The role of objective measures of suprasegmental features in judgments of comprehensibility and oral proficiency in L2 spoken discourse

Okim Kang & Lucy Pickering

In the light of ongoing research into the importance of suprasegmental features in the assessment of non-native speaker discourse, our study uses samples from the iBT TOEFL test to investigate the relationship between NS listener judgments of non-native oral proficiency and comprehensibility and the results of a detailed profile of objectively-measured suprasegmental features. Our findings suggest that prosodic features play a significant role in listener ratings and we consider some of the possible teaching implications.

The role of suprasegmentals in spoken discourse

For some time now, our field has been concerned with the relative importance of segmental and suprasegmental features in judgments of comprehensibility and oral proficiency in English language learners. While proficiency is normally standardized through some kind of established testing, there is no universally agreed-upon definition of what constitutes the construct of comprehensibility. There is, however, a commonly drawn distinction between 'matters of form', usually referred to as intelligibility, and tested using orthographic transcription, and 'matters of meaning' describing the extent to which the message is meaningful in its context. Field (2003) suggests that comprehensibility includes both form and meaning, and it is this more broadly construed idea of comprehensibility that we are interested in here.

In a series of seminal studies, Tracy Derwing, Murray Munro and their associates have prioritized suprasegmental or prosodic components, and concluded that improvement in non-native speaker (NNS) comprehensibility for native speaker (NS) listeners "is more likely to occur with improvement in grammatical and prosodic proficiency than with a sole focus on correction of phonemic errors" (Derwing & Munro, 1997: 15). Prosodic features that researchers have found to be significant include speaking rate (Derwing & Munro, 2001), pause structure (Anderson-Hsieh & Venkatagiri, 1994), nonstandard word stress (Field, 2005), pitch range (Wennerstrom, 1998) and intonation structure (Pickering, 2001).

The comprehensibility and proficiency judgments that many of these studies rely on are the result of asking raters to listen to a language sample and assign to it a number on a scale (e.g., 1 = extremely easy to understand, 9 = extremely difficult to understand.) However, human raters come to the task with a host of possible biases. These may include partiality based on familiarity with different accents (Bent & Bradlow, 2003), listeners' attitudes to speakers' cultural heritage (Rubin, 1992), listener expectation based on ne.g.ative stereotypes (Lindemann, 2003), or listener bias due to attributions of a group membership (Kang & Rubin, 2009), to name just a few. Ideally then, some more objective measure would be preferred that could stand as a more reliable and valid substitute for rater judgments. Thus, our goals were twofold: First to create a suprasegmental profile of our NNS samples using computer assisted analysis and then to compare these profiles to comprehensibility and proficiency ratings given by NS raters.

Data

Our NNS data comprised twenty-six male responses (6 Chinese, 6 Spanish, 8 Korean and 8 Arabic) to an iBT TOEFL integrated task. Students responded to a question that asked them to summarize and demonstrate understanding of a passage they had just read. Each sample is 60 seconds in length. One hundred and eighty eight North American undergraduate students listened to the NNS samples. Using 7-point scales, the listeners rated each sample for oral proficiency in the areas of pronunciation/accent, grammatical accuracy, vocabulary, rate of speech, organization and how well the requirements of the test prompt were met. Additionally, the raters judged comprehensibility of the samples using five 7-point bipolar scales comprising the following:

Easy/hard to understand, incomprehensible/highly comprehensible, needed little effort/lots of effort to understand, unclear/clear, simple/difficult to grasp the meaning.

In order to test how well objective measures of suprasegmental features could predict the listener ratings, a comprehensive suprasegmental profile comprising 29 rate, pause, stress and pitch measures was generated for each of the 26 speech samples. These acoustic measures, listed in Table 1 and elaborated in Table 2, were selected on the basis of previous research in this area and represent the largest set of prosodic variables investigated at one time to date.

Measure	Sub-measures
Rate	Syllables per second Articulation rate Mean length of run
Pause	Phonation time ratio Number of silent pauses Number of filled pauses Mean length of pauses
Stress	Number of prominent syllables per run Proportion of prominent words Prominence characteristics
Pitch	Overall pitch range High, Mid and Low Falling, Rising & Level tone choices Pitch (non-) prominent syllables Pitch on new and given lexical items
Paratone	Number of low terminations Avg. height of onset pitch Avg. height of terminating pitch Avg. paratone pause length

Table 1. Summary of Suprasegmental Measures

Each language sample was transcribed using Brazil's (1997) model of intonation in discourse as this framework has now been used extensively to transcribe a wide variety of NS and NNS English. Speech samples were transferred to a KayPENTAX Model 5400 Computerized Speech Laboratory for computer-assisted analysis. Three acoustic indicators were generated: (a) spectrograms which were used to identify the pricese location of speech as opposed to extraneous noise and to reliably distinguish segmental and pause boundaries, (b) frequency or pitch of fundamental formant (Fo) analyses which were used to calculate all pitch measures, and (c) intensity or volume of speech analyses which were used in conjunction with Fo to identify stress measures and to confirm pause lengths.

Sub-measure	Descriptions
Syllables per second	Number of syllables produced divided by the 60-second sample
Articulation rate	Number of syllables excluding silent pause time
Mean length of run	Stretches of speech bounded by pauses of 100 milliseconds or longer. The length of the run is expressed in syllables, and the number of syllables is divided by the number of runs
Phonation time ratio	Percentage of time within the 60-second sample spent speaking, including filled pauses
Number of silent pauses	Number of silent pauses of 100 milliseconds and longer in the 60 second sample
Number of filled pauses	Number of filled pauses in the 60 second sample. Filled pauses were defined narrowly as non-lexical fillers such as *um, uh, er*, and so on. Repetitions, restarts and repairs were not included in this measure
Mean length of pauses	The total length of silent pause time divided by the number of silent pauses of 100 milliseconds or longer in the 60 second sample
Number of prominent syllables per run	The total number of prominent syllables divided by the total number of runs
Proportion of prominent words	The percentage of prominent words (i.e., those containing prominent syllable(s) out of the total number of words
Prominence characteristics	Percentage of tone units out of the total number of tone units containing a final prominence or termination choice
Overall pitch range	Measurement of Fo maxima and minima and range in Hertz for each task
All tone choices	Identification of tone (rising, falling, or level) and termination (high, mid, or low) on tonic syllables
Pitch (non-) prominent syllable	Measurement of the Fo of five prominent and five non-prominent syllables and calculation of the average Fo for each category
Pitch new item	Measurement of Fo of new lexical item

Pitch given item	Measurement of Fo of following instances of lexical item as given. Where possible, five lexical items were used to calculate the average Fo for each category
Number of low terminations	Total number of low terminations followed by high-key resets
Avg. Height of onset pitch	Average pitch of high-key onsets
Avg. Height of termination pitch	Average pitch of low terminations
Avg. Paratone pause length	Average length of pauses at paratone boundaries

Table 2. Descriptions of Sub-measures

Data analysis

Data were analyzed using a general multiple re.g.ression model, a statistical tool that enables the assessment of the conjoint and unique contributions of predictor variables (in this case the 29 suprasegmental measures) on criterion variables (in this case the comprehensibility and proficiency ratings). Because regression models involving 29 predictor variables would prove difficult to interpret, a preliminary statistical procedure was conducted in order to reduce the number of predictor variables. Hierarchical cluster analysis (HCA) was used for this purpose. That is, the suprasegmental measures were first subject to HCA in order to establish which measures clustered together in terms of distribution. This resulted in five clusters of measures and nine non-clustered predictors:

- *Suprasegmental fluency cluster*: Number of prominent syllables per run, mean length of runs, phonation time ratio, articulation rate, syllables per second, mid-falling tone choices
- *Unit completeness cluster*: proportion of prominent words, average paratone pause length, mean length of silent pauses
- *Boundary marking cluster*: Number of silent pauses, low termination tones
- *Pitch height cluster*: average height of (non) prominent syllables, pitch on new and given lexical items, overall pitch range, paratone onsets & terminations
- *'Um' cluster*: low-level tone choice and mean length of filled pauses
- *Nine non-clustered features*: number of filled pauses, prominence characteristics, high-, mid- and low-rising tones, high- and mid-level tones, high- and low-falling tones

Results

Our final analysis shows that 52% of the variance in the oral proficiency ratings given by the NS listeners could be accounted for by the combined suprasegmental measures listed above.[1] In other words, half of the assessment of a given NS listener as to a given non-native speaker's proficiency can be attributed to the learner's suprasegmental structure. For the comprehensibility ratings, 50% of the variance was accounted for showing that again, half of the judgment made by a given NS listener as to the comprehensibility of a given nonnative speaker was attributable to their performance on these prosodic measures. In addition, we were able to identify the relative importance of each of the 14 predictors (i.e., the five clusters of measures and the nine non-clustered measures.) We found that for both the oral proficiency rating and the comprehensibility rating, the most significant cluster in the prediction was the suprasegmental fluency cluster with the addition of high-rising and, in the case of proficiency rating, mid-rising tones. These predictor variables revealed positive correlation coefficients, indicating directly proportional relations with the final rating scores. The boundary-marker cluster also showed some positive contribution to both oral proficiency and comprehensibility ratings; however, the 'um' factor (so named because of the low pitched filled pauses) showed little relation to either of the ratings.

Discussion

Perhaps the most important finding of the study is the considerable power of the objectively measured suprasegmental features in accounting for the ratings of oral proficiency and comprehensibility of accented speech. Although this is a remarkable amount of variance to be accounted for by one set of features, we should not be surprised. In spoken discourse, prosodic structure is the frame that other linguistic systems build on. Listeners use prosodic cues to confirm if an item is new or one that they are already aware of, to track important information, and to

[1] See Kang, Rubin & Pickering (2010) for tables showing the comprehensive results of the linear regressions

predict when one topic is ending and another is beginning among many other things.

In addition to the role of prosodic features overall, there were several suprasegmental features that were found to be highly positively associated with comprehensibility and proficiency judgments. The first was the suprasegmental fluency cluster which comprised all the rate sub-measures (syllables per second, articulation time, phonation time, and mean length of runs), one stress sub-measure (number of prominent syllables per run), and one pitch measure (mid-falling tone choice.) The inclusion of rate measures as important for perceptions of fluent production concurs with previous studies investigating fluency (Riggenbach, 2000). Similarly, characteristics of prominence structure have also been shown to be a reliable predictor of fluency judgments (Kormos & Denes, 2004).

The association of mid-falling tones is not unexpected as, contextually, this indicates the addition of new information to an ongoing discourse context as would be expected in this task. These tones are also the most common choice to appear in native English speaker discourse. Mid- and high-rising tone choices were also prominently associated with the listener judgments. In NS discourse, these tones are used to 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 socio-cultural context. Previous research has also shown that NNSs tend to overuse falling tones and thus mask these relationships between related propositions.

The boundary-marking cluster (i.e, number of silent pauses and low termination choices) exhibited a positive although less robust relation to comprehensibility and proficiency ratings. Again, this result is consistent with earlier findings that NNSs' production of low termination tones at the end of conceptual units facilitates the comprehension of discourse structure by NS listeners (Pirt, 1990). Interestingly, the 'um' factor (comprising low-level tone choices and mean length of filled pauses) showed little relation to either sets of ratings. Relatively little work seems to have been done on this kind of hesitancy phenomena, although Freed (2000) suggests that the variation exhibited by less fluent learners in pause structure can be highly idiosyncractic. Thus, it may be that these hesitation phenomena reflect more on individual speaking style.

Of course, it is also the case that 50% of the variance in ratings was not accounted for by these measures. Some of the remaining variance in the assessment of oral performance will in all likelihood be explained by accented realizations of phonemes. However, systematic individual differences among raters (rater bias) no doubt also contribute some 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.

Implications for teaching

At this point in the development of pronunciation pedagogy, most would agree that teaching suprasegmentals is at least as important as teaching vowel and consonant production in order to increase the comprehensibility of oral production. But what should we focus on? Studies such as this one give us a road map in terms of prioritizing features. It suggests that focusing on general features of temporal production will increase NS perceptions of NNS fluency. Mean length of runs (stretches of speech bounded by pauses of 100 ms or longer) are clearly important conceptual units for the listener, as is overall rate as measured by syllables per second. This suggests that increased *discourse*-level practice as opposed to *sentence* level practice will be beneficial for the learner. Prominence structure also seems to be important, and particularly, the deleterious effect ofan overuse of prominence within tone units. Pickering (1999) has previously described this as a tendency in lower level language learners that may be exacerbated by 'teacher induced error.' With a teaching focus that prioritizes prominence within units,(i.e., the old teaching standby: stress content words but not function words), we may be contributing to learner confusion regarding prominence characteristics in discourse. In fact, most tone units contain only one or two prominent syllables, which highlight important information that listeners use to track meaning. Too many prominent syllables per unit will obfuscate this information cue.

On the other hand, the analysis suggests that while some tone choices have significant value, this does not apply to the tone choice system as a whole. Thus, it may not be productive to spend a great deal of time practicing the nuances of the tone and key system (as rendered in Brazil's discourse intonation model), but rather to concentrate on overall patterns of tonal change. The importance of differentiating between mid-falling and mid-rising tones has been noted in previous research as critical for both information structure and relationship-building between interlocutors, and our analysis further supports a pedagogical focus in this area.

Finally, we believe that we have demonstrated the high utility of computer-assisted analysis to diagnose and also to teach prosodic structure for those language learners who find themselves in high stakes environments in primarily native speaker contexts.

Okim Kang is Assistant Professor at Northern Arizona University in the USA. Her research focuses on L2 pronunciation, language attitudes, and oral proficiency assessment. She is the winner of the Christopher Brumfit PhD Thesis 2009 award and has received various research grants.

Lucy Pickering is Associate Professor at Texas A&M-Commerce in the USA. Her work explores the pedagogical applications of speech analysis to the pedagogy of English as a second language and the ways in which learners develop competence in relation to prosody.

Email: esllup@langate.gsu.edu

Note: This article is based on Kang, Rubin & Pickering (2010).

References

Anderson-Hsieh, J. & Venkatagiri, H. (1994). Syllable duration and pausing in the speech of Chinese ESL speakers. *TESOL Quarterly*, 28, 807–12.

Bent, T. & Bradlow, A. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, 114,1600–10.

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

Derwing, T. & Munro, M. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 12, 303–13.

Derwing, T. & Munro, M. (2001). What speaking rate do non-native listeners prefer? *Applied Linguistics*, 22,324–37.

Field, J. (2003). The fuzzy notion of 'intelligibility': A headache for pronunciation teachers and oral testers. *IATEFL SIGs Newsletter in memory of Gillian Porter Iardousse*, 34–38.

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

Freed, B. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In *Perspectives on Fluency* , H. Riggenbach (Ed.), 243–66.

Kang, O. & Rubin, D. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, 28,441–56.

Kang, O., Rubin, D. & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal* , 94/4, 554–66.

Kormos, J. & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32,145–64.

Lindemann, S. (2003). Koreans, Chinese, or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics*, 7, 348–64.

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

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly*, 35, 233–55.

Pirt, G. (1990). Discourse intonation problems for non-native speakers. In *Papers in Discourse Intonation,* M. Hewings (Ed.), 145–56. Birmingham University, English Language Research Centre.

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

Rubin, D. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, 33, 511–31.

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