Linguistic Research 41(1): 27-63
DOI: 10.17250/khisli.41.1.202403.002

Native listeners' perceptual assessments of native and foreign-accented speech and their associations with various speech properties*

Jieun Lee[a] · Dong Jin Kim[b]*** · Hanyong Park[b]**
(University of Kansas[a] · University of Wisconsin-Milwaukee[b])

Lee, Jieun, Dong Jin Kim, and Hanyong Park. 2024. Native listeners' perceptual assessments of native and foreign-accented speech and their associations with various speech properties. *Linguistic Research* 41(1): 27-63. When native listeners assess various aspects of both native and non-native speech, do they rely on similar or different speech properties? We investigated this question by conducting two rating tasks with relatively short, spontaneous utterances produced by ten American-English native speakers (L1 speech stimuli) and 21 Korean native speakers (L2 speech stimuli). Fifty-four native English raters evaluated perceptual aspects such as accentedness, fluency, comprehensibility, and pleasantness of either L1 or L2 speech stimuli. The speech stimuli were coded for the twelve speech properties categorized as speed, lexical richness, rhythm, voice quality, and repair fluency. Analyses included correlation and mixed-effects models that allowed to examine the relationships between rated perceptual dimensions and the relative impact of speech properties on L1 and L2 speech evaluations. The findings indicated more robust and stronger relationships among the perceptual dimensions in the L2 rating task compared to the L1 rating task, suggesting that raters are better able to distinguish multiple perceptual dimensions of more familiar speech (L1) as opposed to less familiar speech (L2). Moreover, mixed-effects model analyses revealed that raters assigned the distinct weights to different linguistic features, albeit with some overlap, depending on the type of speech being evaluated. This empirical evidence underscores the possibility that listeners may assess native and non-native speech utterances in different manners. (**University of Kansas · University of Wisconsin-Milwaukee)**

Keywords perceptual dimensions, rating task, acoustic analysis, L1 speech, L2 speech

* We would like to thank the audiences at *the 179th Meeting of the Acoustical Society of America* (ASA 179) for helpful questions and comments. We also wish to thank two anonymous reviewers for their valuable comments and suggestions. The earlier version of this paper was published in *the Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference* (PSLLT 10).
** First author
*** Corresponding author

1. Introduction

Listeners automatically evaluate various aspects of their interlocutors' speech during conversations. These aspects include the degree of accent, comprehensibility, fluency, intelligibility, and pleasantness, among others. In the realm of second language (L2) speech perception research it is well-established that these perceptual dimensions, although interconnected, are distinct constructs in the minds of native listeners (e.g., Munro and Derwing 1995, 1999; Derwing et al. 2004; Trofimovich and Isaacs 2012; Lee et al. 2019). For example, Derwing and Munro (2009) showed that heavily accented L2 learners can still produce speech that is entirely intelligible to native listeners. Previous studies have further sought to identify which specific speech properties of L2 speech contribute to listeners' judgments of accentedness, fluency, comprehensibility, or intelligibility, aiming to elucidate the relationships between these perceptual dimensions (e.g., Munro and Derwing 1995; Towell et al. 1996; Ortega 1999; Kormos and Dénes 2004; Trofimovich and Baker 2006; Kang et al. 2010; Baker et al. 2011; Robinson 2011; Trofimovich and Isaacs 2012; Bosker et al. 2013; Suzuki and Kormos 2020). While there is a consensus that listeners' assessments are linked to various speech properties, the relative importance of these properties can differ for each perceptual dimension.

Despite advancements in understanding how native listeners perceive L2 speech, there is still relatively limited research investigating how listeners evaluate the same perceptual dimensions in native speech (L1) (e.g., Bosker et al. 2013; Kahng 2018). While research on L2 speech assessments often incorporates L1 speech as a point of comparison to ensure that the listeners evaluate L2 speech adequately, the extent of the correspondence between listeners' evaluations of L1 and L2 speech remains an open question. This study seeks to bridge this gap by investigating how native listeners assess accentedness, comprehensibility, fluency, and pleasantness in two different types of speech: native, L1 speech vs. non-native, L2 accented speech. We further aim to explore how listeners weigh L1 and L2 speech characteristics in their evaluations.

1.1 Listeners' assessments of different types of speech

Previous research has yielded somewhat conflicting findings regarding how listeners with the same L1 background evaluate different types of speech, such as native and non-native accented speech. On one hand, it was suggested that listeners employ similar evaluation

criteria for different types of speech when assessing the same perceptual dimension (e.g., Bosker et al. 2014b; Kahng 2018). For example, Bosker et al. (2014b) manipulated pause and speech rate of L1 speech produced by native Dutch speakers and L2 speech produced by non-native Dutch speakers (5 English and 5 Turkish). Native Dutch speakers' ratings of perceived fluency for both L1 and L2 speech were similarly affected by such pause manipulation; adding silent pauses or lengthening pause duration to L1 and L2 speech stimuli led to lower fluency ratings. Conversely, O'Brien (2014) proposed that listeners use different criteria to evaluate different types of speech by relying on different, various acoustic cues. O'Brien (2014) asked English learners of German (i.e., L2 listeners) to rate the accentedness, comprehensibility, and fluency of speech produced by either fellow language learners of German (i.e., L2 accented speech) or German native speakers (i.e., L1 speech). The study found that different speech characteristics predicted the ratings of these perceptual dimensions for L1 speech compared to L2 speech. Listeners were influenced by factors such as corrections/repetitions and morphological errors when evaluating L1 speech, whereas different factors including phonetic realization, filled pauses, speech rate, stress assignment, corrections/repetitions, and lexical errors predicted ratings of all three dimensions of L2 speech. The author concluded that "the extent to which L2 listeners relied upon various cues in the speech of native and non-native speakers did vary, indicating that they used different criteria in their assessment of native versus non-native speech" (O'Brien 2014: 732).

The divergence in findings prompts us to investigate listeners' evaluations of different types of speech, namely L1 vs. L2 speech, for several reasons. First, there is a notable lack of studies addressing this issue. The studies discussed in the previous paragraphs are among the few available on this topic to the best of our knowledge. Second, these studies have certain limitations in providing a comprehensive understanding of listeners' evaluation processes for L1 and L2 speech. Bosker et al. (2014b) raised questions about whether listeners assess L1 and L2 speech based on similar criteria for perceptual dimensions other than fluency. O'Brien's (2014) study left us wondering if native listeners' evaluations would resemble those of L2 listeners. The present study extends previous research by involving native English listeners as the target listener group and having them evaluate English utterances produced by either native English speakers (L1 speech) or L2 learners of English (L2 speech) in terms of accentedness, fluency, comprehensibility, and pleasantness.

1.2 Perceptual assessment of speech and contributions of speech properties

Numerous studies have delved into how listeners evaluate accentedness, fluency, comprehensibility, and pleasantness in utterances by L2 learners (e.g., Munro and Derwing 1995; Derwing and Munro 1997; Kormos and Dénes 2004; Hayes-Harb et al. 2008; Trofimovich and Isaacs 2012; Zetterholm et al. 2017; Lee et al. 2019). These studies have consistently shown that perceived accentedness, fluency, comprehensibility, and pleasantness are related to each other but vary in the strength of their associations (e.g., Munro and Derwing 1995, 1999; Trofimovich and Isaacs 2012). For instance, Lee et al. (2019) introduced pleasantness as a perceptual dimension alongside accentedness, comprehensibility, and fluency and found that listeners' appraisals of L2 pleasantness were best predicted by their judgments on fluency, followed by comprehensibility and accentedness.

Listeners naturally evaluate various aspects of their interlocutor's speech, yet the assessment of 'pleasantness', which reflects listeners' holistic subjective experience, has not received much attention in comparison to other perceptual dimensions such as accentedness, fluency, and comprehensibility, especially in L2 speech research. Nevertheless, a handful of prior studies on how listeners perceive the pleasantness of L2 speech suggest that it could be a significant factor worth exploring in L2 speech and learning research (e.g., Bouchard Ryan et al. 1977; Derwing and Munro 2009; Lee et al. 2019). For instance, Lee et al. (2019) emphasized the importance of improving the pleasantness of L2 speech, pointing out that pleasant L2 speech might elicit more positive feedback from native listeners, subsequently fostering a desire for increased interactions between native and nonnative speakers. Concerning the relationship between pleasantness and other perceptual dimensions, both Bouchard et al. (1977) and Lee et al. (2019) indicated a possible link between L2 pleasantness and L2 fluency. Specifically, Lee et al. (2019) suggested the potential for enhancing L2 pleasantness alongside L2 fluency, as their findings demonstrated that native listeners ratings of L2 pleasantness were best predicted by ratings of L2 fluency. While Derwing and Munro (2009) did not explicitly examine the pleasantness dimension, their research on how L2 speech comprehensibility influences native listeners' preference for interacting with L2 speakers showed listeners' overall preference for more comprehensible L2 speech. The present study aims to expand upon the limited existing literature on pleasantness in speech judgements by investigating

its relationships with other perceptual dimensions and identifying the acoustic features contributing to greater pleasantness in both L1 and L2 speech.

Extensive research on L2 speech perception has investigated how different speech characteristics contribute to listeners' evaluations of speech dimensions, examining whether perceptual dimensions like accentedness, fluency, and comprehensibility, are indeed distinct constructs (e.g., Towell et al. 1996; Ortega 1999; Kormos and Dénes 2004; Trofimovich and Baker 2006; Kang 2010; Baker et al. 2011; Robinson 2011; Bosker et al. 2013; Suzuki and Kormos 2020). These studies have employed various acoustic measurements to capture speech characteristics, such as speed, stress, rhythm, voice, and intonation (e.g., Munro and Derwing 2001; Kang et al. 2010; Trofimovich and Isaacs 2012). The general finding of previous studies is that the relative contributions of speech properties differ when evaluating L2 accentedness, fluency, and comprehensibility. This suggests that listeners are tuned into different speech characteristics depending on the dimension they are assessing (Trofimovich and Isaacs 2012; O'Brien 2014). For instance, Kang (2010) demonstrated that native listeners' judgments of accentedness in L2 English speech were best predicted by pitch range and word stress measures, while speaking rates mainly predicted comprehensibility judgments.

A few studies have suggested that various speech properties in both L1 and L2 speech have differential effects on listeners. For example, Bosker et al. (2014a) found that when listeners were presented disfluent L1 speech with filler um while viewing pictures of high-frequency objects (e.g., a hand) and low-frequency object (e.g., a sewing machine), they anticipated references to low-frequency objects. However, when they listened to disfluent L2 speech with um, they did not exhibit the same anticipation of low-frequency referents. This suggests that native and non-native disfluencies can have different effects on listeners, and listeners adjust their predictive strategies based on speaker identity. Bosker and Reinisch (2015) investigated native listeners' perceptions of native and non-native speech rates. Their findings indicated that non-native speech is implicitly perceived as faster than temporally-matched native speech, indicating that the additional cognitive load of listening to an accent (i.e., non-native speech) accelerates rate perception.

2. The current study

Our review of the literature reveals a wealth of studies focused on native listeners' evaluations of various speech dimensions and the relative contributions of speech properties to those evaluations, particularly within the context of L2 or accented speech. However, there is a notable gap in our understanding of how similarly or dissimilarly native listeners assess the same aspects of speech (e.g., fluency) for the different types of speech (L1 vs. L2). While previous research in L2 speech has identified unique but overlapping speech properties contributing to judgments with varying degrees of associations, it remains largely unexplored whether similar speech properties are employed by native listeners while evaluating L1 speech. Our study aims to address this gap by investigating how native listeners evaluate both L1 and L2 speech and the contributions of various speech preparties to these evaluations.

In the current study, we conducted two independent rating tasks, one targeting L1 speech and the other L2 speech. Native English listeners were tasked with rating accentedness, comprehensibility, fluency, and pleasantness of relatively short, spontaneous utterances in English, spoken by native speakers of American English or native speakers of Korean. We then investigated the relative contributions of 12 acoustic measurements of both L1 and L2 speech utterances on listeners' ratings of four dimensions. This allowed us to explore whether listeners assign different relative importance to acoustic properties depending on the type of speech being evaluated. This study was guided by the following research questions:

RQ 1. Do the relationships between native listeners' evaluations of accentedness, comprehensibility, fluency, and pleasantness differ depending on the type of speech (i.e., L1 vs. L2)?

RQ 2. Are speech properties that contribute to listeners' evaluations of perceptual dimensions different depending on the type of speech?

Regarding the first research question, we hypothesized that all four perceptual dimensions would exhibit varying degrees of relation in both L1 and L2 speech evaluations. While some perceptual dimensions of L1 speech were expected to show stronger correlations with each other in L2 speech evaluations, we refrained from making a definitive prediction regarding the relative strength of relationships between perceptual

dimensions in L1 speech dimensions due to the scarcity of prior research simultaneously examining listeners' assessments of multiple aspects of different speech types.

Concerning the second research question, we hypothesized that listeners would weigh the relative importance of speech properties differently during L1 and L2 speech processing. L1 speech and L2 speech inherently differ in various aspects, such as disfluencies, errors (e.g., repetitions/corrections, stress/morphological/lexical/syntactic errors) (O'Brien 2014), and lexical complexity (Foster and Tavakoli 2009). Given these inherent differences between L1 and L2 speech, we expected that the prominent speech properties contributing to evaluations across all four dimensions would vary between L1 and L2 speech processing.

3. Method

3.1 Stimuli

Twenty-one L1-Korean adult learners of English (fifteen female, six male; mean age of 27.1 years, range of 20-47 years) and ten L1-American English speakers (seven female, three male; mean age of 26.3 years, range of 18-40 years) produced the stimuli as L2[1] and L1 talkers, respectively. The L2 talkers were speakers of the Seoul/Gyeonggi dialect of Korean and had lived in the US (mean length of residence of 2.82 years, range of 0.25-5.66 years) at the time of recording the stimuli. The L1 talkers were born and raised in the Midwest region of the United States. None of the talkers reported hearing or speech problems.

The L1 and L2 talkers were recorded at different locations but under a similar environment. All recordings took place in a sound-attenuated room using a Shure SM-10A microphone with a sampling rate of 44.1 kHz. The talkers initially read the English passage of *The North Wind and the Sun*, and were later asked to retell the story in English in their own words at their normal speech rate, without prior knowledge that they would be retelling it. For the stimuli selection, two tokens per talker were chosen, excluding the first and last sentences of each retelling. All selected tokens were sentence-length to maintain manageable lengths for the rating tasks.

1 L2 talkers were the subset of participants who participated in a series of cognitive, perception, and production tasks (Darcy et al. 2015).

The selected tokens underwent a modification process. First, filled pauses were removed, with only *uhs* and *ums* considered as filled pauses (Lee 2018). Other disfluencies, such as repetition, replacements, reformulations, hesitations, and false starts were retained. Second, any unfilled pauses (i.e., silence) lasting over three seconds were adjusted by shortening the pause to three seconds. For example, if an unfilled pause was 3.04 seconds, the excessive 0.04 seconds were removed. The pause manipulation was implemented to prevent extremely long pauses in sentence-length stimuli from skewing fluency ratings and affecting the relationships between perceptual dimensions with fluency. Such a three-second pause manipulation was processed only for two tokens out of forty-two L2 stimuli (approximately 4.76% of the total stimuli) and none of the L1 stimuli were altered.

In total, we had 20 tokens (2 tokens per each of the 10 L1 talkers) of L1 stimuli with an average length of 8.7 seconds (*SD* = 2.5 sec). For the L2 stimuli, we had 42 tokens (2 tokens per each of the 21 L2 talkers) with an average length of 8.2 seconds (*SD* = 2.0 sec), which were used for the subsequent rating tasks.

3.2 Speech properties

We analyzed both the L1 and L2 speech samples for 12 speech properties, categorized into five distinctive categories that have been previously employed in the literature (e.g., Kormos and Dénes 2004; Kang 2010; Trofimovich and Isaacs 2012; Bosker et al. 2013). These categories include the following: speed, lexical richness (comprising lexical fluency and lexical variation)[2], rhythm, voice quality, and repair fluency.

We incorporated voice quality as a category, as it has received limited attention in assessing the relations between perceived L1 and L2 speech and their acoustic correlates. Furthermore, voice quality would be an important category to assess pleasantness of speech. Following previous literature (e.g., Scherer and Oshinsky 1977; Ilie and Thompson 2006; Kang et al. 2010; Mori et al. 2011), we included f0 mean, f0 range, and Cepstral Peak Prominence (CPP) as voice quality measurements. These measurements have often been used to assess the relationship between perceived overall voice quality

2 The concept of lexical richness entails the use of varied and sophisticated vocabulary (Saito et al. 2017). Previous literature has identified certain linguistic variables associated with lexical richness, including token and type frequency. Token frequency closely corresponds to our measurements for lexical fluency, whereas type frequency mirrors our measurements for lexical variation.

and acoustic-phonetic correlates. The overall tendency of f0 (f0 mean) has been studied as an indicator of speakers' emotional states (e.g., higher f0 levels indicating happiness/joy, confidence, anger, and fear) and listeners' perception of the speaker (e.g., Giles 1970; Scherer and Oshinsky 1977). Additionally, Järvinen et al. (2017) found that experienced vocologists perceived L2 speech, particularly that of less experienced L2 learners, as having poorer overall voice quality and higher pitch compared to their L1 speech. f0 range, a frequently studied voice quality measure, represents the dispersion of f0 by subtracting the lowest from the highest possible f0 (Buder 2000). For example, Niebuhr et al. (2018) analyzed various f0 and voice quality characteristics in entrepreneurial speeches by L2 speakers of English and found a positive correlation between f0 range and listeners' perceived charisma in those L2 speech samples. Cepstral Peak Prominence (CPP) quantifies signal periodicity and harmonic energy and can serve as a measure of overall voice quality (Feng et al. 2021). CPP is widely used to distinguish normal voices from dysphonic voices (e.g., Watts and Awan 2011) and is considered as a robust correlate of perceived breathiness and a precise predictor of breathiness ratings (e.g., Hillenbrand and Houde 1996; Klug et al. 2019). Maryn et al. (2009)'s meta-analysis on the relationship between perceived voice quality and several acoustic-phonetic correlates confirmed the validity of cepstral measures, suggesting that these measures are potentially the most accurate acoustic correlates of overall voice quality.

For the speech property measurements, first, we transcribed all stimuli with the involvement of two of the authors, and an additional individual, not associated with this study, verified the transcriptions to ensure the accurate representation of both native and non-native speakers' intended messages. Next, all words (tier 2 in Figure 1), pauses (tier 2 in Figure 1 represented as |), and segments (consonants and vowels as in tier 3 in Figure 1 with text annotation) of speech stimuli were coded using Praat (Boersma and Weenink 2017). This coding process was initially carried out independently by two trained phoneticians. Any discrepancies that emerged during this initial coding were addressed through consensus. An example of stimulus coding is presented in Figure 1. It should be noted that all 12 speech properties were measured for each speech sample, indicating that each token had its own 12 values for speech measures. The specific measurements are described in Table 1. For measurements under the speed category, higher values indicate faster speech rates and more syllables.

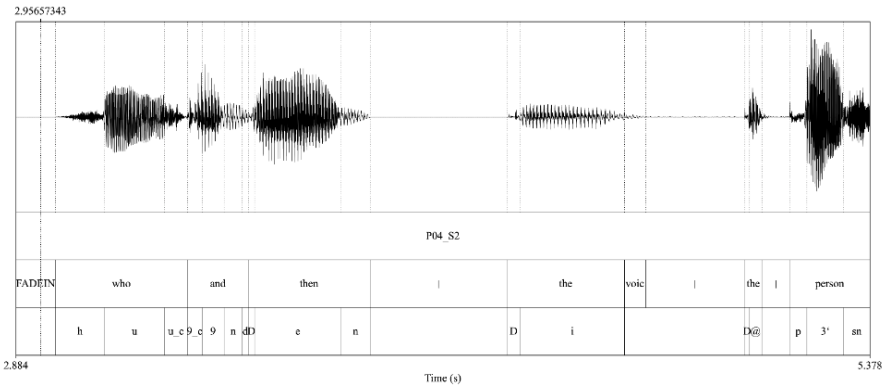

Figure 1. Partial example of stimuli coding. The sound file is from one of L2 talkers. The transcription is the excerpt of the sound file producing that 'who and then the the person a the part who will take off the traveler that guy will be that guy be the the winner'. This sound file includes repair fluency errors such as repetitions (e.g., ***the*** the person) and replacements (e.g., ***a the*** part).

Table 1. Summary of measures for speech properties.

Category		Speech properties	Descriptions
Speed (e.g., Kormos and Dénes 2004; Kang 2010; Trofimovich and Isaacs 2012; Bosker et al. 2013)		Speech rate	Calculated by dividing the total number of syllables by the total length of each speech sample in seconds (including pauses).
		Mean Length of Run (MLR)	Calculated as a mean number of syllables produced between two adjacent pauses of 100 ms or longer.
Lexical Richness (e.g., Iwashita et al. 2008; Bulté and Housen 2012; Saito et al. 2016a, 2016b; Saito et al. 2017; Suzuki and Kormos 2020)	Lexical fluency	Number of words	Calculated by counting the total number of words produced in each speech stimulus.
		Number of syllables	Calculated by counting the total number of syllables produced in each speech stimulus.
	Lexical variation	Word Types	Calculated by counting the total number of morphologically unique word types in each stimulus, using Tool for the Automatic Analysis of Cohesion (Crossley et al. 2016).
		Ratio of content words' syllable numbers to total syllable numbers	Calculated by dividing the number of syllables of content words by the total number of syllables. Higher values indicate more content words in a

	(CT/TN σ #)	speech sample.
Rhythm (e.g., Grabe and Low 2002; Nava and Zubizarreta 2008; Baker et al. 2011; Gut 2012; Zhou and Nagle 2018)	normalized Pairwise Variability Index (nPVI)	$100 \times \left[\sum_{k=1}^{m-1} \left\lvert \frac{d_k - d_{k+1}}{(d_k + d_{k+1}/2)} \right\rvert / (m-1)\right]$ Higher values indicate a more rhythmic speech sample.
	Ratio of function words' syllable duration to content words' syllable duration (FN/CT σ dur)	Calculated by dividing the mean syllable duration of function words by the mean syllable duration of content words. Higher values indicate less durational difference between function words' syllables and content words' syllables in a speech sample.
Voice Quality (e.g., Scherer and Oshinsky 1977; Buder 2000; Ilie and Thompson 2006; Kang et al. 2010; Mori et al. 2011)	Cepstral Peak Prominence (CPP)	Derived via voicesause (Shue et al., 2011) and averaged over each stimulus. Higher CPP values reflect a higher degree of regularity or periodicity and lower degree of breathiness in speech signals, while lower CPP values can be found in less periodic or aperiodic and more breathy signals (Hillenbrand et al. 1994).
	f0 mean	Calculated by averaging the f0 values of the extracted f0 contours. Higher values indicate an overall higher f0 tendency in a speech sample.
	f0 range	Calculate by measuring the difference between the f0 maxima and minima values. Higher values indicate a wider f0 range in a speech sample.
Repair Fluency (e.g., O'Brien 2014; Lee 2018; Saito et al. 2018; Suzuki and Kormos 2020)	Repair fluency	Calculated by counting the number of repairs (repetitions, replacements, reformulations, hesitations, or false starts) appeared in each stimulus. Higher values indicate more instances of repairs in a speech sample. 1. Repetitions: words, phrases, or clauses that are repeated with no modification whatsoever to syntax, morphology, or word order (e.g., s*o one day when a* ***one one*** *one man came north wind blow the wind as hard as possible*) 2. Replacements: lexical items that are immediately substituted for other lexical items (e.g., *but when the north wind*

	blew its wind the person ***the male the guy*** *try to put on his cloak more strongly*) 3. Reformulations: phrases or clauses that are repeated with some modification to syntax, morphology, or word order (e.g., *the the sun sun just* ***maked*** *sun* ***made*** *the traveler warm and he took of* ***her his*** *coat and the sun won the battle*) 4. Hesitations: initial phoneme or syllable(s) uttered one or more times before the complete word is spoken (e.g., *so the north wind wanted to show his* ***pow*** *his* ***power***) 5. False starts: utterances that are abandoned before completion and that may or may not be followed by a reformulation (e.g., ***however the sun try to*** *however the sun just said taking off the man's cloak*)

3.3 Differences between L1 and L2 stimuli

Following the measurement of the 12 speech properties of the stimuli, we conducted a series of independent t-tests to examine differences between L1 and L2 stimuli. As shown in Table 2, L1 and L2 speech stimuli showed their differences in speech rate, mean length of run, number of words, number of syllables, word types, CPP, and repair fluency. These findings suggest that in comparison to L2 talkers, L1 talkers produced speech stimuli characterized by a faster speech rate, greater lexical richness, reduced periodicity (indicating a more breathy voice), and fewer instances of repairs.

Table 2. Summary of independent t-test for speech measures between L1 and L2 stimuli. Means and standard deviations (in parentheses) are provided.

Speech Properties	L1 Stimuli (Native Speakers of English)	L2 Stimuli (Korean Learners of English)
Speech rate	3.0 (0.9)**	2.3 (0.4)
Mean length of run	7.4 (4.3)***	3.0 (1.1)
Number of words	22.0 (9.1)*	16.7 (5.0)
Number of syllables	25.2 (10.9)*	19.4 (5.6)

Word Types	17.1 (6.3)**	12.5 (2.7)
CT/TN σ #	0.5 (0.1)	0.6 (0.1)
nPVI	68.4 (11.7)	65.7 (15.9)
FN/CT σ dur	0.8 (0.2)	0.8 (0.3)
CPP	18.5 (1.2)	19.4 (1.6)
f0 mean	166.0 (41.0)	171.0 (39.3)
f0 range	139.4 (71.7)	119.0 (46.7)
Repair fluency	0.0 (0.1)***	0.1 (0.1)

Note. $*p < .05$. $**p < .01$. $***p < .001$

3.4 Raters

A total of fifty-four monolingual speakers of American English participated as raters (38 female, 16 male; mean age of 22.8 years, range of 18-43 years). All raters completed a language background questionnaire, revealing that none of them identified as bilingual, and none reported any hearing or speech impairments. To account for potential influences on accentedness judgments, especially when evaluating L1 speech stimuli, we included a question asking raters to self-identify their English accent. The questionnaire results showed that 46% of raters self-reported as having no accent, considering themselves to possess a standard American English accent. Meanwhile, 54% of raters self-identified as having a Midwestern accent, specifying their accents as stemming from the Midwest, Wisconsin, Milwaukee, or Illinois (Chicago). Notably, 93% of all raters were born in the Midwestern regions of the United States and all raters were attending college in the Midwest at the time of their participation in the study. These findings indicate that the majority of raters shared a similar native language dialect background.

Twenty-four raters completed the rating task with L1 speech samples, while the remaining 30 raters completed the rating task with L2 speech samples. All raters were college students residing in the Midwestern area of the United States and received extra course credit for their participation.

3.5 Rating tasks

Two separate rating tasks were conducted, one involving L1 speech stimuli and the other L2 speech stimuli. The rating tasks were administered using Praat (Boersma and Weenink 2017) with high-quality headphones at the University of Wisconsin-Milwaukee Phonetics

Laboratory. The raters were randomly assigned to either the L1 or L2 training tasks (i.e., either with L1 or L2 stimuli). One of the authors provided brief and general definitions for each perceptual dimension, along with the rating scales. Raters were allowed to ask questions if they found anything unclear. During this instructional phase, we avoided associating specific speech features (e.g., speech rate) with particular perceptual dimensions (e.g., fluency). This was done to prevent raters from evaluating a speech dimension by solely focusing on one aspect of speech while ignoring others. Instead, our goal was to investigate how various speech features collectively influence raters' judgments.

The rating dimensions were defined as follows based on the definitions used in previous research (e.g., Munro and Derwing 1995; Derwing and Munro 1997; Kormos and Dénes 2004; Trofimovich and Isaacs 2012; Zetterholm et al. 2017): Accentedness—how different the speaker's accent is from standard American English[3]; Fluency—how fluent or disfluent the speaker is; Comprehensibility—how easy or difficult it is to understand the sentence; Pleasantness—how pleasant or unpleasant your experience of listening to the sentence is. The rating scales were 9-point Likert scales, with higher scores indicating better performance across all dimension ratings. The scales were as follows: accentedness (1 = very strong foreign accent, 9 = no foreign accent); comprehensibility (1 = impossible to understand, 9 = very easy to understand); fluency (1 = very disfluent, 9 = very fluent); pleasantness (1 = very unpleasant, 9 = very pleasant).

Prior to the main rating task, a practice session was conducted to familiarize raters with the task procedure and each rating scale. The rating task was divided into two separate sessions with a mid-session break. During each session, raters were instructed to rate two different dimensions using a 9-point Likert scale after listening to each stimulus. For example, if a rater rated pleasantness and accentedness during the first session, they then rated fluency and comprehensibility in the second session for the same stimuli set with different randomization. This design was implemented to prevent raters from experiencing confusion when dealing with the definitions of multiple perceptual dimensions simultaneously. This was important given that the raters in this study were

3 The term "Standard American English" is commonly linked to an accent that is perceived as regionally and socially neutral (e.g., Preston 1996; Bonfiglio 2010). Within the U.S, people treat the Midwest as the locus of "neutral" or "unmarked" speakers (Clopper and Pisoni 2006a, 2006b), which coincides with the language background of most of our raters.

novices without prior experience with speech rating tasks. The entire rating task took approximately thirty minutes to complete.

3.6 Data analysis

To assess whether raters could discriminate L1 and L2 stimuli along each of the four rating dimensions, we ran mixed-effects model analyses using the *lme4* package in *R* (R Core Team 2018). Separate analyses were performed for each rating dimension. In these analyses, the dependent variables were the ratings, the fixed effect variable was the type of stimuli (i.e., L1 vs. L2), and the random effect variable was intercepted for raters.

Regarding the first research question, we conducted Pearson correlation analyses to examine the relationships among perceptual dimensions in L1 and L2 rating tasks. Fisher *r*-to-*z* transformations (Bonferroni adjusted) were additionally conducted to explore potential statistical differences in the strength of correlation coefficients between perceptual dimensions in L1 and L2 speech rating tasks (Yuan et al. 2013; Saito et al. 2016). Specifically, we aimed to determine whether the associations between perceptual dimensions (e.g., fluency and pleasantness) exhibited varying strengths of associations depending on the type of rating task (L1 vs. L2).

To evaluate the contributions of various speech properties in the perception of different perceptual dimensions of speech (Research question 2), we analyzed the rating results using a mixed-effects model analysis. Eight independent analyses (4 perceptual dimensions × 2 types of speech) were conducted. As fixed effects, we entered the ten measures mentioned in section 3.1 into the model. Among the 12 measures, MLR and number of words were not included because both MLR and number of words showed their high correlations with speech rate and number of syllables ($r > .80$). It is important to note that all the measures were transformed into *z*-scores before they were entered in the model (Jassem 1971; Menn and Boyce 1982). Rater intercepts were included as random effects. We obtained *p*-values of ten fixed effects by multiple Likelihood ratio tests with the full model and the model without the effect in question and found the best-fitted models. In conclusion, a total of eight best-fitted models were established (4 perceptual dimensions × 2 types of speech).

4. Results

In the result section, we first present the findings of the rating tasks and explore the relationships between accentedness, comprehensibility, fluency, and pleasantness for both L1 and L2 speech. Subsequently, we examine the relative contributions of ten speech properties to the rating results and present the best-fit models for each perceptual dimension for L1 and L2 speech stimuli.

4.1 Relationships between ratings of perceptual dimensions by type of speech

This section addresses the first research question: "Do the relationships between native listeners' evaluations of accentedness, comprehensibility, fluency, and pleasantness differ depending on the type of speech (i.e., L1 vs. L2)?" Figure 2 illustrates the distribution of ratings for L1 and L2 speech rating tasks. Overall, raters assigned higher scores to L1 stimuli compared to L2 stimuli. The results of mixed-effect model analyses showed significant main effects of speech type (L1 vs. L2), indicating that raters judged L1 stimuli to be significantly less accented (β = 2.57, t = 9.28, p < .001), more comprehensible (β = 2.85, t = 10.04, p < .001), more fluent (β = 2.74, t = 9.07, p < .001), and more pleasant (β = 1.21, t = 3.73, p < .001) than those produced by non-native speakers.

To examine the relationships among perceptual dimensions in both L1 and L2 perception, we conducted Pearson correlation analyses. As shown in Table 3, the analyses revealed significant positive correlations between accentedness, fluency, comprehensibility, and pleasantness ratings of both L1 and L2 stimuli. This indicates that as the ratings of one dimension (e.g., accentedness) increased, the ratings of other dimensions (e.g., comprehensibility, fluency, and pleasantness) also increased. Notably, both L1 and L2 fluency and comprehensibility ratings exhibited the strongest positive correlation (r > .50).

As shown in Table 3, the correlation strengths between perceptual dimensions were stronger in L2 speech ratings than in L1 speech ratings. For example, the correlation coefficients between pleasantness and fluency ratings were 0.35 for L1 speech and 0.56 for L2 speech. We conducted Fisher r-to-z transformations to examine whether perceptual dimensions show different strength of their associations depending on the type of speech being evaluated. The results revealed that some perceptual dimensions showed stronger

correlations in the ratings of L2 speech compared to the ratings of L1 speech. Specifically, the associations between accentedness and comprehensibility (z = 4.75, p < .001), pleasantness and accentedness (z = 3.65, p < .001), pleasantness and fluency (z = 4.37, p < .001), and pleasantness and comprehensibility (z = 5.03, p < .001) were significantly stronger in the L2 rating task compared to the L1 rating task.

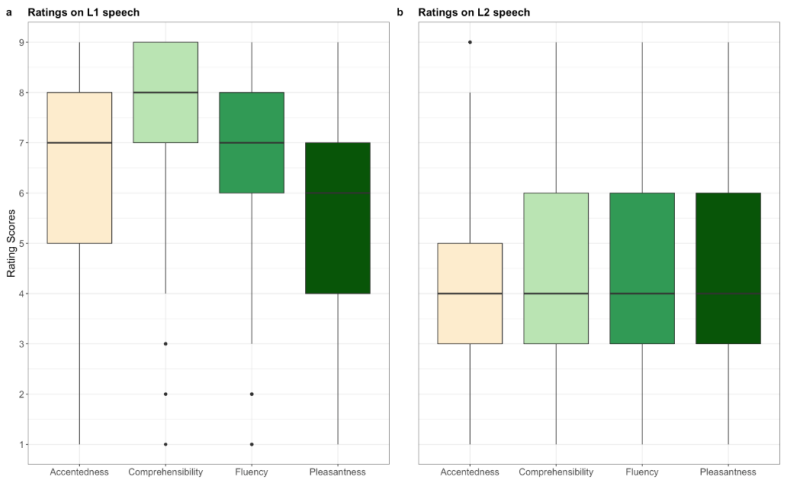

Figure 2. Boxplots of rating task results on (a) L1 and (b) L2 speech stimuli by perceptual dimensions (accentedness, comprehensibility, fluency, and pleasantness). Boxplots: shaded region indicates interquartile range; whiskers extend to extreme values; solid bar indicates median; points indicate outliers.

Table 3. Means and standard deviations for L1 and L2 speech ratings and intercorrelations (Pearson r) between perceptual dimensions; (1) Accentedness, (2) Comprehensibility, (3) Fluency, (4) Pleasantness.

	M	*SD*	1	2	3	4
L1 Speech						
1. Accentedness	6.4	2.2	-			
2. Comprehensibility	7.5	1.6	.26***	-		
3. Fluency	6.7	2.1	.32***	.52***	-	
4. Pleasantness	5.7	21.	.24***	.32***	.35***	-
L2 Speech						
1. Accentedness	3.8	1.9	-			
2. Comprehensibility	4.7	2.1	.48***	-		
3. Fluency	4.1	2.1	.38***	.59***	-	

4. Pleasantness	4.6	2.0	.42***	.51***	.56***	-

Note. *p < .05. **p < .01. ***p < .001

4.2 Speech measures and ratings

Although all four perceptual dimensions showed significant positive correlations to each other in both L1 and L2 speech perception, the ten speech measurements can differ in their contributions to the ratings of L1 and L2 speech stimuli. In this section, we compare L1 and L2 best-fit models for the same perceptual dimension, addressing our second research question: "Are speech properties that contribute to listeners' evaluations of perceptual dimensions different depending on the type of speech?"

4.2.1 Accentedness

Tables 4 and 5 summarize the best-fit models for accentedness ratings of L1 and L2 speech stimuli, respectively. For L1 accentedness ratings, speech properties related to Speed, Rhythm, Lexical Richness, and Voice Quality significantly influenced the ratings. Notably, faster L1 speech was perceived as less accented. Concerning rhythm, stimuli with a wider range of vowel durational variability (nPVI) were rated as less accented, but the syllable duration ratio of function words to content words had a different effect. Stimuli with a higher syllable duration ratio of function words to content words were also perceived as less accented, despite indicating greater duration variations in speech due to shorter durations on function words. The possible reasons for this finding are discussed in the General Discussion section. Additionally, L1 stimuli with more syllables were perceived as less accented, and those with lower f0 mean and wider f0 range received lower accentedness ratings. For L2 accentedness ratings, properties related to Speed, Lexical Richness, Voice Quality, and Repair Fluency were significant. Raters assigned better accentedness ratings (i.e., less foreign accent) to L2 speech stimuli with faster speech rate, a greater variety of word types, lower f0 mean, and/or fewer disfluencies.

Common factors affecting both L1 and L2 speech ratings included faster speech and lower f0 mean, linked to reduced perceived accent. However, some factors, such as rhythm-related speech characteristics (FN/CT σ dur and nPVI) and f0 range, were specific to L1 accentedness ratings, while repair fluency only affected L2 accentedness ratings. Despite lexical richness-related properties influencing both types of speech, they differed

between the models (CT/TN σ # and number of syllables for L1, word types for L2).

Table 4. The best-fit model summary of speech measures of L1 speech stimuli for accentedness ratings.

Category	Measurements	*Estimate*	*SE*	*t*	*p*
	(Intercept)	6.40	0.24	26.67	< .001
Speed	Speech rate	1.01	0.15	6.62	< .001
Rhythm	FN/CT σ dur	0.30	0.10	3.12	.002
	nPVI	0.52	0.12	4.18	< .001
Lexical Richness	CT/TN σ #	0.35	0.12	3.05	.002
	Number of syllables	-1.32	0.19	-7.10	< .001
Voice Quality	f0 mean	-0.59	0.10	-6.24	< .001
	f0 range	0.76	0.11	7.19	< .001

Table 5. The best-fit model summary of speech measures of L2 speech stimuli for accentedness ratings.

Category	Measurements	*Estimate*	*SE*	*t*	*p*
	(Intercept)	3.81	0.15	26.04	< .001
Speed	Speech rate	0.30	0.05	6.42	< .001
Lexical Richness	Word type	0.36	0.05	7.95	< .001
Voice Quality	f0 mean	-0.30	0.05	-6.68	< .001
Repair Fluency	Repair fluency	-0.40	0.05	-8.61	< .001

4.2.2 Comprehensibility

Tables 6 and 7 summarize the best-fit models for comprehensibility ratings of L1 and L2 speech stimuli, respectively. For L1 comprehensibility, speech properties related to Speed, Rhythm, Lexical Richness, and Voice Quality played an important role. Listeners found L1 stimuli more comprehensible when they were spoken at a faster rate, exhibited greater duration variability (higher nPVI), had a richer lexicon (more unique word types), and/or contained fewer syllables. Additionally, stimuli with lower f0 mean and wider f0 range were linked to higher comprehensibility.

With respect to L2 comprehensibility, six speech measures from the Speed, Lexical Richness, Voice Quality, or Repair Fluency category contributed to higher ratings. Faster speech rate, less duration reduction on function words (higher ratio of function words' syllable duration to content words' syllable duration), fewer number of syllables, and/or greater word type diversity led to better comprehensibility. Lower f0 mean and reduced

repair fluency were also associated with higher comprehensibility.

Across both speech types, speech rate, word types, number of syllables, and f0 mean commonly influenced comprehensibility ratings, with faster speech, richer vocabulary, fewer syllables, and lower f0 promoting higher comprehensibility. Notably, nPVI and f0 range only impacted L1 comprehensibility ratings, while repair fluency exclusively affected L2 comprehensibility ratings.

Table 6. The best-fit model summary of speech measures of L1 speech stimuli for comprehensibility ratings.

Category	Measurements	*Estimate*	*SE*	*t*	*p*
	(Intercept)	7.47	0.20	36.95	< .001
Speed	Speech rate	0.72	0.11	6.90	< .001
Rhythm	nPVI	0.16	0.08	2.16	.031
Lexical Richness	Word types	0.73	0.13	5.57	< .001
	Number of syllables	-1.34	0.17	-7.98	< .001
Voice Quality	f0 mean	-0.18	0.07	-2.75	.006
	f0 range	0.24	0.07	3.40	.001

Table 7. The best-fit model summary of speech measures of L2 speech stimuli for comprehensibility ratings.

Category	Measurements	*Estimate*	*SE*	*t*	*p*
	(Intercept)	4.72	0.18	25.93	< .001
Speed	Speech rate	0.52	0.05	9.82	< .001
Lexical Richness	FN/CT σ dur	0.14	0.05	2.91	.004
	Word types	0.43	0.10	4.52	< .001
	Number of syllables	-0.22	0.11	-2.02	.043
Voice Quality	f0 mean	-0.25	0.05	-4.95	< .001
Repair Fluency	Repair fluency	-0.61	0.07	-8.67	< .001

4.2.3 Fluency

Tables 8 and 9 summarize the best-fit models for fluency ratings of L1 and L2 speech stimuli. For L1 fluency, seven speech measures in the Speed, Rhythm, Lexical Richness, or Voice Quality categories significantly influenced ratings. Faster speech rate, higher nPVI, more content words, more unique word types, fewer syllables, lower f0 mean, and/or wider f0 range were associated with higher L1 fluency ratings.

For L2 fluency, seven speech measures from Speed, Rhythm, Lexical Richness, Voice

Quality, and Repair Fluency categories significantly contributed to fluency ratings. L2 stimuli with faster speech rate, higher ratio of function words' syllable duration to content words' syllable duration, more content words, and/or more unique word types were rated with higher fluency scores. In addition, lower number of syllables, lower f0 mean, and lower repair fluency were linked to higher fluency ratings.

Common factors influencing fluency ratings for both L1 and L2 speech included faster speech rate and measures related to lexical richness (more content words, richer vocabulary, and fewer syllables). However, some differences existed between the models. nPVI and f0 range exclusively contributed to L1 fluency ratings, while the ratio of function words' syllable duration to content words' syllable duration and repair fluency solely contributed to L2 fluency.

Table 8. The best-fit model summary of speech measures of L1 speech stimuli for fluency ratings.

Category	Measurements	*Estimate*	*SE*	*t*	*p*
	(Intercept)	6.70	0.23	28.69	< .001
Speed	Speech rate	1.22	0.14	9.05	< .001
Rhythm	nPVI	0.36	0.10	3.48	.001
Lexical Richness	CT/TN σ #	0.21	0.09	2.20	< .028
	Word types	1.15	0.16	7.12	< .001
	Number of syllables	-2.11	0.23	-9.33	< .001
Voice Quality	f0 mean	-0.30	0.08	-3.65	< .001
	f0 range	0.41	0.09	4.69	< .001

Table 9. The best-fit model summary of speech measures of L2 speech stimuli for fluency ratings.

Category	Measurements	*Estimate*	*SE*	*t*	*p*
	(Intercept)	4.08	0.19	20.97	< .001
Speed	Speech rate	0.63	0.05	11.94	< .001
Rhythm	FN/CT σ dur	0.10	0.05	2.22	.027
Lexical Richness	CT/TN σ #	0.16	0.05	3.35	.001
	Word types	0.51	0.19	5.68	< .001
	Number of syllables	-0.29	0.10	-2.78	.006
Voice Quality	f0 mean	-0.28	0.05	-6.12	< .001
Repair Fluency	Repair fluency	-0.61	0.06	-9.52	< .001

4.2.4 Pleasantness

Tables 10 and 11 summarize the best-fit models for pleasantness ratings of L1 and L2 speech stimuli. For L1 pleasantness, speech measures in the Speed, Rhythm, Voice Quality, and Repair Fluency were included in the model. Lower number of syllables and higher nPVI, indicating greater duration variation in vowel length, were associated with higher pleasantness ratings. Voice quality measures, including lower f0 mean, wider f0 range, and lower CPP, were linked to higher pleasantness ratings. The association of CPP indicates that listeners perceived a less periodic and/or breathier voice as more pleasant L1 speech. The presence of disfluencies in L1 speech had a negative impact on pleasantness ratings, with more repairs resulting in lower ratings.

For L2 pleasantness, speech measures in the Speed, Lexical Richness, Voice Quality, and Repair Fluency were relevant. Faster speech rate and more unique word types were associated with high pleasantness scores. Regarding voice quality related measures, f0 mean and CPP contributed to L2 pleasantness ratings. Lower f0 mean and lower CPP values were linked to higher pleasantness ratings. Lastly, repair fluency was negatively associated with L2 pleasantness ratings, suggesting that raters gave negative pleasantness scores to L2 speech with more repairs.

Both L1 and L2 pleasantness, ratings were influenced by f0 mean, CPP, and repair fluency, with lower f0 mean, CPP, and repair fluency positively associated with pleasantness ratings. The most notable difference of the best-fit models for pleasantness ratings, compared to other dimensions (i.e., accentedness, comprehensibility, and fluency), was the contribution of the CPP measure. CPP was associated with perceived pleasantness for L1 and L2 speech but not with other perceptual dimensions. Considering that higher CPP values are often associated with highly periodic (harmonic) signals and more pressed/normal phonations, while lower CPP values are linked with more breathy types of phonations (Hillenbrand et al. 1994; Shue et al. 2010), the contributions of lower CPP values to higher pleasantness ratings were unexpected. This unexpected finding of lower CPP values being linked to more pleasant speech is discussed in the following section. It is noteworthy that repair fluency had an impact solely on L1 speech ratings and did not influence any other L1 perceptual dimensions.

Table 10. The best-fit model summary of speech measures of L1 speech stimuli for pleasantness ratings.

Category	Measurements	*Estimate*	*SE*	*t*	*p*
	(Intercept)	5.66	0.22	25.56	< .001
Rhythm	nPVI	0.38	0.11	3.51	< .001
Lexical Richness	Number of syllables	-0.44	0.14	-3.16	.002
Voice Quality	f0 mean	-0.21	0.09	-2.33	.02
	f0 range	0.26	0.11	2.45	.02
	CPP	-0.39	0.11	-3.58	< .001
Repair Fluency	Repair fluency	-0.39	0.09	-4.22	< .001

Table 11. The best-fit model summary of speech measures of L2 speech stimuli for pleasantness ratings.

Category	Measurements	*Estimate*	*SE*	*t*	*p*
	(Intercept)	4.57	0.19	24.61	< .001
Speed	Speech rate	0.58	0.05	11.86	< .001
Lexical Richnes	Word typesr	0.17	0.05	3.61	< .001
Voice Quality	f0 mean	-0.18	0.05	-3.87	< .001
	CPP	-0.12	0.05	-2.59	.01
Repair Fluency	Repair fluency	-0.50	0.05	-10.18	< .001

5. Summary

Figures 3 and 4 provide a summary of all best-fit models for L1 and L2 speech ratings, respectively. In our previous Pearson correlation analyses, comprehensibility and fluency ratings consistently exhibited the strongest correlation, regardless of the speech type (L1 or L2). This pattern is also evident in Figures 3 and 4, where many of the significant measures in the comprehensibility models are shared with the fluency models.

In Figures 3 and 4, the speech measurements highlighted in red were included in all four models for accentedness, comprehensibility, fluency, and pleasantness ratings. However, the specific measures that were commonly included differed depending on the type of speech being rated.

For L1 speech ratings (Figure 3), stimuli with higher vowel duration variability (nPVI), fewer syllables, lower f0 mean, and wider f0 range were perceived as more fluent, comprehensible, pleasant, and less accented. For L2 speech ratings (Figure 4), stimuli with faster speech rates, a greater variety of unique word types, lower f0 mean,

and lower repair fluency were considered more fluent, comprehensible, pleasant, and less accented. Two notable differences between commonly contributing measures for L1 and L2 speech ratings stand out. First, in L1 speech ratings, rhythm-related measures like nPVI and f0 range played a significant role, whereas they had no significant impact on any of the L2 perceptual dimensions. Second, word type, indicating lexical richness, and repair fluency had a more pronounced influence on L2 speech ratings compared to L1 speech ratings. These findings suggest that speech properties have differential effects on how raters perceive L1 and L2 speech. Depending on the type of speech, different speech properties become more salient than others and influence overall judgments.

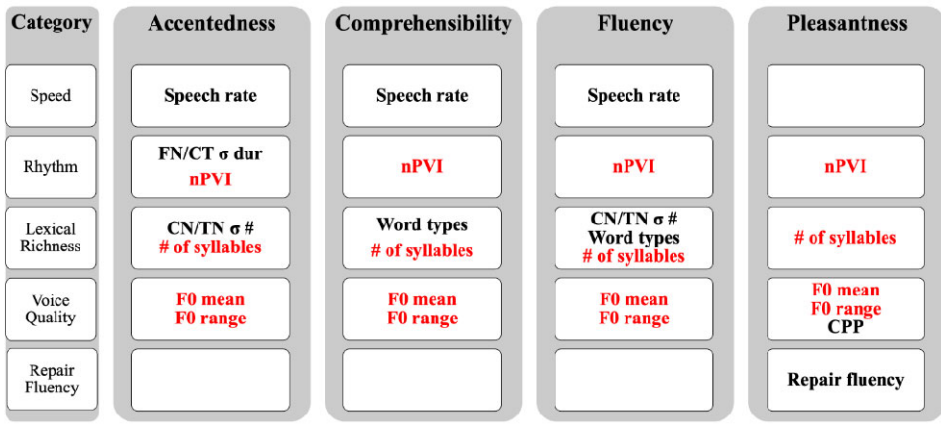

Figure 3. Summary of accentedness, comprehensibility, fluency, and pleasantness best-fit models of speech measures of L1 speech stimuli. Speech measures, which are included in all four models, are in red.

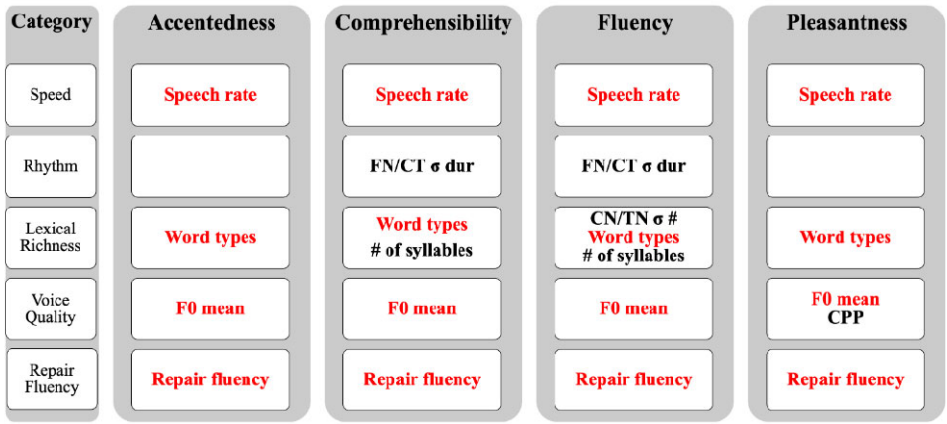

Figure 4. Summary of accentedness, comprehensibility, fluency, and pleasantness best-fit models of speech measures of L2 speech stimuli. Speech measures, which are included in all four models, are in red.

6. General discussion

6.1 Relationships between ratings of perceptual dimensions by the type

This study aimed to compare the results of two rating tasks, which targeted either L1 or L2 speech. It investigated whether the relationships between listeners' evaluations of accentedness, comprehensibility, fluency, and pleasantness differed depending on the type of speech being assessed.

The results in this study reaffirmed the previous findings that L2 perceptual dimensions are related to each other with varying degrees of strength and that some dimensions show closer relationships than other pairs of perceptual dimensions (e.g., Derwing et al. 2004). The ratings of L2 accentedness, comprehensibility, fluency, and pleasantness showed significant but different strength correlations with one another, and L2 comprehensibility and fluency ratings demonstrated the highest correlation strength. Importantly, this study provided empirical evidence supporting the hypothesis that perceptual dimensions of L1 speech also exhibit similar associations to those observed in L2 speech with the strongest relationship between L1 fluency and comprehensibility ratings.

Notably, the associations among perceptual dimensions were stronger in the ratings of L2 perceptual dimensions than in the ratings of L1 perceptual dimensions (Table 3). One interpretation of this result is that native listeners in this study may have a better ability to distinguish between the concepts represented by the four rating dimensions when evaluating the speech of fellow native speakers. This could be due to their higher familiarity with L1 speech compared to L2 speech. It's possible that the raters in this study, who were native English listeners, had limited exposure to foreign-accented speech, which may have influenced their ability to maintain distinct concepts for the four rating scales during the L2 rating task. Instead, when assessing L2 speech samples across four different rating scales, their judgments may have been influenced by an overall impression of global L2 proficiency. This suggests that while raters may possess the ability to apply different criteria to evaluate each rating scale separately, their lack of experience with foreign-accented speech may result in a more holistic impression of perceived L2 proficiency, leading to similar rating scores for all four rating scales. In a similar vein, Huang (2013) argued that raters' ability to separate rating dimensions depends on their familiarity with the target speech. Huang's study demonstrated that raters with high familiarity with foreign-accented English (e.g., ESL teachers) were better able to distinguish between language dimensions during the rating task than raters with less familiarity with foreign accents or without experience teaching English as a second language.

Supporting evidence for raters' varying ability to separate rating dimensions depending on the familiarity of the target speech is also found in the comparison of L1 best-fit models in Figure 3. Raters relied on different speech measures when evaluating each perceptual dimension of L1 speech, with some speech measures uniquely included in specific L1 best-fit models.[4] As for the L2 rating task, we expected to observe a similar tendency if raters were capable of differentiating between rating dimensions during the evaluations. However, raters seemed to associate unique speech properties to a lesser extent in L2 speech evaluations. All four best-fit models in Figure 4 tended to include similar speech measures. Speech rate, word type, f0 mean, and repair fluency

4 It should be noted that the ratio of function words' syllable duration to content words' syllable duration was not considered in a comparison of best-fit models. In spontaneous speech, it is common that some function words are even lengthened instead of being shortened, where speakers slow down their articulation during the function word while searching for the proper content word to follow (O'Shaughnessy 1995). Therefore, the ratio of function words' syllable duration to content words' syllable duration may not accurately represent the duration reduction on function words as intended.

were commonly included speech measures, and only one or two additional/unique speech measures attributed to the ratings of L2 comprehensibility, fluency, or pleasantness. These findings suggest that raters exhibit clearer distinctions between rating dimensions during L1 speech evaluations. While we do not dismiss the raters' ability to distinguish between rating dimensions in L2 speech, it appears that when assessing L2 speech, they may face challenges in maintaining distinct definitions for each dimension. Consequently, they may rely more on similar speech measures across all L2 rating dimensions.

6.2 Comparisons of L1 and L2 best-fit models

6.2.1 Similarities

In this study, we explored the speech measures that characterize less accented, more fluent, more comprehensible, and more pleasant L1 and L2 speech. We developed eight best-fit models to predict ratings for these dimensions in both L1 and L2 speech, allowing us to compare the salient speech characteristics in the perception of target speech across different rating dimensions.

There were some similarities between L1 and L2 best-fit models. Lower f0 mean, associated with low-pitched voices, was preferred in both L1 and L2 speech, suggesting a common preference for low-pitched speech. Moreover, both L1 and L2 speech were rated more positively when spoken at faster pace. This result reinforces a common finding that speech samples with a slower speech rate are often rated as more accented and less comprehensible L2 speech (e.g., Munro and Derwing 1995; Kang et al. 2010; O'Brien 2014). However, note that fast L2 speech does not always positively impact raters' judgments. Derwing and Munro (2001) suggested that there is a point beyond which an increase in speech rate is detrimental. In their study, the fastest L2 speech samples received worse ratings than samples with moderate rates close to an estimated optimal value. The stimuli used in our study had relatively slower speech rates than the optimal rates in Dewring and Munro (2001). This is a possible explanation for why the faster speech rates might consistently result in positive rating scores of all rating dimensions.

Both L1 and L2 speech with higher lexical richness received higher rating scores. This supports previous findings that the more affluent, more varied lexical content of L2

speech is associated with higher comprehensibility ratings (e.g., Trofimovich and Isaacs 2012). Raters perceived more lexically complex/richer L1 and L2 speech stimuli as less accented, more fluent, more comprehensible, and more pleasant speech. Compared to accentedness and pleasantness ratings, more lexical richness-related measures influenced L1 and L2 comprehensibility and fluency ratings. This result may provide supporting evidence to the idea that an L2 speaker who reaches a certain threshold of phonological, lexical, and grammatical ability can be highly comprehensible but still being fairly accented (Munro and Derwing 1995).

One interesting finding was that the CPP measure, related to breathiness of speech, influenced pleasantness ratings for both L1 and L2 speech. Lower CPP values, associated with breathy voices, were perceived as more pleasant. This aligns with previous research that suggests breathier voices are often perceived as more attractive and feminine (e.g., Van Borsel et al. 2009; Klug et al. 2019; Hejná et al. 2021). For example, Xu et al. (2013) reported that vocal attractiveness rated by listeners was enhanced by breathiness, indicating that listeners preferred breathy female and male voices. Although Xu et al. (2013) did not explicitly measure CPP of speech samples, together with other previous studies, we speculate that the prevalence of female speakers in our study (70% for L1 talkers and 71% for L2 talkers) may have contributed to this preference for breathier voices. Furthermore, L2 proficiency may have played a role in the breathiness of L2 speech. Lack of experience in speaking L2 may cause an increase in psycho-physiological stress and mental effort, which may lead to more muscle tension and, therefore, to increased pressedness of voice in L2 speech compared to L1 speech (Järvinen et al. 2017)[5]. Thus, it might be possible that breathier L2 speech samples in the current study sounded more natural and less nervous to the raters and resulted in more positive pleasantness ratings.

6.2.2 Differences

In the following discussion, we delve into the differences between the speech properties in L1 and L2 models that contributed to ratings across all four dimensions. Our primary

5 It should be noted that Järvinen et al. (2017) categorized L2 speakers' voices into two groups based on whether they reported experiencing increased vocal fatigue when shifting from L1 to L2. Those L2 speakers who did not experience increased vocal fatigue in L2 considered themselves more experienced in speaking L2, compared to those who did experience increased vocal fatigue.

goal is to understand whether raters weighed speech characteristics differently depending on whether they were evaluating L1 or L2 speech.

First, the rhythm-related measure, nPVI, was exclusively included in the L1 models. It had a significant impact on L1 speech ratings but did not influence ratings of any perceptual dimensions of L2 speech. This finding was unexpected because nPVI values between L1 and L2 speech samples were not significantly different (p = .514) (Table 2). It is possible that L2 speakers employed somewhat different speech strategies from L1 speakers or exhibited speech errors that eventually made their speech rhythm patterns appear similar to those of native speakers in terms of nPVI (Ross et al. 2008). To investigate this, we looked at the durational differences between tense and lax English vowel contrasts (e.g., English /i-ɪ/ contrast) as one possible source of duration variability in speech. Native English speakers typically maintain this duration contrast, with tense vowels being longer than lax vowels both in clear and conversational speech (e.g., Smiljanic and Bradlow 2008). L1 speakers in our study retained this durational difference between English tense /i/ and lax /ɪ/ vowels (109 ms vs. 80 ms, p = .04); however, L2 speakers did not exhibit the same duration contrast as L1 speakers (172 ms vs. 156 ms, p = .21). This preliminary analysis suggests that native-like nPVI of L2 stimuli might be from non-nativelike speech patterns and speech errors due to low L2 proficiency. Though it is still speculative, duration variability shown by nPVI might not be sufficient for the raters to consider for evaluating L2 speech (Ross et al. 2008).

Secondly, repair fluency had a stronger impact on L2 speech ratings compared to L1. L2 speakers in our study produced more repairs than L1 speakers, possibly due to the difficulty of the retelling task, which may have increased cognitive demand and led to more disfluencies. Consequently, due to more frequent instances of disfluencies and repairs, the raters may have been more generally affected by repair fluency when evaluating L2 speech than L1 speech. Another possible explanation for the stronger impact of repair fluency on L2 speech ratings may stem from differences L1 and L2 speakers differ in their preferences for specific disfluency markers in language planning. This is related to the distinct repair strategy patterns observed in L2 speakers, which are different from those of native speakers. For instance, Yu et al. (2018) showed that compared to English native speakers, Chinese learners of English used repetition rather than restart as their second primary repair strategy.

Lastly, the f0 range measure, related to pitch variability, was only included in L1 models and did not significantly impact L2 models. This was surprising based on

previous findings of the significant influence of pitch variations on L2 speech ratings. For instance, Kang (2010) reported that L2 speakers who presented a wider pitch range when speaking in L2 (i.e., English) were rated as less accented speakers than ones with narrower pitch ranges. However, our results showed that f0 range did not show a significant effect on any of the L2 perceptual dimensions. This suggests that L2 speakers in our study may not have effectively implemented native-like stress patterns in their speech. In a preliminary analysis, we examined whether the maximum f0 in L2 speech samples corresponded to the primary syllable(s) of the content words, which should be emphasized according to English stress patterns. According to Trofimovich and Baker (2006) in English, pitch peak often corresponds to a high-value tonal accent associated with a prominent syllable, usually in the most prominent word in an intonation phrase. Only nine out of twenty randomly selected L2 stimuli correctly placed the highest f0 on the primary syllable of the informative content word in their speech. Many L2 speakers in our study emphasized most of the words in a sentence, including function words, making it challenging for listeners to process the speech. This suggests that L2 speakers may not effectively use pitch variations to convey stress patterns, potentially explaining the lack of significance for f0 range measure in the L2 best-fit models.

7. Conclusion

The current study has investigated the similarities and differences between native listeners' evaluations of L1 and L2 speech. The rating experiments revealed that listeners consistently rated L1 speech higher than L2 speech, in line with previous research. Pearson correlation analysis showed that the perceptual dimensions in both L1 and L2 speech evaluations were related to each other, but the associations were stronger in L2 speech evaluations possibly due to listeners' lesser familiarity with L2 speech. The study identified specific speech properties that contributed to L1 and L2 speech ratings. Some properties were exclusive to either L1 or L2 models. The findings suggest that native listeners tend to be influenced by different aspects of speech depending on the type they are evaluating. Our study has implications for understanding the roles of certain overlapping speech properties in shaping listeners' holistic impressions of L1 and L2 speech, with some properties only affecting one type of speech.

Finally, the study's results hold significance for second language acquisition in

educational settings. The speech properties that most influenced native listeners' evaluations of L2 speech were speech rate, word types, f0 mean, and repair fluency. This suggests that focusing on improving these specific aspects of L2 speech may lead to overall improvements in fluency, comprehensibility, accentedness, and pleasantness. Enhancing L2 speech in these areas can contribute to more effective language learning and communication.

References

Baker, Rachel E., Melissa Baese-Berk, Laurent Bonnasse-Gahot, Midam Kim, Kristin J. van Engen, and Ann R. Bradlow. 2011. Word durations in non-native English. *Journal of Phonetics* 39(1): 1-17. https://doi.org/10.1016/j.wocn.2010.10.006

Boersma, Paul and David Weenink. 2017. *Praat, a system for doing phonetics by computer* [Computer program]. Version 6.0.43, retrived from http://www.praat.org./.

Bonfiglio, Thomas Paul. 2002. *Race and the rise of standard American (Vol. 7).* Berlin; New York: Walter de Gruyter. https://doi.org/10.1515/9783110851991

Bosker, Hans Rutger, Anne-France Pinget, Hugo Quené, Ted Sanders, and Nivja H. de Jong. 2013. What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing* 30(2): 159-175. https://doi.org/10.1177/0265532212455394

Bosker, Hans Rutger, Hugo Quené, Ted Sanders, and Nivja H. de Jong. 2014a. Native 'um's elicit prediction of low-frequency referents, but non-native 'um's do not. *Journal of Memory and Language* 75: 104-116. https://doi.org/10.1016/j.jml.2014.05.004

Bosker, Hans Rutger, Hugo Quené, Ted Sanders, and Nivja H. de Jong. 2014b. The perception of fluency in native and nonnative speech. *Language Learning* 64(3): 579-614. https://doi.org/10.1111/lang.12067

Bosker, Hans Rrtger and Eva Reinisch. 2015. Normalization for speech rate in native and non-native speech. In The Scottish Consortium for ICPhS 2015 (eds.), *Proceedings of the 18th International Congresses of Phonetic Sciences (ICPhS)*, 3024. Glasgow, UK: The University of Glasgow. August 10-14.

Bouchard Ryan, Ellen, Carranza Miguel A., and Robert W. Moffie. 1977. Reactions toward varying degrees of accentedness in the speech of Spanish-English bilinguals. *Language and Speech* 20(3): 267-273. https://doi.org/10.1177/002383097702000308

Buder, Eugene H. 2000. Acoustic analysis of voice quality: A tabulation of algorithms 1902-1990. In Raymond D. Kent and Martin John Ball (eds.), *Voice quality measurement*, 119-244. San Diego, CA: Singular.

Bulté, Bram and Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex

Housen, Ineke Vedder, and Folkert Kuiken (eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA,* 21-46. Amsterdam; Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/lllt.32.02bul

Clopper, Cynthia G and David B. Pisoni. 2006a. Effects of region of origin and geographic mobility on perceptual dialect categorization. *Language Variation and Change* 18(2): 193-221. https://doi.org/10.1017/S0954394506060091

Clopper, Cynthia G. and David B. Pisoni. 2006b. The nationwide speech project: A new corpus of American English dialects. *Speech Communication* 48(6): 633-644. https://doi.org/10.1016/j.specom.2005.09.010

Crossley, Scott A., Kristopher Kyle, and Danielle S. McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods* 48: 1227-1237. https://doi.org/10.3758/s13428-015-0651-7

Darcy, Isabelle, Hanyong Park, and Chung-Lin Yang. 2015. Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences* 40: 63-72. https://doi.org/10.1016/j.lindif.2015.04.005

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

Derwing, Tracey M. and Murray J. Munro. 2009. Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *The Canadian Modern Language Review* 66(2): 181-202. https://doi.org/10.3138/cmlr.66.2.181

Derwing, Tracey M., Marian J. Rossiter, Murray J. Munro, and Ron I. Thomson. 2004. Second language fluency: Judgments on different tasks. *Language Learning* 54(4): 655-679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Feng, Chuyao, Eva van Leer, Mackenzie Lee Curtis, and David V. Anderson. 2021. I-vector Based within speaker voice quality identification on connected speech. *arXiv:2102.07307 [Cs, Eess]*. http://arxiv.org/abs/2102.07307

Foster, Pauline and Parvaneh Tavakoli. 2009. Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning* 59(4): 866-896. https://doi.org/10.1111/j.1467-9922.2009.00528.x

Giles, Howard. 1970. Evaluative reactions to accents. *Educational Review* 22(3): 211-227. https://doi.org/10.1080/0013191700220301

Grabe, Esther and Ee Ling Low. 2002. Durational variability in speech and the rhythm class hypothesis. In Carlos Gussenhoven and Natasha Warner (eds.), *Laboratory Phonology 7*, 515-546. Berlin; New Work: De Gruyter Mouton. https://doi.org/10.1515/9783110197105.2.515

Gut, Ulrike. 2012. Rhythm in L2 speech. *Speech and Language Technology* 14(15): 83-94.

Hejná, Míša, Pavel Šturm, Lea Tylečková, and Tomáš Bořil. 2021. Normophonic breathiness in Czech and Danish: Are females breathier than males? *Journal of Voice* 35(3): 498.e1-498.e22. https://doi.org/10.1016/j.jvoice.2019.10.019

Hillenbrand, James, Ronald A. Cleveland, and Robert L. Erickson. 1994. Acoustic correlates of breathy vocal quality. *Journal of Speech Language and Hearing Research* 37(4): 769-778. https://doi.org/10.1044/jshr.3704.769

Hillenbrand, James and Robert A Houde. 1996. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech and Hearing Research* 39(2) 311-321. https://doi.org/10.1044/jshr.3902.311

Ilie, Gabriella and William Forde Thompson. 2006. A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception* 23(4): 319-330. https://doi.org/10.1525/mp.2006.23.4.319

Iwashita, Noriko, Annie Brown, Tim McNamara, and Sally O'Hagan. 2008. Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics* 29(1): 24-49. https://doi.org/10.1093/applin/amm017

Järvinen, Kati, Anne-Maria Laukkanen, and Ahmed Geneid. 2017. Voice quality in native and foreign languages investigated by inverse filtering and perceptual analyses. *Journal of Voice* 31(2): 261.e25-261.e31. https://doi.org/10.1016/j.jvoice.2016.05.003

Jassem, Wiktor. 1971. Pitch and compass of the speaking voice. *Journal of the International Phonetic Association* 1(2): 59-68. https://doi.org/10.1017/S0025100300000256

John, Van Borsel, Janssens Joke, and Marc De Bodt. 2009. Breathiness as a feminine voice characteristic: A perceptual approach. *Journal of Voice* 23(3): 291-294. https://doi.org/10.1016/j.jvoice.2007.08.002

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

Kang, Okim. 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, Okim, Don Rubin, and Lucy Pickering. 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

Klug, Katharina, Christin Kirchhübel, Paul Foulkes, and Peter French. 2019. Analysing breathy voice in forensic speaker comparison Using acoustics to confirm perception. In Sasha Calhoun, Paola Escudero, Marija Tabain, and Paul Warren (eds.), *Proceedings of the 19th International Congress of Phonetic Sciences,* 795-799. Canberra: Australasian Speech Science and Technology Association Inc.

Kormos, Judit and Mariann Dénes. 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

Lee, Jieun, Dong Jin Kim, and Hanyong Park. 2019. Native listener's evaluations of pleasantness, foreign accent, comprehensibility, and fluency in the speech of accented talkers. In John Levis, Charles Nagle, and Erin Todey (eds.), *Proceedings of the 10th Annual Pronunciation in Second Language Learning and Teaching Conference*, 168-178. Ames, IA: Iowa State University.

Lee, Shinhye. 2018. *Effective planning in real-time speaking test tasks*. PhD Dissertation. East Lansing, MI: Michigan State University.

Maryn, Youri, Nelson Roy, Marc De Bodt, Paul Van Cauwenberge, and Paul Corthals. 2009. Acoustic measurement of overall voice quality: A meta-analysis. *The Journal of the Acoustical Society of America* 126(5): 2619-2634. https://doi.org/10.1121/1.3224706

Menn, Lise and Suzanne Boyce. 1982. Fundamental frequency and discourse structure. *Language and Speech* 25(4): 341-383. https://doi.org/10.1177/002383098202500403

Mori, Hiroki, Tomoyuki Satake, Makoto Nakamura, and Hideki Kasuya. 2011. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication* 53(1): 36-50. https://doi.org/10.1016/j.specom.2010.08.002

Munro, Murray J. and Tracey M. Derwing. 1995. Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech* 38(3): 289-306. https://doi.org/10.1177/002383099503800305

Munro, Murray J. and Tracey M. Derwing. 1999. Foreign accent, comprehensibility, and intelligibility in the speech of Second Language learners. *Language Learning* 49(s1): 285-310. https://doi.org/10.1111/0023-8333.49.s1.8

Munro, Murray J. and Tracey M. Derwing. 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

Nava, Emily and Maria Luisa Zubizarreta. 2008. Prosodic transfer in L2 speech: Evidence from phrasal prominence and rhythm. *Proceedings of Speech Prosody 2008*, 335-338. May 6-9. https://doi.org/10.21437/SpeechProsody.2008-75

Niebuhr, Oliver, Radek Skarnitzl, and Lea Tylečková. 2018. The acoustic fingerprint of a charismatic voice-Initial evidence from correlations between long-term spectral features and listener ratings. *Proceedings of Speech Prosody 2018*, 359-363. June 13-16. http://doi.org/10.21437/SpeechProsody.2018-73

O'Brien, Mary Grantham. 2014. L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning* 64(4): 715-748. https://doi.org/10.1111/lang.12082

Ortega, Lourdes. 1999. Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition* 21(1): 109-148. https://doi.org/10.1017/S0272263199001047

Preston, Dennis R. 1996. Where the worst English is spoken. In Edgar W. Schneider (ed.), *Focus on the USA*, 297-360. Amsterdam; Philadelphia: John Benjamins Publishing

Company. https://doi.org/10.1075/veaw.g16.16pre

Robinson, Peter. 2011. *Second Language task complexity: Researching the cognition hypothesis of language learning and performance*. Amsterdam; Philadelphia: John Benjamins Publishing. https://doi.org/10.1075/tblt.2

Ross, Tristie, Naja Ferjan, and Arvaniti Arvaniti. 2008. Speech rhythm and its quantification in L2. Presented at *the 1st Southern California workshop on Phonetics/Phonology (SCOPHO)*. Pomona College. November 1.

Saito, Kazuya, Meltem Ilkan, Viktoria Magne, Mai Ngoc Tran, and Shungo Suzuki. 2018. Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics* 39(3): 593-617. https://doi.org/10.1017/S0142716417000571

Saito, Kazuya, Pavel Trofimovich, and Talia Isaacs. 2016. Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics* 37(2): 217-240. https://doi.org/10.1017/S0142716414000502

Saito, Kazuya, Pavel Trofimovich, and Talia Isaacs. 2017. Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics* 38(4): 439-462. https://doi.org/10.1093/applin/amv047

Saito, Kazuya, Stuart Webb, Pavel Trofimovich, and Talia Isaacs. 2016a. Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition* 19(3): 597-609. https://doi.org/10.1017/S1366728915000255

Saito, Kazuya, Stuart Webb, Pavel Trofimovich, and Talia Isaacs. 2016b. Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations. *Studies in Second Language Acquisition* 38(4): 677-701. https://doi.org/10.1017/S0272263115000297

Scherer, Klaus R. and James S. Oshinsky. 1977. Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion* 1: 331-346. https://doi.org/10.1007/BF00992539

Shue, Yen-Liang, Gang Chen, and Abber Alwan. 2010. On the interdependencies between voice quality, glottal gaps, and voice-source related acoustic measures. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura (eds.), *Proceedings of Eleventh Annual Conference of the Internatioanl Speech Communication Association,* 34-37. September 26-30. https://doi.org/10.21437/Interspeech.2010-5

Smiljanic, Rajka and Ann R. Bradlow. 2008. Stability of temporal contrasts across speaking styles in English and Croatian. *Journal of Phonetics* 36(1): 91-113. https://doi.org/10.1016/j.wocn.2007.02.002

Suzuki, Shungo and Judit Kormos. 2020. Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition* 42(1): 143-167. https://doi.org/10.1017/S0272263119000421

Towell, R., R. Hawkins, and N. Bazergui. 1996. The development of fluency in advanced learners of French. *Applied Linguistics* 17(1): 84-119. https://doi.org/10.1093/applin/17.1.84

Trofimovich, Pavel and Wendy Baker. 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

Trofimovich, Pavel and Talia Isaacs. 2012. Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition* 15(4): 905-916. https://doi.org/10.1017/S1366728912000168

Watts, Christopher and Shaheen N. Awan. 2011. Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts. *Journal of Speech, Language, and Hearing Research* 54(6): 1525-1537. https://doi.org/10.1044/1092-4388(2011/10-0209)

Xu, Yi, Albert Lee, Wing-Li Wu, Xuan Liu, and Peter Birkholz. 2013. Human vocal attractiveness as signaled by body size projection. *PLoS ONE* 8(4): e62397. https://doi.org/10.1371/journal.pone.0062397

Yu, Jue, Jiena Chen, Shengyi Wu, and Ye Feng. 2018. Disfluency in Chinese L2 spontaneous speech: Patterns and interactions. In Katarzyna Klessa, Jolanta Bachan, Agnieszka Wagner, Maciej Karpiński, and Daniel Śledziński (eds.), *Proceedings of Speech Prosody 2018*, 863-867. June 13-16. https://doi.org/10.21437/SpeechProsody.2018-174

Yuan, Zhongshang, Hong Liu, Xiaoshuai Zhang, Fangyu Li, Jinghua Zhao, Furen Zhang, and Fuzhong Xue. 2013. From interaction to co-association—A Fisher *r*-to-*z* transformation-based simple statistic for real world genome-wide Association Study. *PLoS ONE* 8(7): e70774. https://doi.org/10.1371/journal.pone.0070774

Zetterholm, Elisabeth and Åsa Abelin. 2017. Swedish and Somali listeners' attitudes towards L2 Swedish speech. *Scandinavian Philology* 15(2): 193-203. https://doi.org/10.21638/11701/spbu21.2017.203

Zhou, Ziwei and Charles Nagle. 2018. Acoustic correlates of L2 Spanish judgments of accentedness and comprehensibility: A mixed-effects modeling approach. *The Journal of the Acoustical Society of America* 143: 1949. https://doi.org/10.1121/1.5036387

Jieun Lee
Visiting Assistant Professor
Department of Linguistics
University of Kansas
1541 Lilac Lane
Lawrence, KS 66045, USA
E-mail: jieunlee@ku.edu

Dong Jin Kim
Ph.D. Candidate
Department of Linguistics
University of Wisconsin-Milwaukee
2522 East Hartford Ave.
Milwaukee, WI 53211, USA
E-mail: kim252@uwm.edu

Hanyong Park
Associate Professor
Department of Linguistics
University of Wisconsin-Milwaukee
2522 East Hartford Ave.
Milwaukee, WI 53211, USA
E-mail: park27@uwm.edu

Received: 2023. 12. 01.
Revised: 2024. 02. 20.
Accepted: 2024. 02. 20.