judgments (Kang, 2008; Kormos & Denes, 2004). The association of a mid-falling tone with a general fluency factor was not unexpected as, contextually, this tone signifies the addition of new information to an ongoing discourse context and, therefore, is related to ideational fluency. Mid-falling tones are the most common ones to appear in native English speaker discourse (Hewings, 1995; Pickering, 2001). Overall, just as Kormos and Denes argued, it seems that fluency is an intonational phenomenon as well as a temporal one.

Mid-rising and high-rising tones were also prominently and positively associated with both proficiency and comprehensibility judgments. In the context of discourse production, this is a plausible result. As Hewings (1995) and Wennerstrom (1994) found, NNSs tend to use low-falling tones between related propositions (i.e., to display a paucity of mid- and high-rising tones), whereas rising and mid-level tones would be anticipated by NS listeners. Overuse of falling tones by NNSs can convey to NSs negative impressions of speaker arrogance or overassertiveness (Gumperz, 1982) as well as erode comprehensibility. Rising tones, in contrast, can convey shared background between speaker and listener—both within the context of the discourse itself (i.e., the first part of a continuing utterance) and within the broader sociocultural context (i.e., what we can expect our listeners to already know as part of a shared context; Brazil, 1997). Thus NNSs' underutilization of rising tones may contribute to the impression that NNSs are oriented toward the language itself rather than toward their listeners (Pirt, 1990). The boundary-marking cluster (i.e., number of silent pauses and low termination tone choices) exhibited positive relations to comprehensibility and proficiency ratings. It is well known that NSs of English tend to use low pitch levels accompanied by longer pauses at topic-final boundaries, whereas they use high pitch levels at the initiation of a new topic and use middle levels at points of continuation (Nakajima & Allen, 1993). The

current finding is consistent with earlier research that NNSs' production of low termination tones facilitates the comprehension of discourse structure by NS listeners (Pirt, 1990). Certainly the second component of boundary marking (number of silent pauses) is likewise important for recognizing junctures between idea units. NNSs' pauses, according to previous studies, are more frequent, longer, and less regular than those of NSs (Anderson-Hsieh & Venkatagiri, 1994; Pickering, 1999; Riggenbach, 1991; Rounds, 1987). In the present results, however, sheer frequency of silent pauses created a positive impression among listeners, perhaps because silent pausing at least precluded filled pausing.

The pitch height factor comprised a cluster of pitch variation parameters that included the average height of prominent and nonprominent words, the average height of given and new lexical items, the average height of paratone onsets and terminations, and the overall pitch range. Given the particular array of constituents, we interpret this cluster to reflect individual differences in voice pitch, not the use of discourse-level pitch structure, as a cue to topic openings and closings or to distinguish the informational value of individual lexical items (Cutler et al., 1997). In other words, it is important that the present findings not be confused with the well-established conclusion that restricted pitch range may adversely affect NNSs' comprehensibility (Mennen, 1998; Pickering, 1999; Wennerstrom, 2000).

The present findings regarding pitch, we believe, speak to pitch height, not to pitch range, per se; that is, the pitch height cluster in this study reflects how high or low one's voice is pitched, and men with higher voices in our study fared worse than those with lower voices. The pitch height factor was inversely proportional to ratings of both proficiency and comprehensibility. Although this negative relation seems to be counterintuitive at first glance, it is understandable in light of the literature on speech style and social evaluation. Speakers' idiosyncratic vocal characteristics affect the way in which others interpersonally perceive and evaluate them (Giles & Powesland, 1975). For example, idiosyncratic characteristics, such as vocal "thinness," are perceived as an indication of social, physical, and mental immaturity (Addington, 1968). Listeners tend to rate deep-pitched speakers as more powerful, confident, and stronger than high-pitched speakers (Bradac, Cargile, & Hallett, 2001). Moreover, male and female voices with very high pitch (high F_0) come across as effeminate and immature (Tusing & Dillard, 2000). Thus, raters affected by these social perceptions may consider NNSs' high-pitch speech to be less proficient and difficult to comprehend.

It is interesting that the "um" factor showed little relation to either comprehensibility or oral proficiency. The "um" factor encompassed low-level tone choice and mean length of filled pause. These filled pause hesitation phenomena may reflect more on individual speaking style and cognitive load (Goldman-Eisler, 1968) than on language proficiency. Fulcher (1996) similarly noted that low- and high-proficiency learners create different impressions on listeners, not because their incidence of hesitation is different but because they hesitate for different reasons. In previous studies, by the same token, advanced and low-intermediate learners of English were not distinguishable on the basis of temporal feaures such as number of filled pauses or mean length of filled pauses (Kang, 2008; Kormos & Denes, 2004).

Perhaps the most important finding of this study pertains to the overarching potency of objectively measured suprasegmental features in accounting for ratings of oral proficiency and comprehensibility. Suprasegmental features accounted for about 50% of the variance in mainly naïve raters' assessments of oral proficiency ($R^2 = .51$) and comprehensibility ($R^2 = .50$). These effect sizes are extraordinarily large when one considers the variables not at all included among the predictors in this study—for example, level of vocabulary usage, task fulfillment, grammatical felicity, and the like. Additionally, of course, the present study did not at all measure the impact of the accuracy of segmental production (accuracy of consonant and vowel production).

There has been a debate concerning ranges about the relative salience of segmental versus suprasegmental production for oral proficiency (Goodwin, 2001). Although the present study did not examine realization of phonetic segmentals, its findings nevertheless lend support to the fundamental importance of suprasegmental patterns in comprehensibility (e.g., Derwing & Rossiter, 2003). Moreover, our conclusion about the salience of suprasegmentals is derived from analysis of speech in context (i.e., responses on the TOEFL oral proficiency examination) rather than from pronunciation of isolated sentences, as in much of earlier research.

However, what might account for the other 50% of the variance in comprehensibility that the present study did not? Some of the remaining variance in the assessment of oral performance might indeed be explained by accented realization of phonemes. However, systematic individual differences among raters (rater bias) no doubt also

contribute significant explanatory power. Further research is needed to determine the relative contribution to NNSs' speaking proficiency scores of different rater characteristics (such as experience with NNS speech or special training in linguistics) versus measurable features of speaker pronunciation.

CONCLUSION

Evaluations of NNS speech involves nonlinguistic factors (e.g., rater attitudes, contexts, identity) as well as numerous linguistic factors (grammar, lexis, phonetic accuracy, etc.). Moreover, because the speech signal is inherently ambiguous, it needs to be interpreted in a dynamic fashion, different from situation to situation (Hughes, 2004). Notwithstanding those caveats, the results of this study revealed that suprasegmental features alone can collectively account for about 50% of the variance in proficiency and comprehensibility ratings. Moreover, unlike some other studies that have relied on expert judgments to ascertain accentedness (e.g., Anderson-Hsieh & Koehler, 1988; Derwing & Munro, 1997), the present study demonstrated the utility of computer-assisted instrumental acoustic analysis for objectively measuring suprasegmental aspects of accent. Although the process can be laborious, it is the only way of avoiding the sort of tautological regression that uses human judgments of speech as the criteria for assessing bias in human judgments of speech. Furthermore, the results of the HCA of those acoustic/suprasegmental features confirmed that NNSs failed to integrate tonal patterns (i.e., relative height and movement of tones) into broader patterns of intonation, stress, or fluency. However, some coherent groupings of pronunciation features did appear. These included marking boundaries between idea units, conveying overall fluency, and manifesting height of vocal pitch.

Past research about speech characteristics of NNSs did identify certain differences between high-profiency and low-profiency speakers, or between NSs and NNSs, in the acoustic analysis of speech production. Nevertheless, the current study was novel in the sense that it utilized speech samples of actual high-stakes oral proficiency testing with an extensive collection of suprasegmental parameters, and it investigated their impact on L2 proficiency assessment. The results lend support to the view that suprasegmental errors contribute as much or more to perceived accentedness than do segmental errors (Anderson-Hsieh et al., 1992; Anderson-Hsieh & Koehler, 1988; Munro & Derwing, 1995).

The present study pointed to suprasegmental fluency, the use of mid-rising tone choices and high-rising tone choices, and control over high vocal pitch as especially potent in determining perceived proficiency and comprehensibility. Use of features that mark boundaries between idea units was also important. Together, these factors reflect coherence in the expression of thought.

In pronunciation instruction, improved comprehensibility, rather than phonological accuracy, is the most important goal (Celce-Murcia, Brinton, & Goodwin, 1996; S. Jenkins, 2000). The findings of the present study support approaches to comprehensibility instruction that focus extensively on prosody (e.g., Grant, 2001). Teaching and practicing suprasegmental aspects of production (e.g., intonation, stress, rhythm, rate, and volume) may result in meaningful enhancement in perceived oral proficiency. ELLs may do well to focus especially on pausing silently and using falling tones at the end of idea units, on maintaining fluency within runs (i.e., avoiding pauses within idea units), and on using rising tones appropriately to achieve sentence focus. At the same time, NNSs might be counseled to exploit the lower registers of their voices, as consistent pitch height appears to undermine rated proficiency.

REFERENCES

Addington, D. W. (1968). The relationship of selected vocal characteristics to personality perception. *Speech Monographs*, *25*, 492–503.

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.

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning*, *38*, 561–613.

Anderson-Hsieh, J., & Venkatagiri, H. (1994). Syllable duration and pausing in the speech of Chinese ESL speakers. *TESOL Quarterly*, *28*, 807–812.

Beckman, M. (1997). A typology of spontaneous speech. In Y. Sagisaka, N. Campbell, & N. Higuchi (Eds.), *Computing prosody* (pp. 7–26). New York: Springer.

Bent, T., & Bradlow, A. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, *114*, 1600–1610.

Bond, Z. (1999). *Slips of the ear: Errors in the perception of casual conversation*. San Diego, CA: Academic Press.

Bradac, J. J., Cargile, A. C., & Hallett, J. S. (2001). Language attitudes: Retrospect, conspect, and prospect. In W. P. Robinson & H. Giles (Eds.), *The new handbook of language and social psychology* (pp. 137–154). New York: Wiley.

Brazil, D. (1997). *The communicative value of intonation in English*. Cambridge: Cambridge University Press.

Brazil, D., Coutlthard, M., & Johns, C. (1980). *Discourse intonation and language teaching*. London: Longman.

Celce-Murcia, M., Brinton, D., & Goodwin, J. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge: Cambridge University Press.

Chaudron, C. (1988). *Second language classrooms*. Cambridge: Cambridge University Press.

Cheng, W. (2004). "Well thank you David for that question": The intonation, pragmatics and structure of Q & A sessions in public discourses. *Journal of Asia TEFL, 1*, 109–133.

Cutler, A., Dahan, D., & Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*, 141–201.

Derwing, T. M. (1990). Speech rate is no simple matter. *Studies in Second Language Acquisition*, *12*, 303–313.

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

Derwing, T. M., & Munro, M. J. (2001). What speaking rates do non-native listeners prefer? *Applied Linguistics*, *22*, 324–227.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*, 379–397.

Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning*, *13*, 1–17.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thompson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Testing*, *54*, 655–679.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*, 399–423.

Freed, B. F. (2000). Is fluency, like beauty, in the eyes of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243–265). Ann Arbor: University of Michigan Press.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*, 208–238.

Gallego, J. C. (1990). The intelligibility of three nonnative English speaking teaching assistants: An analysis of student-reported communication. *Issues in Applied Linguistics*, *1*, 219–237.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin*, *106*, 516–524.

Gayle, G. (1980). *A descriptive analysis of second language teaching styles in the oral approach*. Ottawa, Canada: University of Ottawa Press.

Giles, H., & Powesland, P. F. (1975). *Speech style and social evaluation*. London: European Association of Experimental Social Psychology.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press.

Goodwin, J. (2001). Teaching pronunciation. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (pp. 117–123, 130). Boston: Heinle & Heinle.

Grant, L. (2001). *Well said: Pronunciation for clear communication*. Boston: Heinle & Heinle.

Gumperz, J. J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press.

Hewings, M. (1995). The English intonation of native speakers and Indonesian learners: A comparative study. *IRAL, 3*, 251–265.

Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, *33*, 575–591.

Hughes, R. (2004). Testing the visible: Literate biases in oral language testing. *Journal of Applied Linguistics*, *1*, 295–309.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review*, *64*, 555–580.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics*, *23*, 83–103.

Jenkins, S. (2000). Cultural and linguistic miscues: A case study of international teaching assistant and academic faculty miscommunication. *International Journal of Intercultural Relations*, *24*, 477.

Kang, O. (2008). The effect of rater background characteristics on the rating of international teaching assistants speaking proficiency. *Spaan Fellow Working Papers*, *6*, 118–206.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145–164.

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System*, *32*, 505–524.

Lindemann, S. (2003). Koreans, Chinese, or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics*, *7*, 348–364.

Lippi-Green, R. (1997). *English with an accent: Language, ideology, and discrimination in the United States*. New York: Routledge.

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

Mennen, I. (1998). Second language acquisition of intonation: The case of peak alignment. *Chicago Linguistic Society*, *34*, 327–341.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech

of second language learners. *Language Learning*, *45*, 73–97.

Nakajima, S., & Allen, J. (1993). A study on prosody and discourse structure in cooperative dialogues. *Phonetica*, *50*, 197–210.

Pickering, L. (1999). *An analysis of prosodic systems in the classroom discourse of native speaker and nonnative speaker teaching assistants*. Unpublished doctoral dissertation, Gainesville, University of Florida.

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly*, *35*, 233–255.

Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes*, *23*, 19–43.

Pirt, G. (1990). Discourse intonation problems for nonnative speakers. In M. Hewings (Ed.), *Papers in discourse intonation* (pp. 145–156). Birmingham, AL: English Language Research.

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*, 191–215.

Rajadurai, J. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes*, *26*, 87–98.

Riggenbach, H. (1991). Towards an understanding of fluency: A microanalysis of nonnative speaker conversation. *Discourse Processes*, *14*, 423–441.

Riggenbach, H. (Ed.). (2000). Perspectives on fluency. Ann Arbor: University of Michigan Press.

Ross, S. (2001). Program-defining evaluation in a decade of eclecticism. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 167–196). Cambridge: Cambridge University Press.

Rounds, P. (1987). Characterizing successful classroom discourse for NNS teaching assistant training. *TESOL Quarterly*, *21*, 643–672.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, *33*, 511–531.

Rubin, D. L., & Smith, K. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of non-native English speaking teaching assistants. *International Journal of Intercultural Relations*, *14*, 337–353.

Schuetze-Coburn, S., Shapley, M., & Weber, E. G. (1991). Units of intonation in discourse: A comparison of acoustic and auditory analyses. *Language and Speech*, *34*, 207–234.

Smith, L., & Nelson, C. (1985). International intelligibility of English: Directions and resources. *World Englishes*, *4*, 333–342.

Swerts, M., & Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech*, *37*, 21–43.

Tench, P. (1996). *The intonation systems of English*. London: Cassell.

Tusing, K. J., & Dillard, J. P. (2000). The sounds of dominance: Vocal precursors of dominance during interpersonal influence. *Human Communication Research*, *26*, 148–171.

Vaissiere, J. (1995). Phonetic explanation for cross-linguistic prosodic similarities. *Phonetica*, *52*, 123–130.

Warren, M. (2006, February). *// TOtally CRAzy // ?but NEver mind // What is the communicative role of the level tone?* International Conference: La Comunicazione Parlata (Spoken Communication), University of Naples, Naples, Italy.

Wennerstrom, A. (1994). Intonational meaning in English discourse: A study of nonnative speakers. *Applied Linguistics*, *15*, 399–421.

Wennerstrom, A. (1998). Intonation as cohesion in academic discourse: A study of Chinese speakers of English. *Studies in Second Language Acquisition*, *42*, 1–13.

Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 102–127). Ann Arbor: University of Michigan Press.

Wennerstrom, A. (2001). *The music of everyday speech*. New York: Oxford University Press.

Wichmann, A. (2000). *Intonation in text and discourse*. London: Longman.

Yule, G. (1980). Speakers' topics and major paratones. *Lingua*, *52*, 33–47.