Using lexical stress, speech rate, rhythm, and pauses to characterize and normalize second language speech intelligibility

Abstract

While a range of measures based on speech production, language, and perception are possible for the predication and estimation of speech intelligibility (Kang et al., 2020), what constitutes second language (L2) intelligibility remains under-defined. Prosodic and temporal features (i.e., stress, speech rate, rhythm, and pause placement) have been shown to impact listener perception (Kang et al., 2020), but their relationship with highly intelligible speech is yet unclear. This study aimed to characterize the L2 speech intelligibility. Acoustic analyses, including PRAAT and Python scripts, were conducted on 405 speech samples (30 seconds each) from 102 L2 English speakers with a wide variety of backgrounds, proficiency levels, and intelligibility levels. The results indicate that highly intelligible speakers of English employ between 2-4 syllables per second and that higher or lower speeds are less intelligible. Silent pauses between 0.3 and 0.8 seconds were associated with the highest levels of intelligibility. Rhythm, as measured by Δ syllable length of all content syllables, was marginally associated with intelligibility. Finally, lexical stress accuracy did not interfere substantially with intelligibility until less than 70% of the polysyllabic words were incorrect. These findings inform the fields of first and second language research as well as language education and pathology.

1. INTRODUCTION

With an increase of second language (L2) English speakers engaging in research and teaching at North American higher education institutions, a more concrete understanding of the role of speech features is necessary to understand what may hinder communication. There are three prominent speech constructs in the line of speech perception research, including accentedness (i.e., how a given utterance is different from an L1 variety), comprehensibility (i.e., how easy is a speaker to understand), and intelligibility (i.e., how much does a listener actually understand a given utterance) (see Munro & Derwing, 1995).

While a range of measures based on speech production, language, and perception are possible for the predication and estimation of these speech constructs (Kang et al., 2020), what constitutes them remains under-defined. Prosodic and temporal features (i.e., speech rate, stress, rhythm, and pause placement) have been shown to impact listener perception to a large extent (Kang, 2010; Kang et al., 2020), but their relationships are oftentimes nonlinear. The present study seeks to adopt a linguistics perspective in exploring how different suprasegmental features (i.e., lexical stress, speech rate, rhythm, and pauses) can be used to characterize the three speech constructs.

A. SPEECH RATE

A growing number of studies have investigated the relationship between speech rate and speech perception. Some scholars have found that accentedness was negatively correlated with speech rate, meaning that listeners tended to find faster speech production less accented (Trofimovich & Baker, 2006). Others, however, found a curvilinear relationship between speech rate and perception of accentedness, suggesting that listeners found speech production more accented if it was either too fast or too slow (Munro & Derwing, 2001). Similarly, the relationship between speech rate and comprehensibility also seemed to be curvilinear, where low comprehensibility was observed in overly fast and overly slow speech (Derwing & Munro, 1997; Kang, 2010; Munro & Derwing, 2001). However, very few studies have investigated the relationship between speech rate and intelligibility. Because intelligibility, accentedness, and comprehensibility are related but conceptually different constructs (Derwing & Munro, 1997), future research is needed to bridge this gap in knowledge.

B. SILENT PAUSES

Pauses, including silent pauses (i.e., pauses that are not filled with any linguistic elements) and filled pauses (i.e., pauses that are filled with linguistic elements such as ‘uh’), are natural phenomena in both L1 and L2 speech (Kang et al., 2010). However, studies have found that L2 speakers (especially with lower proficiency) tended to pause inappropriately (i.e., within the boundaries of a thought group, a meaningful unit), longer, and more in quantity (Kormos & Dénes, 2004; Iwashita et al., 2008). Most studies explored the relationship between pause features and accentedness and comprehensibility, but not intelligibility. Trofimovich and Baker (2006) found that pause duration was the most prominent contributor to listener perception of accentedness. Kang (2010) found that the number and duration of pauses could influence listener perception of accentedness and comprehensibility. Additionally, Kang et al. (2010) found that more pauses could actually benefit comprehensibility. Taken together, the relationship between the number or duration of pauses and listener perception could be nonlinear, which warrants future research.

C. RHYTHM

In English, rhythm is about how we use a combination of stressed and unstressed syllables in sentences (Ghanem et al., in press). Sentences generally have strong beats (i.e., stressed syllables) that are longer, louder, and higher in pitch and weak beats (i.e., unstressed syllables),

shorter, less loud, and lower in pitch. There are many ways to maintain rhythm in spoken English, including the use of vowel reduction (i.e., de-emphasize grammatical words like pronouns; Isaacs & Trofimovich, 2012) and linking and connected speech features (Brown & Crowther, 2022). Studies have found that L2 rhythm could substantially influence one's comprehensibility (Isaacs & Trofimovich, 2012); moreover, factors affecting comprehensibility seemed to depend on the comprehensibility level of the speakers (Isaacs & Trofimovich, 2012). Chen and Zechner (2011) examined a number of speech rhythm measures, finding that variance amongst various lengths of vocalic, consonantal, and syllabic deviations had a weak but noticeable correlation with perceived proficiency. A systematic investigation of the relationship between rhythm and intelligibility which considers the intelligibility level of the speaker would illuminate the relationship between the two.

D. LEXICAL STRESS

Multisyllabic words in English follow a stress pattern where (usually) one syllable receives prominence, meaning that it sounds longer, higher, and longer in pitch (Ghanem et al., in press). Many L2 English speakers do not sufficiently differentiate syllable duration between stressed and unstressed syllables, thereby creating an unexpected speech pattern (Setter, 2006). This, coupled with the lack of or misplaced prominence on syllables, can lead to decreased speech comprehensibility and intelligibility (Field, 2005; Hahn, 2004; Zielinski, 2008). The current knowledge would benefit from whether lexical stress influences speech perception in nonlinear fashion, as hypothesized with other linguistic features.

E. THE PRESENT STUDY

This study investigates the relationship of speech perception and prosodic features of L2 speech in that it seeks to provide a) a systematic investigation of factors affecting accentedness, comprehensibility, and intelligibility within a single dataset and b) consideration of the potential nonlinear relationship between linguistic features and speech perception. To bridge these gaps in knowledge, this study aimed to characterize accentedness, comprehensibility, and intelligibility through the lens of suprasegmental features with the following research question:

1. To what extent do suprasegmental features (speech rate, pauses, rhythm, lexical stress) relate to L2 speech perception (intelligibility, accentedness, and comprehensibility)?

2. METHODOLOGY

A. PARTICIPANTS AND RECORDINGS

The present study included a relatively diverse sample of 108 L2 speakers, including 36 (33.33%) undergraduate students who were taking English language courses, 43 (39.81%) graduate students who were undergoing pre-service and in-service training to teach L2 English, and 29 (26.85%) university staff and faculty members who were considered highly intelligible in a separate study. The participants were based in the U.S. and also China who were attending US courses at the time of the study. In this sample, participants spoke diverse L1s, including Mandarin (50%), Spanish (12.96%), Farsi (9.26%), Russian (6.48%), Arabic (5.56%), and others (15.74%).

B. RECORDINGS

A total of 405 naturalistic audio files of spontaneous speech were elicited from the 108 L2 speakers. The elicitation tasks included in-class and at-home assignments, student presentations in class, microteachings, and speaking exams. The recordings are between two to 15 minutes long (M = 6.2 minutes).

C. RATERS

The rater participants in the study included 15 linguistic experts, including 11 L1 English speakers (73.33%) and four highly proficient L2 English speakers (26.67%). All had a postgraduate degree in the field of applied linguistics (or relevant field) and had received training in phonetics and phonology at graduate levels. Prior to the rating, all attended a two-hour training and norming session.

D. ANALYSIS

Data analyses in the present study included human perceptual judgement and acoustic analysis. The 15 trained raters provided perceptual judgement of the recordings in terms of accentedness, comprehensibility, and intelligibility. Accentedness was measured on a 9-point numerical scale (1 = *extremely accented*, 9 = *not accented at all*, normed to 100), as was comprehensibility (1 = *not comprehensible*, 9 = *completely comprehensible*, normed to 100), following the tradition of L2 speech perception studies (Derwing & Munro, 1997). Interrater reliability for accentedness (*ICC*[3, k] = .70) and comprehensibility (*ICC*[3, k] = .79) ratings were reasonably satisfactory. Intelligibility was measured with three tasks: a) numerical scalar rating, b) sentence transcription, and c) phrase transcription. Listeners provided their numerical scalar rating of intelligibility on one scale (How much was understood: 1 – 100%), with satisfactory interrater reliability, *ICC* (3, k) = .80. For sentence transcription, listeners were asked to listen to a sentence once only (5-17 words) and transcribe what they had heard. They were also asked to transcribe different sets of three phrases (3-6 words each), separated by a short silence and a beep sound. The two transcription tasks were automatically graded by the fuzzy string match score compared to a golden transcript (Bosker, 2021). The final intelligibility score used in the analysis was computed by averaging the intelligibility score from the three different measures detailed above.

In addition to human perceptual judgement, acoustic analysis was performed from an instrumental perspective. The analysis in the present study featured suprasegmental features, specifically, speech rate, pauses, rhythm, and lexical stress. Two measures of speech rate were employed in this study: a) articulation rate and b) syllables per second. For articulation rate, we first used a Praat script to automatically mark boundaries of a 30-second window of each file (De Jong & Wempe, 2009). Then, a phonetician within the research group reviewed the boundary and make corrections (if any). Calculation of articulation rate was done following the formula (1 / mean syllable length). Syllables per second was estimated with python-syllables (Day, 2021), and was calculated following the formula (total number of syllables / recording duration).

Silent pauses are defined as pauses between silences of 0.30 seconds or above consulting de Jong and Bosker's (2013) method. Duration of silent pauses were automatically extracted from text grids for analysis. Rhythm in this study was operationalized as the standard deviation of 30 randomly selected syllables from Praat analysis. Last, lexical stress accuracy was quantified by having four trained lexical stress coders of polysyllabic words manually review polysyllabic words and mark them as correct or incorrect. Incorrectly stressed words included those with no obvious stress or all stressed syllables, as well as those with left shift and right shift.

3. RESULTS

E. SPEECH RATE

Loess regression curves were used to analyze data with linguistic variables as the predictors and the three speech constructs as the outcome variables. Results suggested a) that lower articulation rate minimally impacts intelligibility, b) that the optimal articulation rate for intelligibility would be from around 3.5 to 4.25, and c) a high articulation rate could inhibit intelligibility. Similarly, comprehensibility was impacted by lower articulation rate and more impacted by higher articulation rate. In comparison, articulation rate had very minimal impact on accentedness.

A similar trend was observed regarding the relationship between syllables per second and accentedness, comprehensibility, and intelligibility. Lower syllables per second marginally influenced intelligibility, whereas higher syllables per second impeded intelligibility. The optimal average syllables per second for intelligibility was from about 2.75 to 4.00. Comprehensibility, on the other hand, was impeded more significantly by low syllables per second than by high syllables per second. Syllables per second had minimal effect on accentedness. Figure 1 provides more information about the findings outlined above.

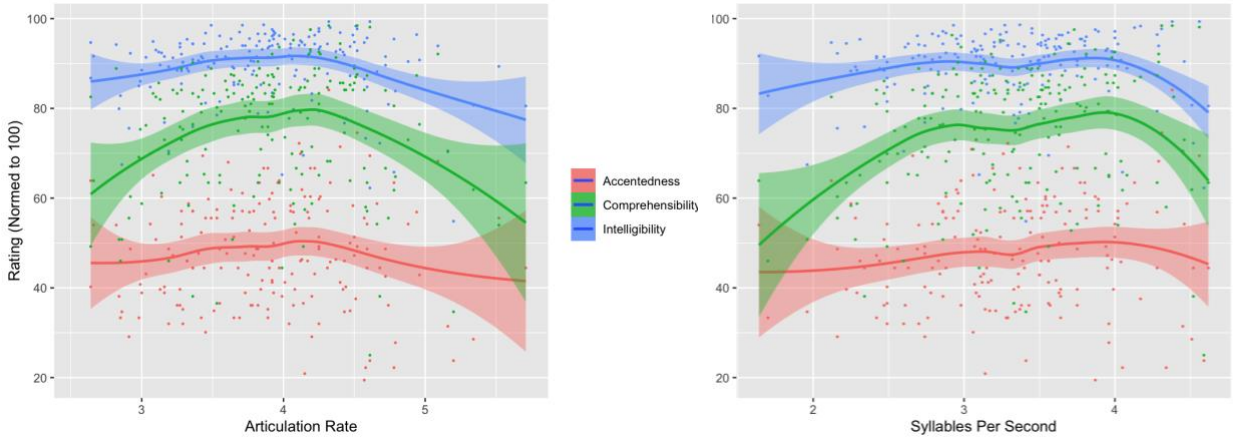

Figure 1. Articulation Rate and Speech Rate.

F. PAUSES

In terms of pauses above 0.30 second, results suggested that the shorter the pauses, the more intelligible the speech were perceived. Specifically, the optimal mean length of pauses for intelligibility should be ideally under 0.75 second. Longer pauses marginally impacted intelligibility. For comprehensibility, shorter pause durations were associated with more comprehensibility. Moreover, comprehensibly showed an observable decrease when the mean pause duration was 1.0 second or longer. No relationship was observed between mean pause length and accentedness, although when the mean pause duration was too long (over 1.50 seconds), the speech was perceived to be more accented. Figure 2 provides more information about the findings outlined above.

G. RHYTHM

Overall, results suggested that rhythm (as measured by the standard deviation of syllable lengths) was marginally related to accentedness, comprehensibility, and intelligibility. For intelligibility, the optimal standard deviation for rhythm is around 1.25; intelligibility slightly decreased when

this value was too high or too low. On the other hand, rhythm did not seem to influence comprehensibility, although when the standard deviation of syllable length was too high (after 1.5), comprehensibility showed a continued downward trend. Last, when the standard deviation of syllable length was under 1.0, speakers were perceived to be more accented. When the value was beyond that, rhythm minimally influence accentedness. Figure 2 provides more information about the findings outlined above.

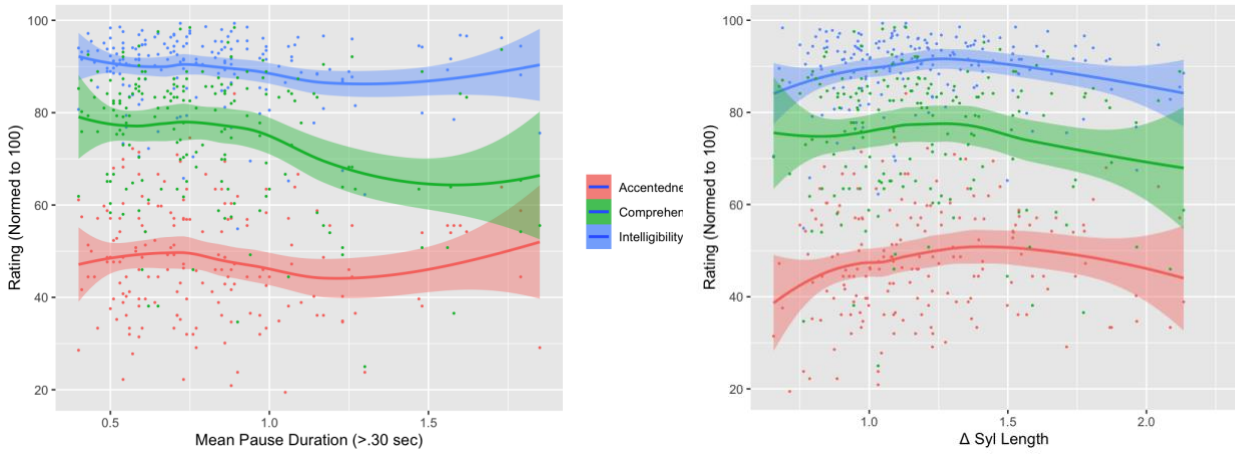

Figure 2. Pauses and Rhythm.

H. LEXICAL STRESS

Lexical stress accuracy differentially influenced one's accentedness, comprehensibility, and intelligibility. Overall, the higher the accuracy, the less accented, more comprehensible, and more intelligible the speakers were perceived. For intelligibility, lexical stress accuracy did not seem to influence it when it was above 70% correct. However, intelligibility was impeded when the accuracy drops below 70%. Second, comprehensibility is relatively stable and positively correlated with lexical stress accuracy. Last, lexical stress accuracy did not seem to be correlated with accentedness when the value was below 90%. However, high lexical stress accuracy (over 90%) was associated with lower degrees of accentedness. Figure 3 provides more information about the findings outlined above.

Figure 3. Lexical Stress Accuracy.

4. DISCUSSION

Overall, different suprasegmental features (i.e., speech rate, silent pauses, rhythm, and lexical stress) were found to be differentially associated with accentedness, comprehensibility, and intelligibility in the present study, oftentimes in a nonlinear fashion. This finding provided empirical support of the argument where accentedness, comprehensibility, and intelligibility are related, but partially independent constructs (Derwing & Munro, 1997; Trofimovich & Isaacs, 2012).

Lower speech rate was found to impact intelligibility minimally. This is not surprising because intelligibility is partially operationalized as sentence / phrase transcription, which required word recognition, and slower speech may not disadvantage intelligibility. On the other hand, higher speech rate was found to inhibit intelligibility. It is possible that higher speech rate was associated with more inaccurate utterances (Brumfit, 1984), which created more processing burdens for the listeners bottom-up processing (Field, 2019). Overall, 2.75–4.0 syllables per second was the ideal speech rate for intelligibility, and the relationship between speech rate and intelligibility was non-linear (Munro & Derwing, 2001). Similarly, the relationship between comprehensibility and speech rate was also in a reverse-U-shaped fashion, where too fast or too slow of a speech would impede comprehensibility (Derwing & Munro, 1997; Kang, 2010). Last, accentedness appeared not to be sensitive to speech rates, although additional data with extremely fast or slow naturalistic speech would provide further evidence for this relationship.

Mean silent pause duration did not seem to affect intelligibility substantially, although shorter pause duration of less than 0.75 second was associated with marginally higher intelligibility. This could be due to the operationalization of intelligibility using transcription tasks, where pauses could give listeners a break before the next words. On the other hand, long pause lengths of over 1.0 second led to significant decrease in comprehensibility. Accentedness did not seem to be related to silent pauses. Taken together, measurement approaches (perceptual scales vs. transcription) tended to reveal differences in word recognition and impressions of difficulty in understanding speakers. It is important to note that although pause length has the potential to contribute to more positive or negative perception, pause location is also important in influencing speech perception (Kahng, 2017). Future research could consider both pause length and location and their relationships with global perception of L2 speech.

Rhythm, as measured by the standard deviation of syllable length of all content syllables, indicates only minor relationships with global perceptions of accentedness, comprehensibility, and intelligibility. Future research could adopt different ways of operationalizing rhythm such as Raw Pairwise Variability Index or VarcoV (White & Mattys, 2007) for a more finely grained perspective into the relationship between rhythm and speech perception.

Lexical stress accuracy did not interfere substantially with intelligibility until less than 70% of the polysyllabic words were incorrect. Moreover, comprehensibility is positively correlated with lexical stress accuracy. Overall, this means that listeners are highly sensitive to lexical stress errors for highly intelligible speakers, which confirms criticality of lexical stress accuracy for intelligibility (Field, 2005; Kang et al., 2020). On the other hand, accentedness did not seem to be related to lexical stress accuracy substantially when the accuracy was below 90%. When the accuracy was above 90%, a clear positive relationship between accentedness and lexical stress accuracy was observed. Taken together, listeners might have adopted a deficit-oriented perspective

in evaluating accentedness, modeled upon nativelike performance, meaning that they only assigned higher ratings when the lexical accuracy was extremely high.

The threshold of intelligibility has been defined as "the lowest requirement for efficiently conveying a message from a native listener's standpoint" (Gimson, 1980). However, to date, very few empirical studies have sought to define the threshold (Kang et al., 2020). The present study contributes to this line of argument and provided empirical data to define this threshold from the perspective of four suprasegmental features. Specifically, ideal speech rates for indelibility were 3.5 – 4.25 (articulation rate) or 2.75 – 4.00 (syllables per second). Moreover, mean silent pause durations of less than 0.75 second, optimal standard deviation of syllable length of about 1.25, and a lexical stress accuracy of above 70% were associated with high intelligibility. These findings provided empirically informed evidence for goal-setting for L2 English learners regarding how to achieve intelligible speech. Underlying these finding was the rationale where intelligible speech, not necessarily nativelike performance, should be prioritized in L2 language learning (Levis, 2018).

ACKNOWLEDGMENTS

This study was funded by EAGER Funding from National Science Foundation (Award Number: 2140414).

REFERENCES

Bosker, H. R., 2021. "Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies". Behavior Research Methods, 53, 1945–1953.

Brown, J. D., Crowther, D., 2022. *Shaping learners' pronunciation: Teaching English connected speech to second language learners*. Routledge.

Brumfit, C., 1984. *Communicative methodology in language teaching: the roles of fluency and accuracy*. Cambridge ; New York, Cambridge University Press.

Chen, L., & Zechner, K. (2011). Applying rhythm features to automatically assess non-native speech. *Proc. Interspeech 2011*, 1861–1864. https://doi.org/10.21437/Interspeech.2011-506

Day, D., 2022. *Syllables: A fast syllable estimator for Python (Version 1.0.6) [Python Package]*. https://pypi.org/project/syllables/

De Jong, N. H., Bosker, H. R., 2013. Choosing a threshold for silent pauses to measure second language fluency. Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech, Stockholm.

de Jong, N.H., Wempe, T., 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*, 385–390. https://doi.org/10.3758/BRM.41.2.385

Derwing, T. M., Munro, M. J., 1997. Accent, intelligibility, and comprehensibility: Evidence from four L1s. Studies in Second Language Acquisition, *19*(1), 1–16. https://doi.org/10.1017/S0272263197001010

Eberhard, D. M., Simons, G. F., Fennig, C. D., 2022. *Ethnologue: Languages of the World*. Dallas, TX: SIL International.

Field, J., 2005. Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*(3), 399–423. https://doi.org/10.2307/3588487

Field, J., 2019. Second language listening: Current ideas, current issues. In J. W. Schwieter & A. Benati (Eds.), *The Cambridge handbook of language learning* (pp. 283–319). Cambridge: Cambridge University Press.

Hahn, L. D., 2004. Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*(2), 201–223. https://doi.org/10.2307/3588378

Isaacs, T., Trofimovich, P., 2012. Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*(3), 475–505. https://doi.org/10.1017/S0272263112000150

Iwashita, Brown, A., McNamara, T., O'Hagan, S., 2008. Assessed Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics*, *29*(1), 24–49. https://doi.org/10.1093/applin/amm017

Kang, O., 2010. Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*(2), 301–315. https://doi.org/10.1016/j.system.2010.01.005

Kang, O., Rubin, D. L., & Pickering, L., 2010. Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*(4), 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Kang, O., Thomson, R., & Moran, M., 2020. Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics*, *41*(4), 453–480. https://doi.org/10.1093/applin/amy053

Kahng, J., 2018. The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*(3), 569–591. https://doi.org/10.1017/S0142716417000534

Kormos, J., Dénes, M., 2004. Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001

Levis, J., 2018. *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press.

Munro, M. J., Derwing, T. M., 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/0023-8333.49.s1.8

Munro, M. J., Derwing, T. M., 2001. Modeling perceptions of the accentedness and comprehensibility of L2 speech: The Role of Speaking Rate. Studies in Second Language Acquisition, 23(4), 451–468. https://doi.org/10.1017/S0272263101004016

Seidlhofer, B., 2011. *Understanding English as a lingua franca*. Oxford University Press.

Setter, J., 2006. Speech rhythm in World Englishes: The case of Hong Kong. *TESOL Quarterly*, *40*(4), 763–782. https://doi.org/10.2307/40264307

Trofimovich, P., Baker, W., 2006. Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 1–30. https://doi.org/10.1017/S0272263106060013

White, L., Mattys, S. L., 2007. Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, *35*(4), 501–522. https://doi.org/10.1016/j.wocn.2007.02.003

Zielinski, B. W., 2008. The listener: No longer the silent partner in reduced intelligibility. *System*, *36*(1), 69–84. https://doi.org/10.1016/j.system.2007.11.004