To be published in *Studies in Second Language Acquisition* (Cambridge University Press)

Studies in Second Language Acquisition

CAMBRIDGE

Automated Assessment of Second Language Comprehensibility: Review, Training, Validation, and Generalization Studies

Kazuya Saito, Konstantinos Macmillan, Magdalena Kachlicka, Takuya Kunihara, & Nobuaki Minematsu

Abstract

Whereas many scholars have emphasized the relative importance of *comprehensibility* as an ecologically valid goal for L2 speech training, testing, and development, eliciting listeners' judgements is time-consuming. Following calls for research on more efficient L2 speech rating methods in applied linguistics, and growing attention towards using machine learning on spontaneous unscripted speech in speech engineering, the current study examined the possibility of establishing quick and reliable *automated* comprehensibility assessments. Orchestrating a set of phonological (maximum posterior probabilities and gaps between L1 and L2 speech), prosodic (pitch and intensity variation), and temporal measures (articulation rate, pause frequency), the regression model significantly predicted how naïve listeners intuitively judged low, mid, high, and nativelike comprehensibility among 100 L1 and L2 speakers' picture descriptions. The strength of the correlation (r = .823 for machine vs. human ratings) was comparable to naïve listeners' inter-rater agreement (r = .760 for humans vs. humans). The findings were successfully replicated when the model was applied to a new dataset of 45 L1 and L2 speakers (r = .827); and tested under a more freely constructed interview task condition (r = .809).

Key words: Automated Assessment, Second Language Speech, Comprehensibility, Pronunciation

Adult second language (L2) speech is generally foreign-accented due to a range of factors, such as the influence of first language (L1) phonetic systems (Flege & Bohn, 2020), perceptual-cognitive aptitude (Saito, 2017), and identity (Sung, 2016). Thus, many scholars have emphasized the importance of setting realistic goals for adult L2 learners, prioritizing understanding over nativelikeness (Munro & Derwing, 1995). There is ample evidence that many L2 speakers are perceived as sufficiently comprehensible regardless of foreign accentedness (Isaacs & Trofimovich, 2012); and L2 speakers can continue to enhance aspects of speech affecting comprehensibility as long as they use the target language, receive feedback, and strive to improve with a view of successful communication (Saito & Akiyama, 2017). Although teachers play an important role in providing immediate feedback, helping students understand their own comprehensibility, and promoting autonomous L2 speech learning, such a resource (provision of feedback) is limited in foreign language classrooms (Muñoz, 2014). In the field of speech engineering, there is a growing amount of research attention towards the development of more robust automatic recognition of not only controlled but also spontaneous L2 speech, and the application of the technology to L2 speech training and evaluation (Fu, Chiba, Nose, & Ito, 2019). Interfacing perspectives from education and speech sciences, we took a first step towards training, validating, and generalizing the automatic assessment of comprehensibility in the context of 190 L1 and L2 speakers of English.

Second Language Comprehensibility

Given that technology allows people to interact worldwide regardless of physical constraints using videoconferencing tools and social networking, attaining adequate L2 speech proficiency is considered a key skill in academic, business, and social settings. On the one hand, some L2 users strive to attain nativelike proficiency (Scales, Wennerstrom, Richard, & Wu, 2006) and traditional teaching syllabi highlight native speakers as an ideal instructional model (Foote, Holtby, & Derwing, 2010). On the other hand, research has convincingly shown that few post-pubertal L2 learners can attain nativelike phonological accuracy and fluency as their L2 system builds on and thus inevitably interacts with their already-developed first language system (Flege & Bohn, 2020). When it comes to English, attaining nativelike phonological proficiency is arguably unnecessary as most interactions take place between L2 users themselves (Pennycook, 2017). Thus, a number of scholars have pointed out the importance of setting more realistic goals for post-pubertal L2 speech learning, such as the enhancement of comprehensibility rather than the attainment of nativelike proficiency (Munro & Derwing, 1995).

From a methodological point of view, simulating behaviors in real-life conversation, L2 comprehensibility is operationalized as listeners' *intuitive* judgements of spontaneous L2 speech on a 9-point scale (*1 = difficult to understand, 9 = easy to understand*). Comprehensibility is thought to represent the amount of listener effort necessary to understand the speakers' message despite the degree of foreign accentedness; and the *process* but not *product* of listeners' understanding (for the further discussion on "intelligibility" rather than "comprehensibility" as a barometer of *actual* understanding and its terminological and methodological issues, see Levis, 2018). According to Derwing and Munro's seminal work, accented L2 speech can be perceived to be comprehensible as certain phonological errors do not hinder listeners' understanding (e.g., Munro & Derwing, 1995). Since Munro and Derwing (1995), there have been a number of follow-up studies showing that listeners' understanding is negatively influenced by certain (but not all) phonological errors, such as the mispronunciation of segmentals with high functional load (Suzukida & Saito, 2019), melodic inaccuracies (Kang, Rubin, & Pickering, 2010), and dysfluencies (Suzuki & Kormos, 2020; for a meta-analysis of the phonological correlates of L2 comprehensibility, see Saito, 2021). Additionally, certain listeners likely assign higher and thus more lenient comprehensibility ratings when they have more familiarity with foreign accents (Kennedy & Trofimovich, 2008), linguistic training (Saito, Trofimovich, & Isaacs, 2017), pedagogical experience (Isaacs & Thomson, 2020), and a greater level of awareness of the importance of L2 comprehensibility (rather than accentedness) (Saito, Tran, Sun, Magne, & Ilkan, 2019).

From a theoretical perspective, it is important to note that comprehensibility can serve as an index of adult L2 speech development. As stated in the interaction account of L2 acquisition (Mackey, 2012), language learning takes place precisely when L2 speakers actively participate in conversational interactions and end up in communication breakdowns due to linguistic errors. L2 speakers work together with interlocutors when comprehensibility is not sufficient by relying on a range of negotiation-for-meaning behaviors, such as clarification requests, and comprehension and confirmation checks. This whole sequence is hypothesized to help L2 speakers become more comprehensible, functional, and proficient users of the target language (for an empirical evidence, see Saito & Akiyama, 2017)[1]. According to longitudinal (Derwing & Munro, 2013) and cross-sectional investigations (Saito, 2015a), L2 learners tend to show quick improvements in comprehensibility and accentedness within the first few years of immersion. Whereas attaining *nativelike* L2 speech proficiency may be limited to certain individuals with earlier ages of acquisition (Saito, 2015b), linguistically similar L1 backgrounds (e.g., Bongaers, van Summeren, Planken, & Schils, 1997) and/or special language aptitude (e.g., He et al., 2013 for phonemic coding; Kachlicka, Saito, & Tierney, 2019 for perceptual acuity), many L2 learners can continue to enhance the comprehensibility aspects of their speech over time as long as they use a target language on a regular basis. Similar learning patterns have been observed in various foreign language classroom settings (e.g., Nagle, 2018).

Due to the significant amount of practical and theoretical relevance, many scholars have promoted comprehensibility as an ecologically valid, more realistic target for adult L2 speech assessment not only in high-stakes testing settings, but also for practitioners (Isaacs, Trofimovich, & Foote, 2017). However, eliciting listeners' comprehensibility judgements is a time-consuming task (e.g., about 1 hour for rating 70 samples in Derwing & Munro, 2013). To

[1] An interaction account of L2 acquisition would indicate that language development occurs when communication breaks down and interlocutors are required to negotiate for meaning (Mackey, 2012). In this paper and elsewhere (e.g., Saito, 2021), therefore, it has been argued that comprehensibility is an index of L2 development. Negotiation-for-meaning episodes likely take place when L2 speakers' comprehensibility is low. Through more interaction and negotiation-for-meaning opportunities, learners with low L2 oral proficiency are pushed to work on the comprehensibility of their speech, even though they still sound foreign-accented. In this regard, it is reasonable to say L2 speech acquisition takes place on the continuum of comprehensibility as learners' speech becomes increasingly easier to understand despite their detectable foreign accentedness. To provide empirical support, Derwing and Munro's (2013) longitudinal dataset showed that the comprehensibility of L2 speakers continued to improve as a function of increased immersion for an extensive period of time (7 years) while their speech remained foreign-accented throughout (cf. Saito, 2015a for cross-sectional evidence; Saito & Akiyama, 2017 for a long-term training study). A reviewer pointed out that what actually matters is intelligibility rather than comprehensibility. Research has indeed shown that speech can be low in comprehensibility yet still intelligible (i.e., even if difficult to understand, speech can be accurately understood). Technically, it should only be speech that is low in intelligibility (which likely entails low comprehensibility) that would initiate a degree of negotiation of meaning. However, we would like to argue that negotiation for meaning can be triggered by comprehensibility and intelligibility alike. In the case of low-level comprehensible but still intelligible speech, interlocutors likely initiate a range of particular interactional moves such as clarification requests and repetition, both of which are termed as "negotiation for meaning" strategies (Mackey, 2012). To our knowledge, few studies have ever explored precisely when negotiation for meaning occurs in accordance with different levels of interlocutors' understanding, i.e., high vs. low comprehensibility and high vs. low intelligibility. This should be considered as an intriguing direction for future research.

this end, where some researchers use shorter stimuli to avoid listener fatigue (e.g., 4.5-10.5 seconds in Derwing & Munro, 1997), others recommend collecting L2 rating data via online platforms to reduce the burden on researchers (e.g., Nagle & Rehman, 2021 for Amazon Mechanical Turk). Therefore, due to the time-consuming nature of the assessment procedure, a growing amount of attention has been given to the idea of using automated L2 comprehensibility scoring (O'Brien et al., 2018). To this end, the current study took an exploratory approach towards examining this topic.

Automatic Assessment of Second Language Speech

To date, research has provided a range of pedagogic techniques that teachers can use (Pennington, 2021) and online learning tools that learners can access (Loewen et al., 2019). To make the most of such opportunities, provision of online feedback plays a critical role for many reasons. First and foremost, learners need to understand the current state of their L2 proficiency, compare it with their target, and make an effort to fill in the gaps accordingly (Lyster & Saito, 2010). Without such adequate self- and teacher-assessments, learners may either overestimate or underestimate their own L2 proficiency and struggle to decide the best course of action, slowing down acquisitional processes (Trofimovich, Isaacs, Kennedy, Saito, & Crowther, 2015).

Although feedback is integral to successful L2 speech learning, the provision of feedback is a resource-intensive task, especially in foreign language classrooms. Not only are students' L2 practice opportunities limited to several hours of weekly classroom instruction (which is typically devoid of conversation activities; Nishino & Watanabe, 2008), teachers also lack ample time to record, listen to and provide feedback on each student's speech, given that they often teach many students at one time (Muñoz, 2014). Thus, many researchers and practitioners alike are interested in the introduction of automated assessment as a pedagogical tool for optimal L2 speech education (O'Brien et al., 2018).

As summarized in Table 1, a range of studies have been conducted to examine the extent to which automated scoring can simulate human listeners' assessments of L2 oral proficiency. In terms of method, scholars have primarily drawn on correlational analyses. First, L2 speech stimuli were collected via a range of speaking tasks (controlled, spontaneous). Then, the stimuli were evaluated by trained and untrained human listeners for various aspects of L2 oral proficiency rubrics (e.g., pronunciation accuracy, overall speaking proficiency, perceived fluency and comprehensibility). Finally, the stimuli were submitted to a range of automated speech measures (phonological and fluency analyses). To examine the potential of automated L2 speech assessment, researchers explored the strength of the associations between human L2 oral proficiency ratings and automated speech scores. With respect to automated speech measures, there is much methodological variation in the primary studies. Although each study adopted different measures, they could be roughly categorized into three subgroups with a view of methodological syntheses—i.e., fluency, melodic, and phonological measures. Relevant citations below derive from Table 1 (i.e., a summary of automated L2 speech assessment studies).

Fluency Measures. This subcategory refers to a set of outcome measures that tap into the temporal characteristics of speech. As stated in Tavakoli and Skehan's (2005) model, such fluency features can be categorized as speed (e.g., speech and articulation rate), breakdowns (e.g., filled and unfilled pause frequency), and repairs (e.g., repetition and self-correction frequency). Whereas some studies covered all three of these aspects of fluency (e.g., Cucchiarini et al., 2000 for speed, breakdown, and repair), other studies focused on one dimension (e.g., Neumeyer at al., 2000 for speed). Within the same subgroup (e.g., breakdown), the measures were operationalized differently (e.g., Cucchiarini et al, 2000 for the frequency of silent pauses [defined as > 20ms] per sample vs. Kang & Johnson, 2018 for the frequency of silent pauses [defined as > 100ms]).

Melodic Measures. This subcategory refers to a set of outcome measures related to the *varied* use of pitch and intensity to mark stress and intonation. Some scholars have adopted different types of pitch measures, such as tone choices and relative pitch (Kang & Johnson, 2018). Others have also adopted both pitch and amplitude measures, such as pitch and intensity variation (e.g., van Santen et al., 2009). It is important to acknowledge that we do not yet know how to conduct the automated analyses of word and sentence stress *accuracy*. In the field of L2 speech, such analyses require linguistically trained coders to examine how each instance of word and sentence stress has been marked (with higher pitch, longer duration, and/or greater intensity) in a contextually appropriate manner (e.g., Isaacs & Trofimovich, 2012 for the *manual* analyses of word stress and intonation accuracy).

Phonological Measures. Thanks to advancements in automatic speech recognition (ASR) technology, one notable change in this line of work concerns the development, sophistication, and application of machine algorithms to calculate phonetic accuracy scores (i.e., qualitative measures; Zechner & Evanini, 2020). In ASR, speech classes are generally defined in a more sophisticated way compared to linguistically-driven phonemes. Speech samples are first segmented into small frames (e.g., approximately 20ms) and converted into spectrum (i.e., different frequencies) via a Fast Fourier Transform. Since acoustic realization of phoneme /x/ depends on its phonemic environment, /x/ is further divided into a full set of /a/-/x/+/b/. This means acoustically realized phoneme /x/ produced after /a/ and before /b/. When the number of linguistically-driven phonemes is N, the full set of /a/-/x/+/b/ has N^2 variants. All of them are considered as the speech classes related to /x/.

In the classical ASR framework based on Hidden Markov Models (HMM), a number of spectrograms were collected from thousands of L1 speakers in order to build an acoustic model for each speech class. Since the spectrogram was used as a fundamental speech representation, it inevitably conveyed a range of extra-linguistic factors, such as speaker, age, gender, and microphone. These factors in turn clouded the accuracy of ASR.

Recently, a more sophisticated framework for ASR has been used based on Deep Neural Networks (DNN). Various frameworks have been proposed with DNN and one of them uses posteriorgram (rather than spectrogram) as a fundamental speech representation. A posteriorgram is defined as a temporal sequence of posterior probability distributions of classes, demonstrating the probability that an input observation belongs to a particular class. Unlike spectogram the extra-linguistic factors are well suppressed. Whereas the spectrograms of "hello" generated by a male speaker and a female speaker can be substantially different, the posteriorgrams of the same speech samples can be indistinguishable. If we use the linguistically-driven phonemes as the speech classes, a posteriorgram can be viewed as a probabilistic version of its phonemic transcript (i.e., how similar the speech signals are to each phonemic class). Unlike the spectrogram (which provides an acoustic representation of speech), a posteriorgram can serve as a linguistic or phonetic representation of speech.[2]

Once ASR models are established in a target language, they are applied to analyze the acoustic profiling of L2 speech data. Looking at the posteriorgram distance between L1 and L2 speech data, for example, some studies have explored the overall acoustic similarity of L2 realizations of phonemes to their L1 equivalents. For a more detailed account of posteriorgram-

[2] Whereas scholars initially used HMM for more precise and reliable ASR, DMM has become increasingly dominant in more recent literature. For detailed accounts of HMM and DNN models, see **Supporting Information A**.

based L1 vs. L2 phoneme categorizations, see Shen et al. (2021) and **Supporting Information-B**. To further increase the accuracy of ASR, another intriguing idea concerns evaluating the quality of L2 speech via L1 and L2 corpus data. For example, the gap among Japanese speakers of English in Fu et al. (2019) was assessed using the outcomes from ASR trained on L1 Japanese data and those from L2 English data. L2 English speech with low word-error rate in both L1 Japanese and L2 English ASR (the gap is nil or small), could be considered highly proficient. In contrast, if word error rate was low in L1 Japanese ASR but high in L2 English ASR (the gap is large), such samples could be considered less proficient.

Overall Findings. Once again, we would like to remind the readers that the primary studies operationalized both L2 speech ratings and automated measures in a substantially different fashion. Thus, the intention of the methodological synthesis is to provide overall patterns. All in all, while human listeners' agreement is generally strong (r = .7-.9), the relationship between human and automated scoring is comparable (r = .6-.8). Earlier studies demonstrated strong correlations between human and machine evaluations of L2 speech elicited from controlled speech tasks (e.g., word and sentence reading). More recent studies have explored the replicability of the findings, focusing on more spontaneous speech samples (e.g., monologues, interviews, picture descriptions). Thus far, scholars have convincingly shown that L2 speech proficiency scores were strongly associated with the results of automated analyses of temporal features in L2 speech (e.g., Ginther, Dimova, & Yang, 2010).

Table 1

Summary of 11 Key Studies on Automated L2 Speech Assessment

	Speaking materials	Human listeners	Automated assessment	Findings
Cucchiarini, Strik, & Boves (2000)	Speakers • 60 L2 Dutch speakers (Beginner, Intermediate, Advanced) • 20 L1 Dutch speakers Materials • 10 sentences (1 minute per speaker)	Listeners • 9 trained listeners (phoneticians, speech therapists) Rubrics • Intuitive judgements of fluency (10-point scale)	Fluency measures • Speed • Breakdown • Repair	Machine vs. humans • *r* = .94 Human listeners • *α* = .76-.97
Neumeyer, Franco, Digalakis, & Weintraub (2000)	Speakers • 100 L2 speakers Materials • 3000 sentences	Listeners • 5 trained listeners Rubrics • Expert judgements of pronunciation accuracy (5-point scale)	Phonological measures • HMM log-likelihood scores • Phone-normalized scores • Segment classification Fluency measures • Segment duration (speed) • Normalized syllable timing (speed)	Machine vs. humans • *r* = .8 Human listeners • *r* = .7-.8
Cucchiarini, Strik, & Boves, 2002[a]	Speakers • 60 L2 Dutch speakers (beginner, intermediate) Materials • 60 monologues	Listeners • 10 trained listeners Rubrics • Intuitive judgements of fluency (10-point scale)	Fluency measures • Speed • Breakdown • Repair	Machine vs. humans • *r* = .8 Human listeners • *r* = .7-.8
Moustroufas & Digalakis (2007)	Speakers • 20 L2 English speakers Materials 114 sentences	Listeners • 3 trained listeners Rubrics Expert judgements of pronunciation accuracy (5-point scale)	Phonological measures • HMM log-likelihood scores	Machine vs. humans • *r* = .7-.8 Human listeners *r* = .6-.8
van Santen, Prud'hommeaux, & Black (2009)	Speakers • 15 autistic children Materials • A range of speaking tasks	Listeners • 5 naïve listeners Rubrics • Affect judgements	Melodic measures • Fundamental-frequency-based differences	Machine vs. humans • *r* = .8 Human listeners • *r* = .7-.8

		• Minimal pair judgments	• Amplitude-based differences Fluency measures • Duration-based differences	
Zechner, Higgins, Xi, & Williamson (2009)	Materials • 7000+ spontaneous speech samples	Listeners • Trained TOEFL listeners Rubrics • Speaking proficiency scores (5-point scale)	Phonological measures • HMM acoustic model scores (segmental accuracy) Fluency measures • Speed • Breakdown • Repair	Machine vs. humans • *r* = .57-.68 Human listeners • *r* = .74-.94
Ginther, Dimova, & Yang (2010)	Speakers • 150 L2 English speakers Materials • 150 monologues	Listeners • 2 trained listeners Rubrics • Expert judgements of oral proficiency (4-point scale)	Fluency measures • Speed • Breakdown • Repair	Machine vs. humans • *r* = .7-.8 Human listeners • *r* = .6-.8
Kang & Johnson (2018)	Speakers • 120 L2 English speakers (intermediate to advanced) Materials • 120 monologues	Listeners • 2 trained listeners Rubrics • Expert judgements of oral proficiency (100-point scale)	Melodic measures • Prominence • Intonation Fluency measures • Speed • Breakdown	Machine vs. humans • *r* = .718
Chen et al. (2018)	Speakers • 1000 L2 English speakers Materials • 7000 monologues	Listeners • Trained TOEFL listeners Rubrics • Expert judgements of oral proficiency scores (5-point scale)	Phonological measures • HMM acoustic model scores (segmental accuracy) Fluency measures • Speed • Breakdown	Machine vs. humans • *r* = .77 Human listeners • *r* = .88
Fu, Chiba, Nose, & Ito (2019)	Speakers • 202 L2 English speakers (controlled) • 630 L1 English speakers (controlled)	Listeners • 5 trained listeners (controlled) • 3 trained listeners (semi-controlled)	Phonological measures • DNN acoustic model scores	Machine vs. humans • *r* = .826 (controlled) • *r* = .799 (spontaneous) Human listeners

	• 13 L2 English speakers (semi-controlled) • 14 L2 English speakers (spontaneous) Materials • 190 sentence reading tasks • 26 constrained interactive tasks • 28 spontaneous conversation tasks	• 3 trained listeners (spontaneous) Rubrics • Expert judgements of segmental proficiency scores (5-point scale)		• r = .7-.8
Shen et al. (2021)[b]	Speakers • 100 L2 English speakers Materials • 100 picture descriptions tasks	Listeners • 10 native speakers Rubrics • Intuitive judgements of fluency (9-point scale)	Phonological measures • DNN acoustic model scores	Machine vs. humans • r = .8-.9 (spontaneous) Human listeners • r = .7-.8

Note. [a] features Experiment 2; [b] serves as a pilot study for the current study (see the Method section)

Motivation for the Current Study

Throughout the past 15 years, many scholars have extensively examined what contributes to naïve listeners' perceptions of foreign-accented yet sufficiently comprehensible speech (Munro & Derwing, 1995). Such intuitive judgements of comprehensibility, intelligibility, and communicative competence (rather than nativelikeness) are believed to play a key role in communicative success in today's globalized society in which most communication in English takes place between L2 speakers (Pennycook, 2017). From a theoretical standpoint, comprehensibility serves as a crucial index of adult L2 speech development as learners' oral proficiency continues to become comprehensible rather than nativelike as a function of increased practice and immersion experience (Derwing & Munro, 2013; Saito, 2015a, 2015b). The quick, reliable, and automatic assessment of L2 comprehensibility has been strongly called for among practitioners (to help students achieve comprehensible L2 speech via feedback; Trofimovich et al., 2015) and researchers (to assess the different stages of adult L2 speech learning; Isaacs et al., 2017).

To advance the agendas of L2 comprehensibility and automatic assessment research, the current study serves as a first attempt to explore the extent to which machine-based assessments can simulate naïve listeners' comprehensibility ratings of 100 L1 and L2 speakers' semi-spontaneous speech (Study 1). Subsequently, we further delved into the validity, replicability, and generalizability of the regression model when it was applied to new speakers (Study 2); and tested under a more freely constructed interview task condition (Study 3).

Whereas some prior work has demonstrated the potential of automated L2 speech assessment (e.g., *r* = .6-.7 for automated vs. human assessments in Table 1), they have exclusively focused on controlled speech where transcripts are available. The current investigation looks at the potential of automated assessment in spontaneous speech. Although L2 speech research has shown that listeners attend to segmental, melodic, and temporal information while judging the quality of foreign-accented speech, most of the existing studies have adopted either phonological, melodic, or fluency measures. Based on the methodological synthesis presented above (cf. Table 1), the current study adopted all the measures (phonological, melodic, *and* fluency) within the same model. As revied earlier, these measures included (a) speed and breakdown analyses for the fluency measures (e.g., Cucchiarini et al, 2000 for articulation rate and pause ratio); (b) variation analyses for the melodic measures (e.g., van Santen et al., 2009 for pitch and intensity variation); and (c) posterior-based analyses for the phonological measures (e.g., Shen et al., 2021 for posterior probabilities and gaps).

Research Questions

The following three research questions and predictions were formulated:

1. To what degree can automatic comprehensibility scoring simulate native listeners' intuitive judgements of various levels of comprehensibility (Study 1)?
2. Can automated fluency scoring predict different levels of comprehensibility when applied to new datasets (Study 2)?
3. Can automated fluency scoring predict different levels of comprehensibility when applied to new task contexts (Study 3)?

Predictions

As shown in the previous studies (e.g., *r* = .7-.8; Saito, 2021 for a meta-analysis), about 50-60% of the variance in L2 comprehensibility can be explained by a range of *manual* phonological measures (linguistically trained coders' analyses of segmental and melodic accuracy and temporal fluency). Thus, we expected to find relatively large correlation coefficients (*r* = .7-.8) between the *automated* phonological (maximum posterior probabilities, posterior gaps to natives), melodic (pitch and intensity variation), and fluency measures (articulation rate, pause ratio) and native listeners' L2 comprehensibility judgements (R1). It was also predicted that such findings can be replicated under new speaker (R2) and task conditions (R3).

Study 1: Model Training Phase

Speakers

As a part of the investigators' larger project, the team established a speech dataset of 1000+ Japanese learners with a wide range of proficiency and experience levels both in Japan (i.e., beginner-to-intermediate learners with relatively limited opportunities to use L2 in English-as-a-Foreign-Language settings) and in Canada (i.e., intermediate-to-advanced learners who use their L2 on a daily basis in English-as-a-Second-Language settings). Their speech was elicited using a range of speaking tasks. The dataset was derived from a series of L2 speech projects that the team has been working on over the past 10 years; parts of the cross-sectional and longitudinal analyses have been reported elsewhere (Saito, 2015a, 2015b for experienced Japanese speakers; Saito & Hanzawa, 2016 for inexperienced Japanese speakers).

Given the length of time required to collect comprehensibility judgements (e.g., more than one hour including explanation, training, and rating for 50 30-sec samples), to avoid listener fatigue, only samples of a single speaking task (picture description) from 100 speakers were used. Four subgroups of speakers were carefully selected to represent various levels of comprehensibility. While all of them started learning L2 English in Japan from Grade 7 (13-14 years of age), they differed substantially in terms of the length and timing of their immersion experience.

In conjunction with both cross-sectional and longitudinal evidence of a significant relationship between increased Length of Residence (LOR) and enhanced comprehensibility (e.g., Derwing & Munro, 2013; Saito, 2015a), participants' LOR was taken into consideration to create three subgroups (i.e., inexperienced, moderately experienced, and highly experienced Japanese speakers). Of course, LOR may not always reflect the amount of input that L2 learners have actually received. For example, the frequency of L2 use may widely vary among L2 speakers even if they stay in an L2-speaking country for the same period of time, with some choosing to use only L1 rather than L2 (for relevant discussion, see Flege & Bohn, 2021). To use LOR as an index of L2 experience for Moderately and Highly Experienced Japanese Speakers, efforts were made to recruit only those who reported English (rather than Japanese) as their main language of communication at work and/or home. As shown in our precursor projects (e.g., Saito, 2015), LOR served as a significant predictor of L2 comprehensibility. In essence, the current dataset comprised Japanese speakers with substantially different levels of comprehensibility.

.

- **Inexperienced Japanese Speakers of English** (*n* = 10): This group represents low-level L2 comprehensibility. All the participants were first-year university students in Tokyo (*M age* = 20.4 years; *Range* = 18-21 years). While all of them were registered for three hours of English classes a week at the time of the project, they reported little conversational use of the language outside the classroom. None of them had any immersion experience abroad.

- **Moderately Experienced Japanese Speakers of English** (*n* = 40): This group represents mid-level L2 comprehensibility. The participants were residents in Canada (*M age* = 34.7 years; *Range* = 22-48 years). Their lengths of immersion varied widely (*M* = 1.4 years; *Range* = 0.1 to 5 years). They were considered late bilinguals as they had moved to major cities in Canada (Vancouver, Montreal, and Calgary) after puberty (*M age* = 28.3 years; *Range* = 19-40 years).
- **Highly Experienced Japanese Speakers of English** (*n* = 40): This group represents high-level L2 comprehensibility. All of them were late bilinguals with their ages of arrival being after puberty (*M age* = 27.1 years; *Range* =21-36 years) and were long-term residents with at least six years of immersion experience in Canada (*M* = 11.3 years; *Range* = 6-18 years). In addition, they reported extensive and regular use of L2 English as their primary language of communication in various settings. Following the standard in SLA research (see DeKeyser, 2013), they were considered highly experienced attainers who had reached the upper range of L2 speech proficiency.
- **L1 Baseline** (*n* = 10): This group serves as a L1 speaker baseline. A total of 10 L1 speakers of English were recruited in Vancouver, Canada (*M age* = 27.5 years; *Range* = 18-37 years). They reported English as their L1 from birth onwards with both of their parents being L1 English speakers.

Speaking Task

All the participants completed a timed picture description task. Building on a similar task procedure (Munro & Mann, 2005), participants described seven different pictures with five seconds of planning time for each photo. Three key words were provided per photo in order to help low proficiency speakers to produce sufficient lengths of spontaneous speech without too much dysfluency due to insufficient vocabulary knowledge. The first four pictures were used as practice for participants to get used to the task procedure (describing a photo in English with minimal planning), and the last three pictures (A, B, and C) were used for the final analyses.

The pictures depicted (a) a table left out in driveway (key words: *rain, table* and *driveway*), (b) three men playing rock music with guitars (keywords: *three guys, guitar* and *rock music*), and (c) a long road under a cloudy blue sky (keywords: *blue sky*, *road*, and *cloud*). These key words were selected to elicit segmental, melodic, and syllabic structures that are especially difficult for Japanese learners of English. For example, Japanese speakers are likely to neutralize the English /r/-/l/ contrast (“rain, rock, brew, crowd” vs. “lane, lock, blue, cloud”) and substitute borrowed words (i.e., Katakana) by inserting epenthesis vowels between consecutive consonants (/dəraɪvə/ for “drive,” /θəri/ for “three,” /səkaɪ/ for “sky”) and after word-final consonants (/teɪbələ/ for “table,” /myuzɪkə/ for “music”).

All the samples were recorded at individual meetings which took place at a community center, a university lab, and participants’ residences (prior to the covid-19 pandemic) via digital Roland-05 audio recorders at 44.1 kHz sampling rate with 16-bit quantization.

Comprehensibility Judgements

Listeners. A total of 10 L1 listeners of General American English were recruited online (M_{age} = 24.8 years). They were all born in the USA and raised by monolingual parents. While all of them held a BA and/or MA degree, none of them majored in linguistics. Using a 6-point scale (*1 = not at all, 6 = very much*), they all reported that they were strongly familiar with foreign accented English speech (*M* = 5.8, *Range* = 4-6) but their familiarity with Japanese-accented English varied (*M* = 3.6, *Range* = 1-6). In line with Isaacs and Thomson’s (2020) categorization, these listeners can be considered naïve rather than expert. Following Munro and Derwing (1995), L2 comprehensibility was operationalized as naïve listeners’ intuitive judgements of spontaneous speech.

Procedure. Due to the pandemic, all of them completed the rating sessions individually with an investigator via a video-conferencing tool. All the rating sessions were conducted through the Gorilla platform for online research (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020). During the data collection, the investigator engaged with the participants in order to provide training, monitor their rating behaviors, respond to any enquiries, and help solve any technological problems. First, the listeners familiarized themselves with the listening materials (three picture descriptions by 100 L1 and L2 speakers) and the online rating procedure using Gorilla. Then, each listener received instruction on what characterized comprehensibility using the following definition:

- This term refers to how much effort it takes to understand what someone is saying. If you can understand (what the picture story is about) with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.

All the samples were played in a randomized order. Upon hearing each sample once, they made an intuitive judgement using a 9-point scale (*1 = difficult to understand, 9 = easy to understand*). No repeat button was provided.

First, the listeners practiced the rating procedure using three samples (not included in the main dataset). By listening to the speech dataset that the research team established (beyond the current study), the author carefully identified the three practice samples which represented the three subgroups of Japanese speakers (inexperienced, moderately experienced, highly experienced). After the practice session was completed, the listeners proceeded to judging the comprehensibility of the 100 samples. Each session took around 2 hours with a five-minute intermission a halfway through. In accordance with the recommended quality control measures for online L2 rating data collection (Nagle & Rehman, 2021), the platform was designed so that the listeners had to listen to the full-length of each sample (30 sec) before they rated the comprehensibility. The entire secession was carefully monitored by the investigator via a video-conferencing tool.

Automated Fluency Measures

Following Tavakoli and Skehan's (2005) model of fluency, and the automated fluency measures adopted in Cucchiarini et al (2000), the two major temporal dimensions of L2 speech were measured: speed (speech rate), and breakdown (pause ratio). The repair aspect of fluency (the ratio of self-correction and repetition) was not taken into consideration as recent literature has suggested it is a trait of first language fluency (Duran-Karaoz & Tavakoli, 2020) or/and cognitive individual differences (Zuniga & Simard, 2019).

Using the "To TextGrid (silences)" function in Praat (Boersma & Weenink, 2019), pauses were detected as silences longer than 250ms. The silence threshold (i.e., maximum intensity) was set to -20 dB as some speech samples included background noise. Next, the number of syllables was calculated based on de Jong and Wempe's (2009) Praat script. Nuclei were detected when the following phonetic conditions were met: (a) peak intensity was 2 dB above the median intensity; and (b) it was preceded and followed by 4+ dB of dips in intensity. The praat script did not automatically identify and remove any filled pauses and repetitions (both of which were included in phonation time). All the temporal information was used to calculate the following fluency measures:

(1) **Articulation rate**: This was calculated by dividing the number of syllables by the phonation time (sample duration minus all pauses).

(2) Pause ratio: This was calculated by dividing the number of unfilled pauses by sample duration.

Automated Phonological Measures

To capture the multilayered nature of comprehensibility, it is important to include not only speech features related to the quantity of phonation (fluency), but also those related to the phonological quality of pronunciation (accuracy). As reviewed earlier, a range of previous studies have adopted the posteriorgram-based analyses (DNN-HMM; Chen et al., 2018; Zechner et al., 2009). As preparation for the current investigation, the preliminary project was conducted, wherein a range of analysis methods for posteriorgram-based data were proposed, piloted, and refined with a view of optimal L2 speech assessments (Shen et al., 2021). The two posteriorgram-based analyses were adopted to assess phonological quality: (a) maximum posterior probabilities and (b) posterior gaps to natives. They were found to greatly boost the predictive power of the automatic assessment of perceived fluency (*r* = .8-.9 for machine vs. humans).

First, DNN models were trained with the Wall Street Journal corpus, which featured 37,416 utterances spoken by 123 L1 speakers of General American English and corresponding scripts. Since the current dataset (100 picture descriptions) included noticeable noise, it was necessary to adjust the sound clarity of the corpus. To this end, four levels of babble noises and machine noises (computer-synthesized distortions) we added to the Wall Street Journal corpus (signal-to-noise ratio = 10, 30, 40, and 50 [dB]). As a result, noise-robust English DNN-HMM acoustic models were trained based on the Wall Street Journal corpus (Povey et al., 2011). Once the models were trained for ASR, any input sample can be converted to its posteriogram.

The number of phonemes used in posteiorgrams is generally large (*n* = 2,000-3,000) (for further discussion on the concepts of posterior-based phonemes, see **Supporting information B**). As proposed in Kashiwagi, Zhang, Saito, and Minematsu (2016), the Bhattacharyya distance between two states was calculated directly from their state posterior probabilities through Bayes' theorem; the original dimension of the posteriorgrams (n = 2,000) was reduced to 50. From the posteriorgram of each utterance after pause removal, the two quality measures were calculated:

(3) **Averaged Maximum Posterior Probabilities:** When L1 posteriorgrams were visualized with context-independent phonemes, a posterior vector at each time appeared to be a one-hot vector. This means that the phoneme intended had a probability close to 1.0 and the others had almost 0.0. Here, from a given posteriorgram, the maximum posterior probability was calculated at each time, and then averaged over time. For example, at any point in time, there are always probability scores for 50 phoneme states. Subsequently, the maximum posterior probability score was identified out of the 50 scores, and its corresponding phoneme was assumed to indicate a speaker's intended phoneme. The higher the average was, the more distinct the pronunciation of the utterance was.

(4) **Averaged Posterior Gaps to Natives:** For each speaker, their averaged posterior vector was calculated. Since 10 native speakers were included in the main dataset, a gap from one learner to each native speaker was calculated to generate 10 gap scores in total. The distance was calculated via the Bhattacharyya metric. The averaged gap scores were thought to quantify the proximity of nativelikeness based on the distribution of perceived phonemes. Figure 1 visualizes the averaged posterior vectors and the posterior gaps. The former characterizes quality of pronunciation, represented as location in the feature space, and the latter characterizes distances to the 10 native speakers.

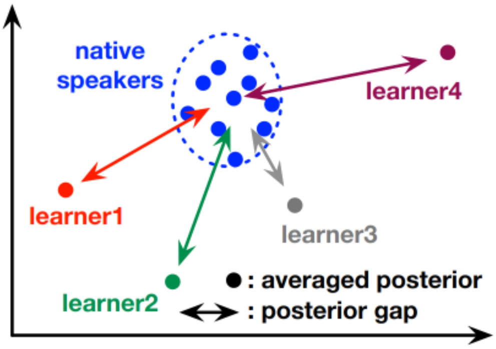

Figure 1

Conceptual Summary of Averaged Posterior and Posterior Gap

Automated Melodic Measures

As reviewed earlier, the previous literature has developed and adopted the automatic analyses of the varied use of pitch and amplitude in L2 speech (e.g., van Santen et al., 2009). In order to assess the melodic variation in participants' speech production, the software openSMILE ('open-source Speech & Music Interpretation by Large-space Extraction') was used (Eyben, Wöllmer, & Schuller, 2010). Given that listeners rely on both pitch and intensity contour to perceive English melody (e.g., Lieberman, 1960), two directly relevant measures were adopted in the current study, i.e., interquartile range values for F0 envelope and loudness:

(5) **Pitch variation**: After extracting fundamental frequencies (pitch frequencies) of given utterances and sorting the frequency values, the difference between 75th and 25th percentiles was calculated. The value indicated the magnitude of pitch variation or dynamics in the utterances. Monotonous speech (typical of beginner-to-intermediate L2 speakers) could be generally characterized as smaller pitch variations.

(6) **Intensity variation**: Similar to pitch variation, loudness was calculated as sequential data, which was defined as normalized intensity raised to a power of 0.3. After sorting the intensity values, the difference between 75th and 25th percentiles was calculated. The lack of intensity variation (and/or pitch variation) could help identify certain L2 English speakers who fail to distinguish between stressed and unstressed vowels.

Results

Listeners' Judgements of Comprehensibility

The first objective of the analyses was to examine how 10 L1 listeners perceived the comprehensibility of 90 Japanese speakers (Inexperienced, Moderately Experienced, and Highly Experienced) and 10 L1 speakers. As stated earlier, these groups were designed to represent various levels of comprehensibility among inexperienced/experienced L2 speakers and L1 speakers. In terms of inter-rater reliability, the listeners demonstrated medium-to-strong agreement ($r = .760$). The strength of the correlations widely varied from $r = .569$ between Listeners 3 and 7 to $r = .832$ between Listeners 6 and 8 (see Table 2). As in many L2 comprehensibility studies (e.g., Munro & Derwing, 1995), their rating scores were averaged across in order to derive one comprehensibility score for each speaker as an index of the listeners' overall ratings. According to the results of a Kolmogorov-Smirnov test, the averaged comprehensibility scores did not significantly differ from normal distribution (*D* = .086, *p* = .096). While all the participants in the L1 baseline group received 9 points, the L2 speakers' comprehensibility scores demonstrated a great deal of individual variation (*M* = 5.01, *SD* = 1.81, *Range* = 1.3-8.5).

Table 2

Inter-Class Correlations Among 10 Listeners' Comprehensibility Judgements (Study 1)

	Listener 2	Listener 3	Listener 4	Listener 5	Listener 6	Listener 7	Listener 8	Listener 9	Listener 10
Listener 1	.868	.694	.730	.712	.854	.767	.851	.755	.797
Listener 2		.696	.803	.758	.843	.778	.842	.795	.776
Listener 3			.614	.754	.716	**.569**	.658	.687	.764
Listener 4				.726	.804	.756	.802	.739	.752
Listener 5					.799	.686	.741	.797	.803
Listener 6						.784	**.832**	.748	.818
Listener 7							.784	.744	.701
Listener 8								.756	.807
Listener 9									.773

Automatic Assessments of Comprehensibility

The second objective of this study was to examine the extent to which L1 listeners' comprehensibility judgements can be tied to a set of automated phonological, melodic, and fluency measures (for descriptive statistics, see Table 3). The results of the normality test (Kolmogorov-Smirnov) demonstrated that whereas most of their fluency, phonological, and melodic scores were comparable to normal distribution ($p > .05$), their posterior gap scores were significantly different from normal distribution ($D = .202$, $p < .001$). Since a severe positive skewness was observed, the posterior gap scores were submitted to inverse transformation, resulting in the scores becoming normally distributed ($D = .080.$, $p = .515$). As such, the distance measure linearly represents the degree of proximity (i.e., nativelikeness). Given that both the dependent (comprehensibility ratings) and predictor variables (fluency, phonological, and melodic measures) followed normal distribution, only linear models were considered in the subsequent statistical analyses.

Table 3 *Descriptive Statistics of Automated Measures*

Measures	Interpretations	*M*	*SD*	95% *CI*	
				Low	Upper
A. Temporal quantity					
Articulation rate	syllables per second	3.618	0.510	3.2790	3.5343
Pause ratio	%	0.406	0.107	0.385	0.428
B. Phonological quality					
Maximum posterior probabilities	Probabilities between 0 (silence) and 1 (only one particular phoneme identified)	0.805	0.030	0.799	0.811
Posterior gaps to natives	Distance to native speaker data between 0 (no distance) and 1 (heavily foreign accented)	0.066	0.023	0.061	0.070
C. Melodic quality					
Pitch variability	Degree of variation (25th vs. 75th percentile) in fundamental frequencies in Hz	69.218	35.550	62.164	76.271
Intensity variability	Degree of variation (25th vs. 75th percentile) in normalized intensity between 0 (flat amplitude) and 1 (varied amplitude)	0.716	0.217	0.673	0.759

To examine how the automated assessments predicted different levels of comprehensibility (treated as ordinal values), a multiple regression model was constructed using the *lm* function in the R statistical environment (R Core Team, 2020). The model comprised listeners' averaged comprehensibility scores as dependent variables relative to the six automated measures (articulation rate, pause ratio, maximum posterior probabilities, inversed posterior gaps, pitch and intensity variability). As for model selection, "a full model" was chosen to maximize the predictive power of the model (with all six predictors entered in the equation). In the current analyses, we did not choose stepwise models (backward or forward selection based on the results of *F* tests) due to the following criticisms. First, the multiple comparisons conflated the occurrences of type 1 and 2 errors. Second, stepwise models are prone to overfitting the data and underestimating the degrees of freedom (for guidelines for the use of multiple regression analyses in applied linguistics research, see Larson-Hall, 2010).

As summarized in Table 4, the following model was tested: Listeners' averaged comprehensibility scores = Intercepts + articulation rate + pause ratio + maximum posterior probabilities + inversed posterior gaps + pitch + intensity variability. The full model significantly explained 67.7% of the variance in the listeners' comprehensibility judgements, $F(6, 93) = 32.507$, $p < .001$, without any clear evidence of multicollinearity problems (Variance Inflation Factor [VIF] = 1.016-1.456). In terms of the standardized β values, perceived comprehensibility was mainly predicted by segmental quality ($\beta = .607$ for posterior gaps, $\beta = .267$ for max posterior probabilities), secondarily by temporal quality ($\beta = -.193$ for pause ratio, $\beta = .123$ for articulation rate), and finally by melodic quality ($\beta = .113$ for intensity variation, $\beta = -.093$ for pitch variation).

The predicted comprehensibility scores were calculated based on the regression model's coefficients in Table 4. For each predictor variable, raw predictor values (i.e., articulation rate, pause ratio, maximum posterior probabilities, inversed posterior gaps, pitch variability, and intensity variability) were multiplied by unstandardized *B* (i.e., 0.524, -3.920, 19.427, 0.262, -0.006, and 1.137). Then, constant (i.e., -15.371) was added to total scores (i.e., the sum of all explanatory variables). Since one L1 speaking participant yielded 9.02, this was adjusted to the upper limit, 9.0.[3] The Pearson correlation coefficients between predicted and human

[3] There was one predicted score beyond 9 (9.02). This is because the predictive power of the model ($R^2 = .677$) was quite strong but not perfect. In addition, the model assumed a linear relationship between

comprehensibility scores were relatively strong, $r = .823$, $p < .001$. The figure here was higher than the 10 listeners' averaged inter-rater agreement ($r = .760$ among Listeners 1-10), and comparable to that of Listeners 6 and 8 who demonstrated the strongest agreement ($r = .832$ for Listeners 6 vs. 8). According to the results of an independent *t*-test, the predicted and human comprehensibility scores did not significantly differ, $t = .002$, $p = .998$, $d = 0.01$.

machine and human ratings without any breakpoints (i.e., linear rather than piecewise regression). Taking a close look at the participant in concern, he was a native speaker of English whose speech was characterized as substantially fast speech rate and low pause ratio.

Table 4

Results of Multiple Regression Analysis Using Automated Measures as Predictors of Listeners' Comprehensibility Scores

Predicted variables	Predictor variables	Unstandardized *B*	Standardized *β*	*F*	*p*	*VIF*
Comprehensibility scores	Constant	-15.371		-4.096	< .001	
	Articulation rate	0.524	.123	1.784	.078	1.359
	Pause ratio	-3.920	-.193	-2.967	.004	1.213
	Max posterior probabilities	19.427	.267	4.493	< .001	1.016
	Inversed posterior gaps	0.262	.607	8.535	< .001	1.456
	Pitch variation	-0.006	-.093	-1.324	.189	1.417
	Intensity variation	1.137	.113	1.656	.101	1.345

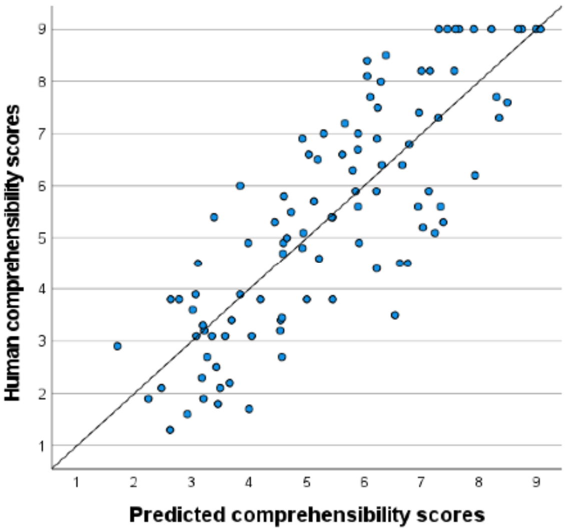

Figure 2

The Relationship Between Human Comprehensibility Scores and Predicted Comprehensibility Scores (*r* = .823)

Study 2: Model Validation Phase

Overall, Study 1 demonstrated (a) that the machine scoring can successfully simulate human listeners' comprehensibility judgements (*r* = .823 for machine vs. humans; *r* = .760 for humans vs. humans); and (b) that such automated comprehensibility assessments can be primarily determined by phonological accuracy (*β* = .607 for inversed posterior gaps; *β* = .267 for max posterior probabilities) and secondarily by fluency features (*β* = -.193 for pause frequency; *β* = .123 for articulation rate). The results here are in line with the existing literature (the relation weights of phonological accuracy over fluency in L2 comprehensibility; see Isaacs & Trofimovich, 2012). Interestingly, none of the melodic measures (pitch and intensity variation) turned out to be significant predictors. This could be due to several reasons. First, the timed picture description task used in the current study may not have elicited sufficiently long sentences, limiting speakers from demonstrating varied and accurate use of melodic cues. Secondly, the relationship between pitch height measures (including pitch range) and L2 comprehensibility may be minimal (*β* = -.10 in Kang, Rubin, & Pickering, 2010). Thirdly, listeners orchestrate a range of different melodic cues (beyond pitch and intensity) to perceive, identify, and encode the targetlikeness of lexical and sentence stress in English (Lieberman, 1960). Study 2 was designed to test the extent to which the automated assessments can predict different levels of L2 comprehensibility when applying the regression formula in Table 5 to a new L2 speech dataset (*N* = 45 timed picture descriptions).

Method

Speakers

A total of 40 Japanese learners of English were recruited from Japan and Canada (17 males, 23 females). 30 of them were university-level students in Japan (M_{age} = 19.8 years; *Range* = 18-26 years). The results of their general English proficiency test scores (Test of English for International Communication) suggested that their L2 oral proficiency varied widely (*M* = 681.5 out of 900; *Range* = 300-980), covering basic to proficient users. The other 10 participants were recruited in Vancouver and Ontario (M_{age} = 41.4 years; *Range* = 35-47 years). All of them were late L2 speakers ($M_{\text{age of arrival}}$ = 27.3 years; *Range* = 21-35 years) who used English regularly and had a great deal of immersion experience in Canada ($M_{length\ of\ immersion}$ = 13.1 years; *Range* = 7-24 years). All of them could be considered highly functional and regular L2 users as they reported English rather than Japanese as their primary language of communication in home and/or work settings. Finally, five L1 speakers of Canadian English were included. The biographical profiles of the participants in Study 2 can be considered comparable to those in Study 1. They were believed to represent different levels of comprehensibility.

Speaking Task

The participants' spontaneous speech in English was elicited via the same task used in Study 1 (timed picture description). The first 10 seconds of three picture descriptions were stored in a single WAV file per participant and used for the comprehensibility judgements.

Comprehensibility Judgements

A total of five L1 English speakers were recruited in London, England (M_{age} = 20.6 years). Using the same rating procedure in Study 1, they participated in individual rating sessions with a researcher using a video-conferencing tool. First, they received a brief summary of project, and explanations of the nature of the dataset, the definition of comprehensibility, and the rating procedure. Next, in order to familiarize themselves with the procedure, they rated the same practice samples used in Study 1, explained their decisions, and received feedback from the researcher. Finally, they proceeded to make comprehensibility judgements of the main dataset (40 minutes without any intermission).

Automated Speech Analyses

The speech data was submitted to the same automated fluency (articulation rate, pause frequency), phonological (maximum posterior probabilities, posterior gaps to natives), and melodic analyses (pitch and intensity variability) as Study 1.

Results

Similar to Study 1, the averaged inter-rater agreement among the five listeners was relatively high, $r = .789$ ranging from $r = .718$ (Listeners 1 vs. 2) to .871 (Listeners 2 vs. 3; summarized in Table 6). Their comprehensibility scores were averaged per speaker to represent the human listeners' overall comprehensibility judgements ($M = 3.67$; $SD = 2.26$; *Range* = 1-9). According to the Kolmogorov-Smirnov test, the resulting scores did not significantly differ from normal distribution, $D = .160$, $p = .179$. To check the degree to which the comprehensibility scores in Study 2 ($n = 45$) differed from those in Study 1 ($n = 100$), a non-parametric test (Mann-Whitney) was conducted. The results showed that the participants in Study 2 ($M = 3.67$) received lower comprehensibility scores than those in Study 1 did ($M = 5.40$), $Z = -4.429$, $p < .001$. Although efforts were made, the two datasets were not comparable. It is noteworthy, however, that the listeners were different between Studies 1 and 2 ($n = 10$ American listeners vs. 5 British listeners). Thus, we consider this to be a tentative pattern.

Using the regression formula in Table 5, the subsequent analyses set out to examine the extent to which the six automated scores (summarized in **Supporting Information-C**) can predict human comprehensibility scores. To this end, the raw maximum posterior gap scores were inversed. All the temporal, phonological, and melodic scores followed normal distribution ($D = .082\text{-}.143$, $p = .281\text{-}.896$). The predicted comprehensibility scores varied between 1.82 and 7.17 ($M = 4.13$, $SD = 1.23$).

As visually summarized in Figure 3, the correlation between human and predicted comprehensibility scores was relatively strong, $r = .827$ ($p < .001$) and can be compared to the averaged correlation coefficients among the five listeners ($r = .789$). An independent t-test did not detect a significant difference between the human listeners' comprehensibility scores and listeners' averaged comprehensibility scores, $t = -1.186$, $p = .239$, $d = 0.252$. To sum, the results indicated (a) that the automated measures successfully simulated the human listeners' comprehensibility scores; and (b) that the difference between the predicted and human listener scores was minimal.

Table 6

Inter-Class Correlations Among Five Listeners' Comprehensibility Judgements (Study 2)

	Listener 2	Listener 3	Listener 4	Listener 5
Listener 1	**.718**	.761	.769	.759
Listener 2		.850	.845	.830
Listener 3			**.871**	.745
Listener 4				.743

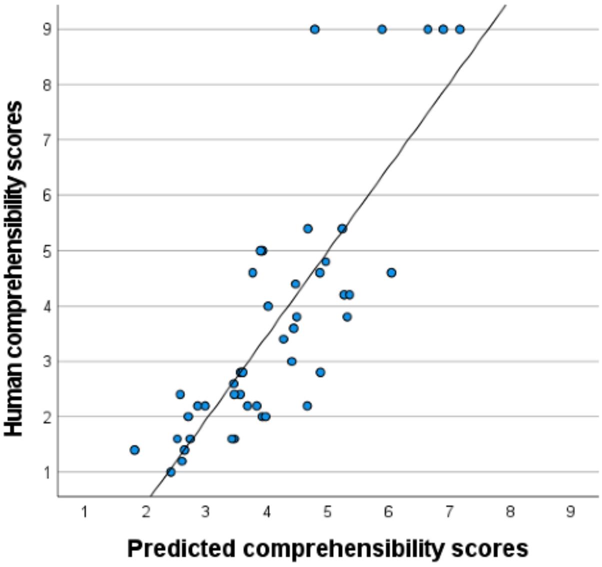

Figure 3

The Relationship Between Human Comprehensibility Scores and Predicted Comprehensibility Scores (*r* = .827)

Study 3: Generalization Study

Study 2 provided some empirical support for the predictive power of the regression model for human listeners' comprehensibility scores. Notably, the findings have thus far been based on a relatively structured, semi-spontaneous speech task (timed picture description) wherein the linguistic content of sample is likely to be highly predictable (e.g., all participants used the same word prompts to describe the same pictures). In Study 3, we aimed to examine the extent to which the regression model can be applied when speech is elicited using a more extemporaneous, free-constructed task.

Speakers and Listeners

The same speakers and listeners from Study 2 (40 Japanese learners of English in Japan and Canada; 5 L1 speakers of English; 5 L1 speaking listeners) participated in Study 3.

Speaking Task

The participants engaged in an oral interview task which was created by the research team based on the procedure in the IELTS interview task (Crowther, Trofimovich, Isaacs, & Saito, 2015). Given that the task format gave participants a certain level of freedom over the content and organization of their speech (rather than describing pictures), participants were encouraged to produce extemporaneous speech with their primary focus on conceptualization (what to say) rather than accuracy (how to say it). In accordance with Skehan's (1998) task taxonomy, the oral interview task can be considered less structured, and more informal and personal. In such tasks, speakers likely use more contextually rich, varied, and idiosyncratic language. This sharply contrasts with the timed picture description task, wherein participants are likely focused on accurately and fluently using language in order to describe the provided content in the most effective and efficient manner.

First, the participants were given a familiar and personal topic to discuss (i.e., *What was the hardest and toughest change in your life?*). To help speakers produce long and meaningful speech, a set of possible discussion points (e.g., *Why was it so challenging?*) were also provided on a topic card. After the participants spent one minute on planning, they spoke for approximately two minutes. Finally, the researcher asked one or two follow-up questions in response to the content of their speech (e.g., *What did you learn from the experience?*). For the task prompts, see **Supporting Information-D**. The first 30 seconds of each sample were carefully cut and saved in a single WAV file.

Comprehensibility Judgements

After the five listeners completed the comprehensibility judgements of the 45 picture description samples in Study 2, they assessed the comprehensibility of 45 interview samples on a 9-point scale using the same procedure as in Studies 1 and 2.

Automated Speech Analyses

The speech data was analyzed via automated fluency, phonological, and melodic measures in Studies 1 and 2.

Results

Similar to Study 1 (r = .760) and Study 2 (r = .789), the averaged inter-rater agreement among the five listeners in Study 3 was relatively high, r = .756. As summarized in Table 7, the strength of the agreement between certain individual listeners varied from "mid" (r = .655 for Listeners 4 vs. 5) to "strong" (r = .869 for Listeners 2 vs. 3). To obtain the listeners' consensus, their comprehensibility scores were averaged per speaker (M = 4.47, SD = 2.05, *Range* = 1-9). The results of the normality test (Kolmogorov-Smirnov) found the resulting comprehensibility scores to be indistinguishable from normal distribution, D = .103, p = .685. In order to calculate the predicted comprehensibility scores, the regression formula in Table 5 was used. Following Studies 1 and 2, the raw maximum posterior gap scores were inversely transformed. The normality test (Kolmogorov-Smirnov) demonstrated that all the automated scores (articulation rate, pause frequency, maximum posterior probabilities and gaps, pitch and intensity variability) followed normal distribution (D = .052-.117, p = .527-.992). The predicted comprehensibility scores varied between 2.12 and 8.85 (M = 4.33, SD = 1.73).

As visually displayed in Figure 4, the predicted comprehensibility scores were strongly associated with the human listeners' comprehensibility scores, r = .809 (p < .001). The correlation coefficients here can be considered similar to those among the five listeners (r = .756). According to the results of an independent t-test, the difference between human and predicted comprehensibility scores did not reach statistical significance, t = -0.360, p = .719, d = 0.07. In a nutshell, the findings here indicated that the automated measures can greatly simulate human judgements of different levels of comprehensibility not only when speech is elicited via a relatively structured task (picture description), but also when the regression model is applied to free, extemporaneous speech (oral interview).

Table 7

Inter-Class Correlations Among Five Listeners' Comprehensibility Judgements (Study 2)

	Listener 2	Listener 3	Listener 4	Listener 5
Listener 1	.719	.707	.695	.726
Listener 2		**.869**	.752	.858
Listener 3			.741	.844
Listener 4				**.655**

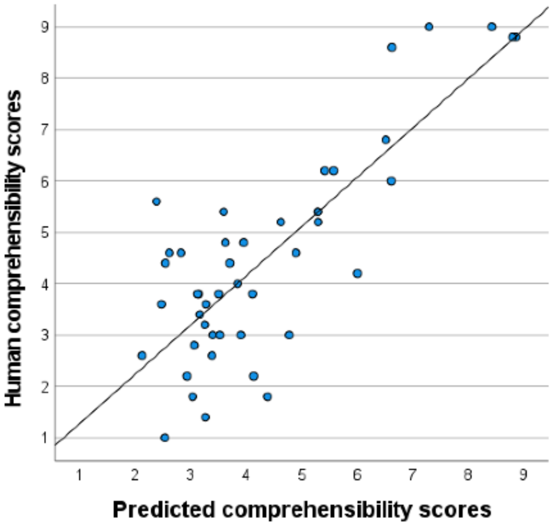

Figure 4

Relationship Between Human Comprehensibility Scores and Predicted Comprehensibility Scores (*r* = .809)

Discussion

Based on a speech corpus of 90 Japanese speakers of English in Japan and North America, and 10 L1 speakers of English, Study 1 showed the extent to which L1 listeners' comprehensibility ratings can be simulated by a set of automated fluency (articulation rate, pause ratio), phonological (posterior probabilities and gaps), and melodic (pitch and intensity variation) measures. The composite model's predicted comprehensibility scores were significantly predictive of human assessments of comprehensibility. The strength of the correlation coefficients (*r* = .823 for machine vs. humans) can be considered comparable to the inter-rater agreement among 10 naïve listeners (*r* = .760 for humans vs. humans). In Studies 2 and 3, we further confirmed the validity and generalizability of the model as it strongly predicted L2 comprehensibility among a *different* group of Japanese speakers' L2 English speech which was elicited not only from the same picture description task (*r* = .827 for machine vs. humans) but also via a more freely constructed oral interview task (*r* = .809 for machine vs. humans).

First and foremost, the findings supported the emerging evidence for the possibility, reliability, and validity of machine assessments of spontaneous speech samples (e.g., *r* = .799 in Fu et al., 2019; *r* = .77 in Chen et al., 2018). Furthermore, our study indicated that such automatic scoring can be used to simulate naïve listeners' *intuitive* judgements of L2 speech comprehensibility, an anchor of communicative success among English speakers in global contexts. We would like to stress the importance of using the recent models of fluency-based measures and DNN-based instruments complementarily. Since the fluency, phonological, and melodic measures tap into different dimensions of L2 speech without notable multicollinearity (VIF < 1.5), they make separate contributions for the model to capture the multifaceted nature of L2 comprehensibility.

As shown in the previous studies (e.g., Isaacs & Trofimovich, 2012), it is important to note that listeners' comprehensibility scores were uniquely related to various areas of language with relatively different weights (the order of importance: phonetic accuracy > fluency > melodic variation). This in turn indicates that comprehensibility is a complex construct as listeners attend to all the available linguistic information in foreign accented speech in order to decipher as much meaning as possible. This supports the importance of including conceptually and methodologically *diverse* measures (phonological, fluency, melody). It is possible that the relatively strong predictive power of the current findings (r = .823) could be ascribed to the fact that we adopted an integrative approach (including fluency, phonological, and melodic measures) rather than single models typically found in previous studies (e.g., Fu et al., 2019 for phonological measures; Ginther et al., 2010 for fluency measures).

Given that the current study took a first step towards examining the potential of automated assessments of L2 comprehensibility using fluency, phonological, and melodic measures in a complementary fashion, a set of promising directions need to be addressed with a view towards future extension studies. the findings of the current study (and those of the existing research in Table 1) exclusively focused on the correlation analyses with an assumption that the relationship between human and machine scoring could be linear. Future studies should further examine the nature of the relationship (e.g., linear vs. quadratic).

Secondly, although the current study exclusively focused on Japanese speakers of English, it would be intriguing to replicate the findings in the context of different L1-L2 pairings. For example, Crowther et al. (2015) showed that L1 listeners were found to attend to different linguistic dimensions while assessing the comprehensibility of L2 English speech among L1 Chinese speakers (segmentals), L1 Hindi speakers (intonation), and L1 Farsi speakers (segmentals, word stress, fluency). Saito and Akiyama (2017) demonstrated that L1 listeners likely prioritized melodic rather than segmental information during the comprehensibility judgments of L2 Japanese speech. Given that the linguistic weights of perceived comprehensibility vary across different L1 and L2 contexts, further research is necessary to see the extent to which the combination of the automated fluency, phonological, and melodic measures featured in the current study (articulation rate, pause ratio, posterior probabilities/gaps, pitch/intensity variation) can be applied to other large-scale datasets.

Third, although the current study included temporal, phonological, and melodic measures within the same model, there is a growing amount of research showing that naïve listeners take into account not only pronunciation, but also lexicogrammatical information while making L2 comprehensibility judgements (Isaacs & Trofimovich, 2012). In fact, previous work has demonstrated that listeners primarily rely on phonological and temporal information (40-50% of the variances) but secondarily on lexicogrammar (10-20% of the variances; Crowther et al., 2015). There is a growing amount of evidence for the relative importance of certain vocabulary features such as lexical appropriateness rather than richness (Isaacs & Trofimovich, 2012) and collocational association rather than frequency (Saito, 2020). One promising direction is to add automated word recognition measures in order to further increase the predictive power of the model that we have developed for the current study (see Kyle & Crossley, 2015 for their discussion on the automated assessments of L2 spoken vocabulary).

Fourth, although efforts were made to include a range of fluency, melodic, and phonological measures in the model, there are a range of other key automated fluency, melodic and phonological measures which future studies should highlight in order to further improve the predictive power of the automated L2 speech assessment. For example, Fu et al. (2019) conducted posteriorgram-based analyses of participants' L1 and L2 speech data. Unlike the current study (using only participants' L2 speech data), future work should compare the distance between the poteriorgram distribution of participants' L1 and L2 speech data. On a related note, the results of the multiple regression analyses in the current study found neither of the melodic *variation* measures (i.e., pitch and intensity variation) to make a significant contribution to the predicted compressibility scores. Given that there is ample research evidence showing that melodic *accuracy* is strongly associated with L2 comprehensibility (e.g., Isaacs & Trofimovich, 2012), future studies should include both melodic variation and accuracy measures. As we acknowledged in the Literature Review section, however, it is important to remember (a) that the existing literature (including the current investigations) has exclusively focused on the *variation* of melody; and (b) that few studies have probed the automatic analyses of the *accurate* use of melody especially in spontaneous speech (where transcripts and model speech are unavailable). Thus, we may need to wait for more future research to develop, validate, and refine the automated analyses of how L2 speakers use various types of melodic cues while marking word and sentence stress in a contextually appropriate manner.

Fifth, we acknowledge that the findings in the current investigation were based on a series of cross-sectional analyses. As Knoch and Chapelle (2017) argued, assessment validation entails much more than correlational examinations. The question of whether machine assessments of spontaneous speech will be suitable for making fine-grained decisions may still be at hand. One promising direction for future replication studies concerns a longitudinal investigation of the development of comprehensibility among L2 speakers over time. There is evidence showing that L2 speech learning takes place on a continuum of comprehensibility in relation to increased immersion experience (e.g., Derwing & Munro, 2013). Thus, researchers can track how L2 leaners' speech behaviors change for a certain period of time (1-5 years of immersion), how listeners can perceive such changes from the point of comprehensibility, and how automated measures can replicate the developmental trajectories of comprehensibility among the same individuals.

Sixth, a reviewer pointed out the significance of the predicted scores calculated by the automated measures. As seen in Figure 3, for example, although native English speakers received the highest comprehensibility scores (i.e., 9), the predicted comprehensibility scores varied to a great extent (as low as 5). Here, we would like to remind the readers that the correlations between the human and automated assessments of L2 comprehensibility were considered relatively strong but not perfect (r = .823, .827, and .809 in Studies 1, 2, and 3). In addition, it is important to remember that the inter-rater correlations among the human raters was considered medium-to-strong (r = .7-.8). This in turn suggests that it is difficult to obtain an *absolute* L2 comprehensibility rating for a particular L2 speech sample because it triggers different types of reactions (thereby different L2 comprehensibility scores) not only between machine and human raters, but also within human raters.

Finally, many EFL students suffer from the lack of speech training opportunities, wherein they can receive individualized feedback on their speaking skills especially regarding comprehensibility (rather than nativelikeness). The current studies shed light on the pedagogical potential of automated assessment of the ecologically valid metric of comprehensibility. Such automated comprehensibility judgements can be used to guide individual students to develop their L2 comprehensibility in accordance with their own goals relative to their current proficiency levels. This will in turn allow learners to engage in more tailored, optimal, and autonomous L2 speech learning in the long run. For example, the descriptive statistics showed that Japanese speakers' comprehensibility levels ranged from 1.3 to 8.5 (relative to native speakers who scored 9 out of 9). Using the values as a rough benchmark for different levels of comprehensibility, inexperienced Japanese speakers can be encouraged to aim at moderately comprehensible speech which corresponds to moderate L2 English proficiency (5-6 out of 9). Similarly, Japanese participants whose speech is already somewhat intelligible/comprehensible can aim to achieve more advanced, adequately comprehensible speech, which corresponds to highly advanced L2 English oral proficiency (7-8 out of 9). Importantly, students should be informed (a) that many L2 learners should aim for mid-to-high L2 comprehensibility to become functional, competent L2 users (6-8 out of 9); and (b) that comprehensibility and accentedness are separate constructs, and even highly accented speech (e.g., 9 out of 9) can be adequately comprehensible. By adopting both quantitative and qualitative analyses, future studies are strongly recommended to further pursue how the use of automated assessment can enrich L2 speech training and then facilitate the different levels of L2 comprehensibility.

Conclusion

Given that eliciting listeners' L2 speech assessment is a time-consuming task, some previous research has begun to examine the automated assessment of linguistically trained raters' general speaking proficiency judgments. However, the analyses of these studies have been restricted to controlled speech samples and/or a limited set of speech measures (phonological fluency or accuracy). Using a speech data of L1 and L2 speakers of English, the current study took a first step towards combining *three* different automated analyses (fluency, phonological, and melodic) in order to simulate how *naïve* listeners *intuitively* perceive the comprehensibility of *spontaneous* speech samples across different task conditions. Findings showed that the composite model can provide predicted comprehensibility scores (r = .809-.827 for machine vs. humans) which are comparable to what different naïve listeners likely agree on (r = .756-.789 for humans vs. humans). The current study supports the use of up-to-date automated assessment of L2 oral proficiency on the continuum of comprehensibility which many researchers have suggested as an ecologically-valid goal for adult L2 speakers and students (comprehensible rather than nativelike).

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388-407. https://doi.org/10.3758/s13428-019-01237-x

Boersma, P., & Weenink, D. (2019). *Praat: doing phonetics by computer [Computer Software]*. Retrieved from http://www.praat.org/

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, *19*(4), 447-465. https://doi.org/10.1017/S0272263197004026

Chen, Z., Qian, Y., & Yu, K. (2018). Sequence discriminative training for deep learning based acoustic keyword spotting. *Speech Communication*, *102*, 100-111. https://doi.org/10.1016/j.specom.2018.08.001

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal*, *99*, 80-95. https://doi.org/10.1111/modl.12185

Cucchiarini, C., Strik, H., & Boves, L. (2000). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, *30*(2-3), 109-119. https://doi.org/10.1016/S0167-6393(99)00040-0

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*(6), 2862-2873. https://doi.org/10.1121/1.1471894

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

DeKeyser, R. M. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning*, *63*(s1), 52-67. https://doi.org/10.1111/j.1467-9922.2012.00737.x

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

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-Year Study. *Language Learning*, *63*(2), 163-185. https://doi.org/10.1111/lang.12000

Duran-Karaoz, Z., & Tavakoli, P. (2020). Predicting L2 fluency from L1 fluency behavior: The case of L1 Turkish and L2 English speakers. *Studies in Second Language Acquisition*, *42*(4), 671-695. https://doi.org/10.1017/S0272263119000755

Eyben, F., Wllmer, M., & Schuller, B. (2010). Opensmile: The Munich versatile and fast open-source audio feature extractor. *MM '10: Proceedings of the 18th ACM international conference on Multimedia*, 1459-1462. https://doi.org/10.1145/1873951.1874246

Flege, J. E., & Bohn, O. S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3-83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002

Foote, J. A., Holtby, A. K., & Derwing, T. M. (2012). Survey of the teaching of pronunciation in adult ESL programs in canada, 2010. *TESL Canada Journal*, *29*(1), 1-22. https://doi.org/10.18806/tesl.v29i1.1086

Fu, J., Chiba, Y., Nose, T., & Ito, A. (2020). Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Communication*, *116*, 86-97. https://doi.org/10.1016/j.specom.2019.12.002

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *37*(3), 379-399. https://doi.org/10.1177%2F0265532210364407

Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., & Reiterer, S. M. (2013). Language aptitude for pronunciation in advanced second language (L2) learners: Behavioural predictors and neural substrates. *Brain and Language*, *127*(3), 366-376. https://doi.org/10.1016/j.bandl.2012.11.006

Isaacs, T., & Thomson, R. I. (2020). Reactions to second language speech: Influences of discrete speech characteristics, rater experience, and speaker first language background. *Second Language Pronunciation*, *6*(3), 402-429. https://doi.org/10.1075/jslp.20018.isa

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

Isaacs, T., Trofimovich, P., & Foote, J. A. (2017). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*(2), 193-216. https://doi.org/10.1177%2F0265532217703433

Kachlicka, M., Saito, K., & Tierney, A. (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language*, *192*, 15-24. https://doi.org/10.1016/j.bandl.2019.02.004

Kang, O., & Johnson, D. (2018). The roles of suprasegmental features in predicting English oral proficiency with an automated system. *Language Assessment Quarterly*, *15*(2), 150-168. https://doi.org/10.1080/15434303.2018.1451531

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

Kashiwagi, Y., Zhang, C., Saito, D., & Minematsu, N. (2016). Divergence estimation based on deep neural networks and its use for language identification. *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/ICASSP.2016.7472716

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 Speech: The role of listener experience and semantic context. *The Canadian Modern Language Review*, *63*(3), 459-489. https://doi.org/10.3138/cmlr.64.3.459

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757-786. https://doi.org/10.1002/tesq.194

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *The Journal of the Acoustical Society of America*, *32*(4), 451-454. https://doi.org/10.1121/1.1908095

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo Case Study. *ReCALL*, *31*(3), 293-311. https://doi.org/10.1017/S0958344019000065

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA. *Studies in Second Language Acquisition*, *32*(2), 265-302. https://doi.org/10.1017/S0272263109990520

Mackey, A. (2012). *Input, interaction, and corrective feedback in L2 learning*. Oxford, UK: Oxford University Press.

Moustroufas, N., & Digalakis, V. (2007). automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech & Language*, *21*(1), 219-230. https://doi.org/10.1016/j.csl.2006.04.001

Muñoz, C. (2014). Exploring young learners' foreign language learning awareness. *Language Awareness*, *23*(1-2), 24-40. https://doi.org/10.1080/09658416.2013.863900

Munro, M., & Mann, V. (2005). Age of immersion as a predictor of foreign accent. *Applied Psycholinguistics*, *26*(3), 311-341. https://doi.org/10.1017/S0142716405050198

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/j.1467-1770.1995.tb00963.x

Nagle, C. (2018). Motivation, comprehensibility, and accentedness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development. *The Modern Language Journal*, *102*(1), 199-217. https://doi.org/10.1111/modl.12461

Nagle, C. L., & Rehman, I. (2021). Doing L2 speech research online: Why and how to collect online ratings data. *Studies in Second Language Acquisition*, *Advance Online Publication*, 1-24. https://doi.org/10.1017/S0272263121000292

Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication*, *30*(2-3), 83-93. https://doi.org/10.1016/S0167-6393(99)00046-1

Nishino, T., & Watanabe, M. (2008). Communication-oriented policies versus classroom realities in Japan. *TESOL Quarterly*, *42*(1), 133-138. https://doi.org/10.1002/j.1545-7249.2008.tb00214.x

O'Brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I., Strik, H., Levis, J. M., Munro, M. J., Foote, J. A., & Levis, G. M. (2018). Directions for the future of technology in pronunciation research and teaching. *Second Language Pronunciation*, *4*(2), 182-207. https://doi.org/10.1075/jslp.17001.obr

Pennington, M. C. (2021). Teaching pronunciation: The state of the art 2021. *RELC Journal*, *52*(1), 3-21. https://doi.org/10.1177%2F00336882211002283

Pennycook, A. (2017). Translanguaging and semiotic assemblages. *International Journal of Multilingualism*, *14*(3), 269-282. https://doi.org/10.1080/14790718.2017.1315810

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The kaldi speech recognition toolkit. *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, US*. Retrieved from https://infoscience.epfl.ch/record/192584

Saito, K. (2015a). Experience effects on the development of late second language learners' oral proficiency. *Language Learning*, *65*(3), 563-595. https://doi.org/10.1111/lang.12120

Saito, K. (2015b). The role of age of acquisition in late second language oral proficiency attainment. *Studies in Second Language Acquisition*, *37*(4), 713-743. https://doi.org/10.1017/S0272263115000248

Saito, K. (2020). Multi- or single-word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, *70*(2), 548-588. https://doi.org/10.1111/lang.12387

Saito, K. (2021). What Characterizes Comprehensible and Native-like Pronunciation Among English-as-a-Second-Language Speakers? Meta-Analyses of Phonological, Rater, and Instructional Factors. *TESOL Quarterly*. https://doi.org/10.1002/tesq.3027

Saito, K., & Akiyama, Y. (2017). Linguistic correlates of comprehensibility in second language Japanese speech. *Journal of Second Language Pronunciation*, *3*(2), 199-217. https://doi.org/10.1075/jslp.3.2.02sai

Saito, K., & Hanzawa, K. (2016). Developing second language oral ability in foreign language classrooms: The role of the length and focus of instruction and individual differences. *Applied Psycholinguistics*, *37*(4), 813-840. https://doi.org/10.1017/S0142716415000259

Saito, K., Tran, M., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech? *Studies in Second Language Acquisition*, *41*(5), 1133-1149. https://doi.org/10.1017/S0272263119000226

Saito, K., Trofimovich, P., & Isaacs, T. (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

Scales, J., Wennerstrom, A., Richard, D., & Wu, S. H. (2012). Language learners' perceptions of accent. *TESOL Quarterly*, *40*(4), 715-738. https://doi.org/10.2307/40264305

Shen, Y., Yasukagawa, A., Saito, D., Minematsu, N., & Saito, K. (2021, January). Optimized prediction of fluency of L2 English based on interpretable network using quantity of phonation and quality of pronunciation. In *2021 IEEE Spoken Language Technology Workshop (SLT)* (pp. 698-704). IEEE.

Skehan, P. (1998). Task-based instruction. *Annual Review of Applied Linguistics*, *18*, 268-286. https://doi.org/10.1017/S0267190500003585

Sung, C. C. M. (2016). Does accent matter? Investigating the relationship between accent and identity in English as a lingua franca communication. *System*, *60*, 55-65.

Suzuki, S., & Kormos, J. (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

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*, *25*(3), 431-450. https://doi.org/10.1177%2F1362168819858246

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp. 239-273). https://doi.org/10.1075/lllt.11.15tav

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition*, *19*(1), 122-140. https://doi.org/10.1017/S1366728914000832

van Santen, J. P.H., Prud'hommeaux, E. T., & Black, L. M. (2009). Automated assessment of prosody production. *Speech Communication*, *51*(11), 1082-1097. https://doi.org/10.1016/j.specom.2009.04.007

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, *30*(2-3), 95-108. https://doi.org/10.1016/S0167-6393(99)00044-8

Zechner, K., & Evanini, K. (Eds.). (2020). *Automated speaking assessment: Using language technologies to score spontaneous speech*. New York: Routledge.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883-895. https://doi.org/10.1016/j.specom.2009.04.009

Zuniga, M., & Simard, D. (2019). Factors influencing L2 self-repair behavior: The role of L2 proficiency, attentional control and L1 self-repair behavior. *Journal of Psycholinguistic Research*, *48*, 43-59. https://doi.org/10.1007/s10936-018-9587-2

Supporting Information-A: Hidden Markov Models (HMM), Gaussian Mixture Models (GMM) and Deep Neural Networks (DNN)

HMM is a state transition model, trained separately for each class, and each of the states in an HMM generates acoustic signals o_t (t is time), whose features exclusively depend on the state. In other words, each state has its own output probability $P(o|c_i)$ and each HMM is trained with training data so that the trained HMM generates the training data with a higher probability. Here, it is usually assumed that $P(o|c)$ follows a Gaussian distribution (i.e., bell curve shape) or a mixture of Gaussian distributions. For ASR, with a given acoustic observation o, posterior probability $P(c|o)$ has to be calculated. For this calculation, $P(o|c)$ is referred to because $P(c|o)$ is proportional to $P(o|c)P(c)$.

On the other hand, DNN is trained so that it can output $P(c|o)$ directly, where DNN functions as a speech classifier and it shows how probably input observation o belongs to class c. Here, no statistical assumption, such as Gaussian assumption, is needed. With DNN, each state of the HMM can have its own DNN classifier, and GMM-HMM has been overcome by DNN-HMM with respect to their performances. By using the DNN acoustic model, any speech input o_t can be converted to its set of class posteriors $\{P(c_i|o_t)\}$, $1 \leq i \leq I$, where I is the total number of the speech classes. By using $\{P(c_i|o_t)\}$ to represent o_t, speech sequence $\{o_t\}$ is converted to a temporal sequence of probability distributions or probability vectors. This probabilistic representation is often called as phone posteriograms.

Supporting Information-B: Calculation of Posterior-Based Phonemes

Notably, the number of phonemes used in posteriorgrams is not equivalent to the number of linguistically-driven phonemes. In automated speech recognition (ASR), a phoneme is often divided into three states (i.e., beginning, intermediate and ending states). In addition, a phoneme is often defined as per surrounding phonemic contexts. For instance, phonemes /x/ can be treated as different categories when they are found in various contexts (e.g., /axb/ vs. /cxd/). If the number of the linguistically-driven phonemes is N, then, the number of the context-dependent phonemes is N^3. Furthermore, since a phoneme is divided into three states, the number of context-dependent phoneme states is $3*N^3$. If $N = 50$, $3*N^3 = 375{,}000$, which is the logically-driven number of the phoneme states, and is the logically-driven dimension of a posterior probability vector in the posteriorgram. However, this number is too huge, and by clustering the phoneme states down into a few thousands, the posteriogram with its dimension being 2,000 to 3,000 are widely used in the current ASR. When applying the posteriorgram to L2 speech assessment, the dimension can be further reduced by bottom-up clustering with Ward's method, where the distance matrix between any pair of the states is used (see Kashiwagi et al., 2016).

Supporting Information-C: Descriptive Statistics of Automated Measures

Descriptive Statistics of Automated Measures: Study 2

	M	*SD*	95% *CI*	
			Low	Upper
A. Temporal quantity				
Articulation rate	3.406	0.424	3.279	3.534
Pause ratio	0.383	0.108	0.350	0.415
B. Phonological quality				
Maximum posterior probabilities	0.802	0.022	0.795	0.809
Posterior gaps to natives	0.086	0.021	0.079	0.093
C. Prosodic quality				
Pitch variability	56.564	29.449	47.716	65.412
Intensity variability	0.655	0.142	0.612	0.698

Descriptive Statistics of Automated Measures: Study 3

	M	*SD*	95% *CI*	
			Low	Upper
A. Temporal quantity				
Articulation rate	4.117	0.621	3.930	4.304
Pause ratio	0.426	0.159	0.378	0.474
B. Phonological quality				
Maximum posterior probabilities	0.771	0.029	0.762	0.779
Posterior gaps to natives	0.075	0.022	0.069	0.082
C. Prosodic quality				
Pitch variability	43.040	18.482	37.487	48.592
Intensity variability	0.591	0.119	0.556	0.628

Supporting Information-D: Oral Interview Materials

Describe the hardest and toughest challenge in your life.

Your story should start with the following words:

One of the hardest/toughest challenges in my life was ______________________

- ➢ Discussion points
  - ✓ When? How old and where were you?
  - ✓ Why did you encounter this challenge?
  - ✓ Why was it so challenging?
  - ✓ Did anybody (e.g., friends, parents) help you?

- ➢ Rounding off questions
  - ✓ What did you learn from this experience?
  - ✓ Would you like to go through the same experience again?