Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Relating pronunciation distance metrics to intelligibility across English accents

Tessa Bent [a,*], Malachi Henry [a], Rachael F. Holt [b], Holly Lind-Combs [b]

[a] Indiana University, Department of Speech, Language and Hearing Sciences, Bloomington, IN, 47408, USA
[b] The Ohio State University, Department of Speech and Hearing Science, Columbus, OH, 43210, USA

ARTICLE INFO

Article history:
Received 20 July 2023
Received in revised form 16 August 2024
Accepted 2 September 2024
Available online 17 September 2024

Keywords:
Speech perception
Intelligibility
L2 accents
L1 accents

ABSTRACT

Unfamiliar accents can cause word recognition challenges, particularly in noisy environments, but few studies have incorporated quantitative pronunciation distance metrics to explain intelligibility differences across accents. To address this gap, intelligibility was measured for 18 talkers -- two from each of three first-language, one bilingual, and five second-language accents -- in quiet and two noise conditions. The relations between two edit distance metrics, which quantify phonetic differences from a reference accent, and intelligibility scores were assessed. Intelligibility was quantified through both fuzzy string matching and percent words correct. Both edit distance metrics were significantly related to intelligibility scores; a heuristic edit distance metric was the best predictor of intelligibility for both scoring methods. Further, there were stronger effects of edit distance as the listening condition increased in difficulty. Talker accent also contributed substantially to intelligibility models, but relations between accent and edit distance did not consistently pattern for the two talkers representing each accent. Frequency of production differences in vowels and consonants was negatively correlated with intelligibility, particularly for consonants. Together, these results suggest that significant amounts of variability in intelligibility across accents can be predicted by phonetic differences from the listener's home accent. However, talker- and accent-specific pronunciation features, including suprasegmental characteristics, must be quantified to fully explain intelligibility across talkers and listening conditions.

1. Introduction

Although there is substantial variability across talkers in the way phonemes and words are realized, listeners are generally able to recover the intended linguistic messages (Heald & Nusbaum, 2014; Kleinschmidt & Jaeger, 2015; Pierrehumbert, 2016). However, some sources of variability can lead to decrements in intelligibility. One factor that can lead to substantial intelligibility challenges is the presence of an unfamiliar accent, whether it be from regional differences across talkers communicating in their first language (L1) or accent differences stemming from L1 influences when communicating in a second language (L2) (Adank et al., 2009; Munro & Derwing, 1995). These communication challenges can be particularly acute when an unfamiliar accent is combined with environmental degradation, such as background noise or reverberation (Clopper & Bradlow, 2008; Munro, 1998; Rogers et al., 2004; Wilson & Spaulding, 2010). Although poorer accuracy has been observed for both unfamiliar L1 and L2 varieties, some research suggests that the pronunciation patterns found in unfamiliar L2 accents are more challenging to overcome than those in unfamiliar L1 varieties (Adank et al., 2009; Bent et al., 2016) with some research even suggesting that listeners recruit different processing mechanisms during the perception of these two accent types (Floccia et al., 2006; Goslin et al., 2012). However, not all studies support this pattern (Bent et al., 2021; Levy et al., 2019) and other research suggests that disruptions to speech processing are similar across less familiar L1 and L2 accents (Floccia et al., 2009). Furthermore, it has long been known that not all L2 speakers, even those with strong L2 accents, are difficult to understand (Munro & Derwing, 1999).

One of the challenges for reconciling these results is that many intelligibility studies do not include precise characterizations of their stimuli. That is, most studies, including our own,

[*] Corresponding author at: Indiana University, Department of Speech, Language and Hearing Sciences, 2631 East Discovery Parkway, Bloomington, IN 47408, USA.
E-mail address: tbent@iu.edu (T. Bent).

https://doi.org/10.1016/j.wocn.2024.101357
0095-4470/© 2024 The Authors. Published by Elsevier Ltd.

that test perception of unfamiliar L1 and L2 accents simply state that the variety is different from the local norm but do not quantify the extent to which the varieties differ from one another (Baese-Berk et al., 2023). Studies may provide general descriptions of how the included accents differ from the local standard (e.g., Bent et al., 2016; Clopper & Bradlow, 2008) but do not typically measure the acoustic–phonetic characteristics of their specific stimuli nor integrate these types of measures into statistical modeling of intelligibility.

Part of the challenge to incorporating pronunciation distance metrics into intelligibility studies is that there is no consensus in the field about how distance from the local variety should be measured. Researchers have taken a range of different approaches to determining what specific acoustic–phonetic characteristics are leading to word recognition decrements across talkers and accents. Central approaches to this problem can be classified into three broad categories, which we describe in more detail below: the transposing approach, the computational acoustic approach, and the phonetic distance approach.

1.1. Transposing approaches

One approach to determining broadly which aspect of L2 accents are causing intelligibility decrements is to synthetically transpose specific dimensions of speech between L1 and L2 speakers. With this approach, researchers can determine how shifting characteristics, such as intonation or rhythm, to make them more or less L1-like impacts intelligibility (Sereno et al., 2016; Tajima et al., 1997; Winters & O'Brien, 2013). The results of these studies have been inconsistent. Changing the rhythmic properties of L2 speech to be more L1-like was shown to increase intelligibility in one study (Tajima et al., 1997) while it decreased intelligibility in another (Winters & O'Brien, 2013). Two studies found that changing intonation patterns to be more L1-like decreased intelligibility (Sereno et al., 2016; Winters & O'Brien, 2013). Thus, synthetic manipulations tend to have negative effects on intelligibility relative to the talker's original productions making it difficult to determine how proximity to L1 norms in these features impacts intelligibility. The manipulations may have introduced unnatural qualities to the speech demonstrating the limitations of this approach. Furthermore, they do not allow for understanding how interactions among multiple phonetic differences or between segmental and suprasegmental features may impact intelligibility for naturally produced speech.

1.2. Computational acoustic approaches

A second class of approaches may be described as computational acoustic approaches. In these designs, researchers have used computational techniques to automatically measure acoustic distance between a talker's and listener's utterances or between a representation of the local L1 accent and L2-accented stimuli and related these measures to human intelligibility scores or accent strength judgements. One example of this approach is from Pinet et al. (2011) who showed that intelligibility could be predicted by the acoustic similarity between a listener's speech patterns and an average of four talkers from the accents in their study that included both L2 (higher and lower proficiency French-accented talkers) and non-local L1 (Irish English) accents. The acoustic measure they employed, ACCDIST (Huckvale, 2007), automatically compares the acoustic similarity of recordings. Although ACCDIST has primarily been employed for automatic accent classification, it showed promise in predicting intelligibility. One limitation of this specific metric is that it only incorporates vowels. Furthermore, at least as employed by Pinet et al. (2011), the method required making recordings of each listener in the study and comparing these recordings to the speech stimuli used in the intelligibility tests. It seems possible that the approach could be adapted to make comparisons with a set of reference speakers representing the local norm, thus making it more generalizable and increasing feasibility across larger participant samples. However, as far as we are aware, this modification in the approach has not yet been applied to studies of intelligibility.

More recent acoustic distance measures have shown substantial promise in predicting human L2 accent strength judgements across a large set of L2 talkers (Bartelds et al., 2020, 2022; Lind-Combs et al., 2023). Two examples of this approach used dynamic time warping with various speech representations as input to quantify distances between target and reference acoustic signals (Bartelds et al., 2020; 2022). These approaches have some of the advantages of ACCDIST but incorporate all segments (i.e., vowels and consonants). In Bartelds et al. (2020), their distance metric based on Mel Frequency Cepstral Coefficients (MFCCs) was related to human accent strength ratings but was not as strong of a predictor as an edit distance approach (see below). Bartelds et al. (2022) found that one of the deep-learning signal-based models, which uses Transformers to extract the acoustic features, was the best predictor of accent strength in comparison to approaches utilizing edit distance or MFCC-based acoustic features. These acoustic approaches have not yet been applied to modeling intelligibility across accents, although there have been a range of related projects modeling speech intelligibility in different types of maskers (for a recent example, see Martinez et al. (2022)). There are many benefits to using these types of signal-based models; however, Bartelds et al. (2022) noted that they are limited by the availability of large training sets in the specific language, as their assessments of Norwegian were less successful due to the lack of training data compared to English (see San et al. (2024) for a recent approach that improves similar models for low resourced languages). These approaches demonstrate the ability to predict L2 accent strength. Future studies should also evaluate whether they can predict intelligibility across a range of L1 and L2 accent varieties.

The literature reviewed in this section is not an exhaustive description of the computational approaches to measuring pronunciation distance. It should also be noted that there are many other automated tools that have been developed over the past two decades for a range of purposes. For example, Witt and Young (2000) developed an automated "Goodness of Pronunciation" measure that can approximate human performance in determining whether L2 speakers have produced phones in error. This tool was developed and has primarily been applied to language learning contexts. There are also hybrid acoustic approaches that use IPA transcription to inform acoustic modeling. For example, Pongkittiphan et al. (2015) found that an approach that incorporated both linguistic features of a word (e.g., syllable structure) as well as phonetic pronunciation distance derived from dynamic time warping informed by IPA transcription could predict if a Japanese-accented English production had very low (<10%) or moderately low (10–30%) intelligibility for L1 American English listeners. This hybrid approach also could predict accent strength ratings using the same dataset as in Bartelds et al. (2020; 2022) (Shi et al., 2015).

1.3. Phonetic distance approaches

A final approach, and the one we incorporate here, is to use measures of phonetic difference between a reference accent or set of stimuli and the target accents. An advantage of these approaches is that they use linguistically motivated units (i.e., phonemes) to make comparisons across talkers. Early attempts at relating intelligibility scores to the number of phonetic errors for L2 speakers did not show significant associations (Munro & Derwing, 1999). However, there is some evidence that this relation may hold for speakers with less advanced proficiency, novice to intermediate learners (Nagels & Huensch, 2020). More recently, Kang et al. (2020) investigated how a range of phonological features impacted comprehensibility and intelligibility for talkers representing several different accent types, including L1 and L2 varieties. The speaker's productions were evaluated using a phonetic measure that calculated all phonetic divergences from standard American English divided by the number of syllables produced. This phonetic measure predicted intelligibility scores and was a stronger predictor than their prosodic and fluency measures. This study suggests that phonetic deviation measures hold substantial promise in predicting intelligibility differences across L2 and L1 accents. The study was limited by including very few stimulus items per talker (four anomalous sentences each), only using L2 listeners for the statistical modeling, and testing in ideal listening conditions (i.e., quiet). These decisions were primarily driven by the motivation for this specific work which focused on applied questions regarding the appropriate speaker parameters for evaluations of second language learners' abilities (e.g., in TOEFL-like tests). Similarly, Nagels et al. (2023) investigated how a range of phonetic variables known to pose challenges for L1 English speakers acquiring Spanish as an L2 related to intelligibility, comprehensibility ratings, and foreign accent strength ratings. Again, the study only included quiet listening conditions and had two utterances per talker, which the authors note as a substantial limitation. Further, the intelligibility scores were highly skewed towards perfect intelligibility scores and therefore the authors could not compute the relation between their variables and a range of intelligibility scores but rather used a binary measure of intelligibility. This analysis showed several significant relations (i.e., with rising intonation as well as one consonant- and one vowel-based variable), but these only explained about 1–3% of the variance.

Other studies have incorporated edit-distance metrics into investigations of L2 accent strength ratings (Bartelds et al., 2020, 2022; Lind-Combs et al., 2023; Wieling, Bloem, et al., 2014), intelligibility across accents (Bent et al., 2021; Levy et al., 2019), mutual intelligibility across related languages (e.g., Beijering et al., 2008; Gooskens & Heuven, 2020; Gooskens & Schneider, 2019), and intelligibility of speech from children with cochlear implants (Sanders & Chin, 2009). These studies build on measures developed within the field of dialectometry, which investigates large sets of linguistic features to characterize geographically or socially conditioned varieties (for reviews see Nerbonne, 2009; Wieling & Nerbonne, 2015). Rather than counting the number of differences or deviations from the assumed local standard, these metrics measure pronunciation distance for a speaker relative to a reference set using phonetic transcription of the target stimuli in comparison to phonetic transcriptions of one or more talkers representing the local standard. These metrics find the optimal alignment between the phonetic transcriptions of the target and reference stimuli and assign penalties according to differences between the two. Several variations have been evaluated in relation to L2 accent strength ratings including standard Levenshtein Distances and Pointwise Mutual Information (PMI) metrics (Bartelds et al., 2020, 2022; Wieling, Bloem, et al., 2014). The PMI metric incorporates the frequency of pronunciations within a corpus to determine penalties for pronunciation differences. Higher penalties are then assigned for less frequent pronunciation patterns with the idea that listeners are likely to assign higher accent strength ratings to pronunciation patterns that are less frequently encountered. Both of these metrics have been shown to be related to human accent strength ratings.

Fewer investigations have incorporated edit-distance metrics into modeling intelligibility across accents. One intelligibility study with child listeners calculated Levenshtein distances for three stimulus talkers: one local L1, one non-local L1, and one L2 (Levy et al., 2019). They used a weighted version of the Levenshtein algorithm in which the penalties for different types of pronunciation differences were adjusted based on theoretical assumptions from the literature about how pronunciation differences should impact word recognition accuracy. They found that the overall intelligibility scores aligned with the Levenshtein scores (i.e., the least intelligible talker had the highest Levenshtein score), but the researchers did not incorporate the distance scores into their statistical models. Building on their work, Bent et al. (2021) used the same hand-calculated weighted Levenshtein distance algorithm to model intelligibility for talkers representing seven different accents under quiet and noise-added conditions with both adults and children. The accents included both L2 varieties and non-local L1 varieties. The Levenshtein scores were a significant predictor of intelligibility in both quiet and noise-added conditions. Furthermore, a model including both Levenshtein distances and talker accent was a better fit than the one using only the Levenshtein distances. Jurado-Bravo (2021) also used Levenshtein distances to model intelligibility across a set of 15 L1 Spanish speakers producing a passage. Their adaptation of the algorithm compared the stimulus items to English as a Lingua Franca (EFL). In this version called the EFL-Levenshtein Distance (EFL-LD), the stimuli were assessed relative to whether they diverge from a set of pronunciation features that Jenkins (2000) identified as essential for L2 English speakers to ensure their intelligibility. In this version, errors such as producing different vowels or /ɹ/ variations are not penalized whereas other errors, such as consonant substitutions, are. In their model, EFL-LD scores significantly predicted intelligibility scores, again suggesting the utility of these types of edit distance metrics for predicting intelligibility. Their adaptation of the algorithm, however, does not lend itself to investigating intelligibility across L1 varieties because their research aims are focused on L2 speakers of English. The lack of penalties for many types of pronunciation differences also could lead to very compressed scores across talkers, limiting the explanatory value of their adaptation. Related work has shown that Levenshtein distances significantly correlate with word intelligibility scores for children with cochlear implants (Sanders & Chin, 2009).

All three studies that have incorporated Levenshtein distances with intelligibility data across accents have either been limited by including only one accent (Jurado-Bravo, 2021) or including multiple accents but only having one talker representing each accent (Bent et al., 2021; Levy et al., 2019). For the latter studies, there was a confound between talker- and accent-specific effects. It will be essential to include not just a wide range of accents into these models, but also to include more than one talker per accent. The inclusion of multiple talkers per accent will address whether variability in intelligibility can be traced to pronunciation features that are characteristic of a specific accent or whether effects that have been described as “accent effects” are talker-level effects. We begin to address this limitation by including two talkers per accent in the current study to preliminarily disentangle talker and accent level effects. In addition, we evaluate two versions of the Levenshtein (edit) distance metric that are automatically calculated via freely available web-based applications: a heuristic version (Bailey et al., 2022) and weighted version (Heeringa et al., 2023). Previous comparisons of Native Discriminative Learning (NDL) pronunciation distances and Levenshtein distances, showed similar utility in predicting accent strength ratings (Wieling, Nerbonne, et al., 2014), but there have not been any comparisons across edit distance measures for predicting intelligibility. Finally, we investigate how different listening environments (i.e., quiet vs. two background noise levels) may influence the relation between edit distances and intelligibility. Nearly all prior studies that have evaluated the predictive value of edit distances for human perception have tested participants in quiet listening conditions (Bartelds et al., 2020, 2022; Beijering et al., 2008; Gooskens & Heuven, 2020; Gooskens & Schneider, 2019; Levy et al., 2019; Lind-Combs et al., 2023; Wieling, Bloem, et al., 2014; Wieling, Nerbonne, et al., 2014).

1.4. Measurements of intelligibility

We must consider how speech perception success is measured if we want to understand how different accents impact perception. Here, we are focusing specifically on speech intelligibility, with the acknowledgement that intelligibility measures only capture part of the perceptual processes involved in speech communication (Baese-Berk et al., 2023; Beechey, 2022). Even with this focus, there are numerous decisions to be made about how to score the data, which may impact our findings and conclusions. For example, studies have employed phoneme, word, or sentence accuracy (Case et al., 2018; Kent et al., 1994). Some studies allow for variation in grammatical morphemes for content words (Yoho & Borrie, 2018) whereas others do not (Nilsson et al., 1994). Some include all words (Spahr et al., 2012) whereas others focus specifically on keywords (Bamford & Wilson, 1979). If we are considering how the edit distance metrics relate to intelligibility scores, it is also important to consider how exactly we are measuring intelligibility. However, few studies have systematically compared results across different scoring approaches (Baese-Berk et al., 2023). A few studies have used more than one scoring method and have found similar intelligibility results across the scoring methods. For example, Case et al. (2018) showed that scoring intelligibility at the word vs. sentence level led to similar results. Likewise, models of word recognition accuracy in Levi (2015) produced similar findings whether the responses were scored at the phoneme or word level. Recently, Bosker (2021) developed an online implementation of fuzzy-string matching for automatically scoring intelligibility data, Token Sort Ratio (TSR). He showed that the TSR scores correlated well with hand-scored percent words correct data and outperformed two other automated scoring approaches (i.e., Levenshtein distance and Jaro distance). TSR scores also showed the highest alignment with two acoustic markers of intelligibility. The advantages of this approach are its consistency and speed. Although this scoring method has substantial promise and is beginning to be adopted (e.g., Babel, 2022), there is a need for more study of how this measure compares to a more traditional percent words correct measure. The comparisons in Bosker (2021) only employed L1 Dutch speakers in two experiments; thus, the impact that different speaker characteristics may have on the use of this measure should continue to be investigated. In the current study, in addition to comparing two edit distance measures for modeling pronunciation distance, we employ two methods for scoring the intelligibility data. The first method is the token sort ratio (the instantiation of fuzzy string matching from Bosker [2021]) and the second is a traditional percent words correct measure. For our percent words correct measure, we used some automation (as described below) but preprocessing of the data included human decision making, such as whether inaccurate words were the result of typos by using spell check and changing homophones to the correct target word. The percent words correct measure is therefore aligned with more traditional scoring approaches for intelligibility studies.

The purpose of the current investigation was twofold. First, we test the predictive value of two edit distance measures for intelligibility across L1 and L2 varieties of English. The software tools selected to compute these measures are freely available web-based calculators that increase accessibility for researchers and speech-language pathologists without the need for software download and installation. Second, we assess whether different intelligibility scoring approaches impact the modeling results. To address these goals, recordings from 18 talkers representing nine English accents with two talkers per accent were presented to adult listeners in a transcription task. We expect that edit distance scores will predict word recognition accuracy consistent with their predictive value for L2 accent strength scores (Bartelds et al., 2020; Wieling, Bloem, et al., 2014) and intelligibility scores (Bent et al., 2021). The comparison across edit distance scores is more exploratory. Relative to the weighted edit distance measure, the heuristic edit distance measures may be less sensitive in predicting intelligibility because all pronunciation differences are given equal weight without adjustment for the frequency of pronunciation differences from the local standard on intelligibility. Additionally, the models using fuzzy string matching may better align with the edit distance measures because both measures are gradient measures (of perception and production, respectively) compared with binary scores for word recognition which do not give listeners credit for correctly perceiving parts of words.

2. Method

2.1. Participants

Listeners included 369 American English monolingual participants between the ages of 18 – 35 years (average = 26). All participants reported typical speech, language, and hearing abilities. The gender of participants included women (n = 203), men (n = 145), non-binary (n = 15), one each of transgender man, transgender woman, and genderfluid. Three participants did not report their gender. For ethnicity, participants indicated that they were Hispanic or Latino (n = 32), not Hispanic or Latino (n = 332) or prefer not to respond (n = 5). For race, participants identified as white (n = 280), Black or African American (n = 35), Asian or Asian American (n = 23), two or more races (n = 21), Native Hawaiian or other Pacific Islander (n = 1), other (n = 7), or prefer not to say (n = 2). Participants were asked to rate their exposure to all accents included in the study on a scale of 1 (no exposure or only brief casual exposure) to 5 (daily at home exposure). Exposure ratings for the accents were calculated for the participants who received these specific accents in their assigned conditions. The average ratings were as follows with range in parentheses: British English (England) = 2.0 (1–4); Scottish English = 1.6 (1–4); French-accented English = 1.7 (1–3); German-accented English = 1.6 (1–4); Hindi-accented English = 1.9 (1–4); Japanese-accented English = 1.6 (1–4); Mandarin Chinese-accented English = 1.7 (1–4); Spanish-accented English = 3.0 (1–4). Although not all participants indicated daily interaction with speakers from the Midland American region, this variety of English was familiar to all participants due to its similarity to Standard American English. Participants who reported daily exposure to one of the other accents in their condition were excluded (and not included in the subject counts above). Participants rated the level of background noise in their environment as 1.9 on average from a scale of 1 (very quiet) to 10 (very loud) (range = 1–8) with nearly all participants (n = 354) reporting ratings of 4 or less.

An additional 25 participants were tested but their data were not usable due to bilingual language status (n = 4), indicating that they did not turn off all devices that make sound (n = 2), frequent exposure to one of the test accents in their assigned condition (n = 7), hearing or language disorder (n = 2), low effort responses (e.g., leaving multiple trials blank in the Midland condition, n = 9) or reporting the background noise in their environment as a 10 (n = 1). An additional 55 participants did not pass the headphone screening (see detail below) (Woods et al., 2017).

Listeners were paid for their participation at $10.00–12.00 per hour. Most participants were paid $10/hour; the hourly payment was increased in June 2022 to align with recommended compensation on Prolific. Participants who failed the headphone screening were paid for the time they put into the study. All experimental procedures took approximately 25 min and were approved by the Institutional Review Board at Indiana University.

2.2. Stimuli

Sixty sentences from the Hearing in Noise Test for Children (Nilsson et al., 1996) were used as the experimental stimuli. These sentences are short declaratives with three to four keywords per sentence. These sentences were produced by 18 talkers representing 9 different accents with two speakers (1 female and 1 male) representing each variety. The varieties included three L1: Midland American English, Southern Standard British English, and Scottish English, with the male speaker from Glasgow and the female speaker from the Highlands. The first language varieties for the five L2 accents included French (from France), Spanish (Colombian dialect), German (from Germany), Japanese (from Japan), and Mandarin (Beijing dialect). There was also one bilingual variety: Hindi-accented English from India. Recordings from all L2 speakers and the Midland American speakers were taken from the Hoosier Database of Native and Nonnative Speech for Children (Bent, 2014). Both Standard Southern British English speakers, the male Scottish speaker, and both Hindi-accented English speakers were recorded at Indiana University. The Scottish female speaker was recorded at Ohio State University. All IU and OSU recordings were made in sound-attenuated booths using identical high-quality recording equipment (Marantz PDM670 digital recorder) and head-mounted microphones (Shure Dynamic WH20XLR headset microphone). All L2 speakers learned English at the age of 12 or later and had been in the U.S. for 4 years or less. The bilingual speakers began learning both English and Hindi before the age of 3 years. The average age for the speakers was 27 with a range from 18 to 53. All stimuli were recorded and presented as.wav files equalized for RMS amplitude using the scale intensity function in Praat. All recordings are available in our OSF repository: https://osf.io/cnxgt/.

2.3. Procedure

Listeners were recruited through Prolific (https://www.prolific.co/). If participants met the inclusion criteria, the study would appear as one for which they were eligible. The inclusion criteria included being a monolingual speaker of English, living in the U.S. with U.S. citizenship, and between the ages of 18–35 with no hearing difficulties. The residency and citizenship filters were used to try to ensure that participants were L1 speakers of American English. Participants were then directed to Qualtrics to complete the consent form, a background questionnaire, and a headphone screening (Woods et al., 2017). The headphone screening included six trials. Each trial had three pure tones and participants indicated which was the quietest. One of the sounds was 180 degrees out of phase across the stereo channels, resulting in phase cancellation.

The task is designed to be relatively easy if the participant is wearing headphones but difficult if listening over speakers. The participants had three opportunities to pass this headphone screening. If they failed the third attempt, they could not continue with testing. Participants used their own computers and headphones to complete all testing.

After completing the questionnaire and headphone screening, participants were directed to Pavlovia, the online testing platform for PsychoPy (Peirce et al., 2019). Participants were randomly assigned to one of three listening conditions: quiet, +4 dB signal-to-noise ratio (SNR), or 0 dB SNR. The noise was an 8-talker babble with talkers matched in gender to the target speech. The female babble track was taken from Van Engen et al. (2014) and a parallel babble file was developed with male speakers using the same materials, talker types, and number of speakers. A randomly selected section of the babble file that was 1 s longer than the sentence was selected as the masker for each item, with the sentence centered in the babble. Within these noise conditions, they were presented with three talkers representing three different accents. All talkers in each condition were matched in gender. All listeners were presented with the Midland American speaker and then assigned to one of the following accent conditions: (1) Japanese-accented English and Standard Southern British English, (2) German-accented English and Scottish English, (3) Mandarin-accented English and Hindi-accented English, or (4) Spanish-accented English and French-accented English.[1] The task began with nine practice trials that included three sentences from each of the talkers in the listener's assigned condition. Then listeners were presented with 60 experimental trials, including 20 sentences from each of their assigned talkers. The sentences were blocked by talker and randomized within a block. Listeners were instructed to listen carefully to each sentence then type in what they heard. They could only listen to each sentence one time and were not provided with any feedback as to the accuracy of their responses. For each accent / listening condition combination, 14–18 participants were tested.

2.4. Analysis

2.4.1. Levenshtein distances

All sentences were phonetically transcribed by two trained research assistants. Inter-transcriber agreement was calculated for a subset of the transcriptions (80 sentences with 10 sentences from 8 of the talkers) and found to be 82.1%. This rate of inter-transcriber agreement is in line with other recent work (Seifert et al., 2020). The two initial transcriptions were compared. Discrepancies were evaluated by a third transcriber and the differences were resolved through further listening, evaluation of the visual representations of the sentences (waveforms and spectrograms in Praat), and discussion with the transcription team. An agreed-upon set of IPA symbols was used for transcription. The transcriptions for each non-Midland speaker were compared to four talkers representing the familiar Midland American English referent. Two of these speakers were also used in the intelligibility testing. An additional two speakers (one male and one female) were included in the referent set to have a slightly more representative sample of the familiar accent variety. These comparisons were used for two edit distance variants. All IPA transcriptions are available to researchers upon reasonable request.

The first edit distance scoring method implemented the Levenshtein Edit Distance App (LED-A; Heeringa et al., 2023) available at https://www.led-a.org/. This application allows the user to implement a variety of methods for calculating Levenshtein distance. Here, we employed the site's PMI-based distance metric as our weighted edit distance measure. This PMI implementation follows Wieling (2012) where segmental distances are weighted based on their frequency in the dataset. The alignments between vowels and consonants are given very high weights to prevent the alignment of consonants and vowels. Other alignments are determined through an iterative process of comparing pairs of transcriptions until the process stabilizes. Ultimately, the weights are scaled from 0 to 1 where the most frequent pairs get scores of 0 and the least frequent get scores of 1. We utilized the option where the scores were normalized for alignment length. Greater detail about the computation of this metric can be found at https://www.led-a.org/docs/PMI.pdf.

The second scoring method employed the Automated Phonetic Transcription Comparison Tool (APTct) available at https://aptct.auburn.edu/ (Bailey et al., 2022). This automated tool is similar to traditional Levenshtein distance algorithms, in which all phoneme differences are assigned a penalty of 1, except for cases in which a vowel is substituted for a consonant or vice versa, in which the penalty is 2; therefore, we refer to this approach as the heuristic edit distance. The algorithm is designed to find the most economical alignment between the strings. More detail about the design and implementation of this scoring can be found in Bailey et al. (2022). For this study, word level transcriptions from the non-local talkers were compared with the four Midland talkers' transcriptions and the lowest distance score among the four comparisons was used for analysis. The approach of using the lower distance score rather than an average across the four reference speakers was determined based on preliminary modeling results suggesting this method resulted in a better fit than the averaging method. Word level scores were then averaged across words in the sentence.

The edit distance metrics were initially compared using two schemes for calculating the reference to identify which method would be a better fit for the perceptual data. The first edit distance scheme incorporated the lowest edit distance value for each word from each of the four referent Midland talkers. Thus, a single referent talker was selected depending on the specific stimulus and variety. The second scheme employed an average of the edit distance values obtained from the four referent Midland talkers for each word. Ultimately, the edit distance metrics performed better during analysis when using the lowest-of-four distances approach, which is implemented here, rather than the average-of-four distances. While the lowest-of-four approach clearly performed better with the current dataset, further evaluation of these approaches could be informative, especially for areas such as automatic speech recognition and natural language processing.

[1] Intelligibility data for the female talker conditions has been presented in Bent & Holt (2018) (condition 1), Bent et al. (2021) (conditions 1, 2, and 3), Bent et al. (2023) (conditions 1, 2, and 3). A preliminary version of this dataset was presented in Bent & Holt (2023).

2.4.2. Intelligibility scoring

Intelligibility scores were calculated two ways. The first method is the traditional binary word scoring method. To calculate these scores, participants' responses were first spell checked and obvious typos, misspellings and homophones were corrected. The transcripts were automatically scored as correct or incorrect using a Python script for each word in the sentence. A strict scoring criterion was applied except that the following substitutions were allowed following HINT scoring procedures: a(n)/the, has/had, have/had, is/was, and are/were. Scores were then calculated for percent words correct (PWC) per sentence. The second scoring method employed was Fuzzy String Matching, specifically the token sort ratio (TSR) from Bosker (2021) using the online implementation (https://tokensortratio.netlify.app). These scores range from 0 to 100. Sentences that match the target exactly are given a score of 100 and those without any matching characters are given a 0. These scores tend to be higher than percent words correct scores with a strict scoring criterion because responses are given credit for partial matches.

3. Results

3.1. Correlations among distance and intelligibility measures

Pearson product-moment correlation coefficients were computed with a Bonferroni correction to account for multiple comparisons to assess the relations between the two edit distance measures and the two measures of intelligibility. The two intelligibility measures – PWC scores and TSR scores – were strongly positively correlated across listening conditions, r (22138) = 0.915, p < 0.001, and within each listening condition with correlation coefficients of r(7318) = 0.837, p < 0.001 in the quiet listening condition; r(7558) = 0.904, p < 0.001 in the + 4 dB SNR condition; and r(7258) = 0.923, p < 0.001 in the + 0 dB SNR condition (Fig. 1). Strong positive correlations were also found between the weighted and heuristic edit distance metrics with a correlation coefficient of r (22138) = 0.724, p < 0.001 (Fig. 2). The reported correlation coefficients suggest strong positive relationships between the two edit distance measures and between the two measures of intelligibility, warranting further examination of these variables with mixed effects modeling to identify the best-fitting model parameters for the behavioral data.

3.2. Mixed effect modeling

Mixed effects modeling analyses were performed using R Version 2023.03.1 + 446 release for macOS (R Core Team, 2023). The statistical code is available at https://osf.io/cnxgt/. To determine the predictive value of the two edit distance scoring methods for predicting intelligibility, four mixed-effects beta regression models were built using the "glmmTMB" package (Brooks et al., 2017). Models incorporated either the fuzzy string matching score (i.e., TSR score) or the PWC score as the outcome variable with one of the edit distance scores. All models included SNR as fixed effects along with their interactions. Maximally designed models included random by-participant and by-item intercepts and slopes for SNR, edit-distance measure, and their interaction. These overspecified models failed to converge, and were then stepped down through the incremental removal of one random effect term from the random effect structure, beginning with the term with the highest correlation to other variables in the random effect structure, until convergence was achieved. The resulting models included by-participant and by-item random intercepts and slopes for the edit distance measures only. Continuous response variables were scaled to values greater than zero and less than one by dividing intelligibility scores by 100. To ensure that data fit into a beta distribution, intelligibility scores of 0 were replaced with a value of 0.001 and scores of 1 were replaced with 0.999. Categorical variables were dummy coded with Midland serving as the referent accent and Quiet as the referent listening environment. Type III Wald chi-square tests were conducted for each model to identify significant effects for edit distance scores, SNR, and their interaction. After evaluating significant effects, model comparisons using the Akaike Information Criterion (AIC) were conducted using the "AICcmodavg" package (Mazerolle, 2023) to determine which of the two edit scores resulted in a better fit for each intelligibility measure. AIC was used rather than other methods for model comparison due to the tendency for log likelihood to covary with AIC and differences in predictor variables, despite each model having the same number of parameters. The two best-fit models were then compared by conducting a seemingly unrelated regression (SUR) equation using the "systemfit" package (Hamann, 2023). SUR allows for two or more models with differing response variables to be compared using R2 and Root Mean Square Error. After assessing the variance explained through the SUR system of equations, a full mixed-effects beta regression model was constructed to examine the relation between accent and intelligibility scores.

3.3. Edit distance with TSR score

Two mixed-effects beta regression models were built with TSR score as the response variable. Type III Wald chi-square tests of the two TSR models (Table 1) revealed significant main effects of both edit distance variant and SNR, which indicates that edit distance and listening condition independently contributed to intelligibility such that higher edit distance and more difficult listening conditions were associated with lower intelligibility. The interaction between both of the edit variants and SNR (Fig. 3) was also significant, indicating that edit distance scores impacted intelligibility differentially for each of the three listening conditions. In the most difficult listening condition, there was a steep decline in intelligibility once edit scores diverged from zero. In contrast, in the quiet condition, high intelligibility was maintained until productions diverged substantially from zero. TSR model selection based on AIC indicated that the heuristic edit distance model (AIC=−51905.51) was the best-fit model with the lowest AIC value compared to the weighted edit distance model (Δ AIC=339.33). A likelihood ratio test comparing the two models was also completed, X^2(13) = 339.33, p < 0.001, with a log-likelihood of 25,966 and a deviance of −51932 which indicated a significantly better fit. Thus, the heuristic edit distance scores were a better fit for predicting intelligibility, as measured by the Token Sort Ratio method.

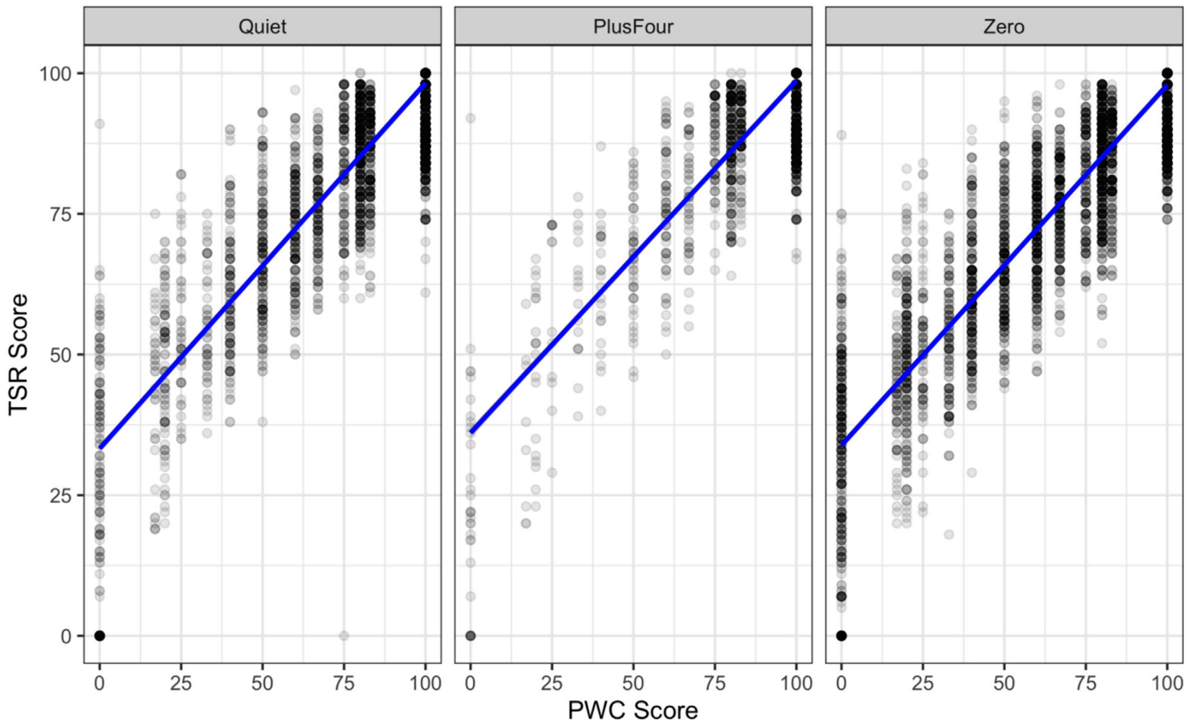

Fig. 1. Scatterplots showing the correlation between Token Sort Ratio (TSR) scores and Percent Word Correct (PWC) scores for the three listening environments. Note: Each data point is made slightly transparent with darker points on the plot indicating more overlap in the data.

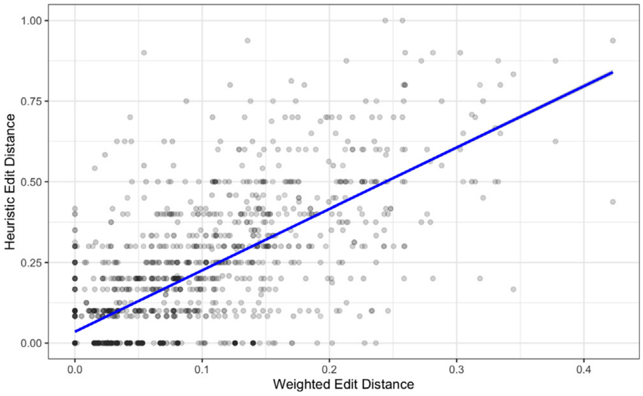

Fig. 2. Scatterplot showing the correlation between the two edit distance measures. Note: Each data point is made slightly transparent with darker points on the plot indicating more overlap in the data.

In addition to these models, we conducted an exploratory analysis using simple correlations between the edit distances and TSR scores to allow for more direct comparison to prior literature in which correlational analyses were used. Again, following prior literature (e.g., Saito et al, 2023), the data input to the correlations were averages for each talker. Specifically, we investigated the relation between the average TSR score for each talker in each listening condition by their average edit distance score (heuristic or weighted) (Fig. 3). Moderate negative relations, approaching or reaching significance, between TSR score and weighted edit distance were observed in quiet, $r(16) = -0.46$, $p = 0.057$, at + 4 dB SNR, $r(16) = -0.49$, $p < 0.05$, and at 0 dB SNR, $r(16) = -0.61$, $p < 0.01$. Moderate, significant negative relationships were also found between TSR score and heuristic edit distance at in quiet, $r(16) = -0.60$, $p < 0.01$, +4 dB SNR, $r(16) = -0.52$, $p < 0.05$, and at 0 dB SNR, $r(16) = -0.58$, $p < 0.05$. Due to the number of correlations in this analysis, the p-values reported should be interpreted with caution. We did not employ Bonferroni correction because Bonferroni correction may be too conservative to be applied with the added analyses; it may excessively inflate Type II error (Armstrong, 2014). These patterns, as well as the parallel analysis below, can be used for broad comparisons to prior work and as a starting point for future studies.

3.4. Edit distance with PWC score

Two parallel models to the analyses with the TSR scores were built to evaluate how the predictive value of the edit distance scores may differ when PWC scores were used as the response variable. Type III Wald chi-square tests of the three PWC models (Table 1) again revealed significant main effects of both edit distance variants and SNR indicating that phonetic distance and listening environment independently influenced the PWC scores. Again, intelligibility decreased as edit distance increased (Fig. 4). Intelligibility also decreased as the SNR became more difficult. The significant interactions between both of the edit distance variants and SNR again show that the influence of phonetic distance varies across different listening environments. PWC model selection based on AIC indicated again that the model with the heuristic edit distance scores ($AIC = -35928.25$) was the best-fit model with the lowest AIC value compared to the weighted edit distance model ($\Delta\ AIC = 391.45$). A likelihood ratio test comparing models, also showed that the heuristic edit distance model was the best fit, X^2 (13) = 391.45, $p < 0.001$, with a log-likelihood of 17,977 and a deviance of -35954 indicating a substantially better fit. Therefore, the heuristic edit distance scores were a

Table 1
Results of Type III Wald Chi-Square Tests for four mixed-effects beta regression models predicting intelligibility scores.

Response Variable	Model	Intercepts, Effects, & Interactions	X^2	*df*	*p*
Token Sort Ratio	Weighted Distance Model	Weighted-Distance by Quiet SNR (Intercept)	7195.23	1	<0.001
		Weighted-Distance	10.45	1	<0.001
		SNR	8.97	2	<0.05
		Weighted-Distance by SNR	142.60	2	<0.001
	Heuristic Distance Model	Heuristic-Distance by Quiet SNR (Intercept)	7123.51	1	<0.001
		Heuristic-Distance	19.69	1	<0.001
		SNR	9.07	2	<0.05
		Heuristic-Distance by SNR	139.93	2	<0.001
Percent Words Correct	Weighted Distance Model	Weighted-Distance by Quiet SNR (Intercept)	5128.75	1	<0.001
		Weighted-Distance	26.66	1	<0.001
		SNR	60.44	2	<0.001
		Weighted-Distance by SNR	170.94	2	<0.001
	Heuristic Distance Model	Heuristic-Distance by Quiet SNR (Intercept)	5285.52	1	<0.001
		Heuristic-Distance	45.36	1	<0.001
		SNR	57.01	2	<0.001
		Heuristic-Distance by SNR	168.21	2	<0.001

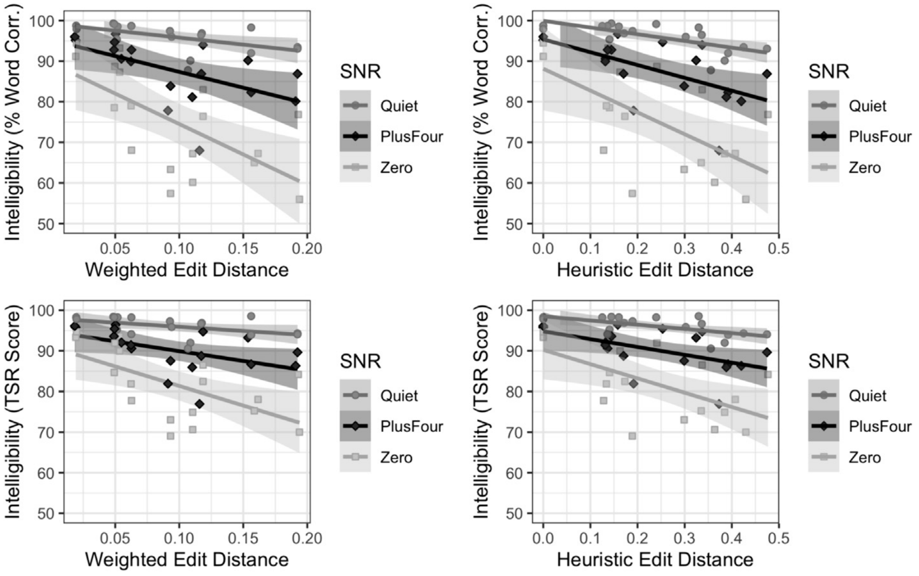

Fig. 3. Scatterplots showing the speaker-level correlations between percent words correct score (top) and TSR score (bottom) and the two edit distance metrics, weighted edit distance (left) and heuristic edit distance (right) for each level of SNR.

better fit for predicting intelligibility when accuracy was calculated with both Percent Words Correct and the Token Sort Ratio.

In parallel with the correlation analysis above, we also calculated correlations for edit distance scores with percent words correct (Fig. 3). Similar to the results of the correlations with TSR scores, significant negative correlations between percent words correct and the weighted edit distance were found in quiet, $r(16) = -0.69$, $p < 0.01$, at + 4 dB SNR, $r(16) = -0.59$, $p < 0.05$, and at 0 dB SNR, $r(16) = -0.61$, $p < 0.01$. Moderate negative correlations were also observed between percent words correct and heuristic edit distance in quiet, $r(16) = -0.55$, $p < 0.05$, at + 4 dB SNR, $r(16) = -0.55$, $p < 0.05$, and at 0 dB SNR, $r(16) = -0.65$, $p < 0.01$.

3.5. SUR model fitting with PWC and TSR

A winning model predicting PWC and another winning model predicting TSR were identified and selected using AIC comparison. However, these two models differ in that they are built around two distinct and different response variables. Therefore, a comparative analysis that accounts for multiple models with differing dependent variables is needed. One such method is to build a SUR model, which consists of two or more regression equations, each having its own dependent variable as well as varying sets of exogenous explanatory variables. To compare the two best-fit models, a system of regression equations was built using the SUR method for model comparison with the systemfit() function (Hamann, 2023), which allows for comparisons between models with separate dependent response variables. To further examine the predictive value of the edit distance scores for intelligibility, R^2 and Root-Mean-Square Error (*RMSE*) values were compared between the weighted edit distance model with TSR score as the response variable (automated model) and the weighted edit distance model with PWC score as the response variable. Very little difference was found between the R^2 values for the two models with the TSR model accounting for approximately

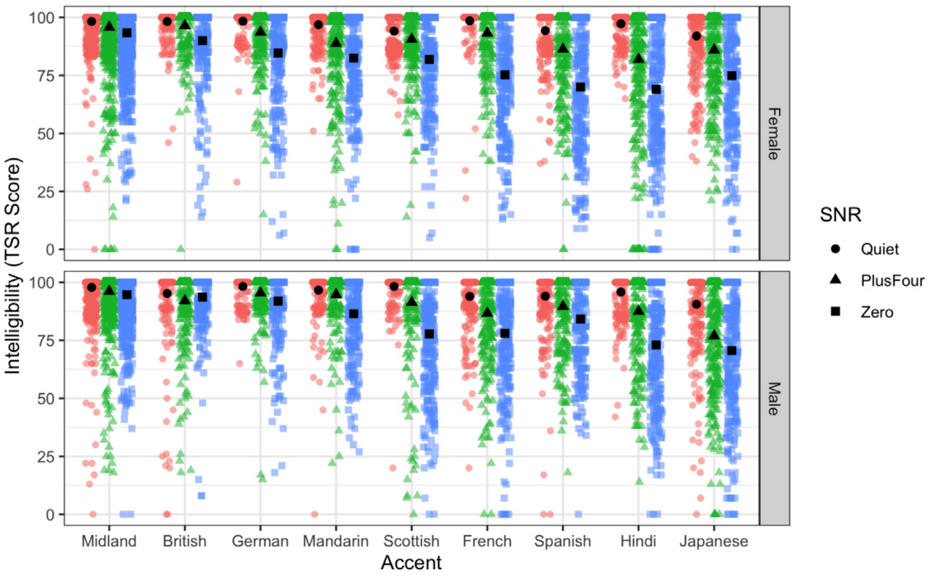

Fig. 4. Mean TSR scores by accent and talker gender (female in top panel; male in bottom panel) for the three listening conditions: quiet (circles, red), +4 dB SNR (triangles, green), and 0 dB SNR (squares, blue). Small symbols represent individual listener's scores for each sentence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6.8% of the variance (R^2 = 0.068397) and the PWC model accounting for slightly more variance with 9.5% of the variance accounted for (R^2 = 0.095271). However, there were differences between the models when considering their predictive quality. The TSR model appears to be the better predictor of intelligibility (*RMSE*=11.5% TSR score) compared to the PWC model (*RMSE*=15.8% PWC score). Considering the lack of difference in the variance explained between the two models and the higher predictive power of the TSR model, the heuristic edit distance score and TSR score were used in a full mixed-effects beta regression model to examine the relation between accent and intelligibility scores.

3.6. Intelligibility scores and talker accent

We next investigated the contribution of talker accent to intelligibility scores. Fig. 4 displays each mean talker's intelligibility in each listening condition. Although the figure is divided by talker gender, we do not include gender in any of our analyses as none of our research questions address the impact of gender and we only have one male and one female talker for each accent. A full mixed-effects beta regression model was built that included fixed effects of heuristic edit distance scores, SNR, and talker accent, as well as three-way interactions between each of the variables. The model also included by-item random intercept and slopes for accent and the heuristic edit distance score, and by-participant random intercept and slope for the heuristic edit distance score. Summary results are shown in Table 2 with full results shown in Appendix A.

Table 2
Output of Type II Analysis of Variance Table with Satterthwaite's method for full model of intelligibility.

Effects & Interactions	X^2	*df*	*p*
Heuristic-Distance	7.27	1	< 0.05
SNR	43.44	2	< 0.001
Accent	137.49	8	< 0.001
Heuristic-Distance x SNR	16.55	2	< 0.001
Heuristic-Distance x Accent	40.77	7	< 0.001
SNR x Accent	483.68	16	< 0.001
Heuristic-Distance x Accent x SNR	21.67	14	0.086

All three main effects were significant. The main effect of edit distance scores arose because items with higher edit distance scores were less intelligible than those with lower edit distance scores. That is, productions that diverged more from the local accent were more difficult for listeners to understand than those closer to the local accent. The main effect of SNR resulted from the highest accuracy in quiet and lowest accuracy in the 0 dB SNR with intermediate performance in the + 4 dB SNR condition. The main effect of accent was due to the differences in intelligibility across accents with highest accuracy for the Midland American English and the Southern Standard British English accents and lowest accuracy for the Hindi and Japanese accents. The two-way, but not the three-way, interactions were significant (Fig. 5). The SNR by Accent interaction arose because some accents were highly intelligible even in the most difficult SNRs (e.g., Midland and British) whereas other accents showed much larger intelligibility declines, particularly in the most difficult SNR. The edit distance by Accent interaction arose because the extent to which edit distance scores and TSR scores were related differed across accents. Although there was a significant interaction between edit distance and accent, the patterns between these two variables were not necessarily consistent for the two talkers representing each accent (Fig. 5) suggesting that interlanguage phonology patterns are not the primary driver of these relations.

3.7. Pronunciation differences post hoc analysis

To further examine the relation between intelligibility and pronunciation distance, a follow-up analysis of the distribution of pronunciation differences (i.e., insertions, deletions, different substitution types) was conducted using Python scripting to

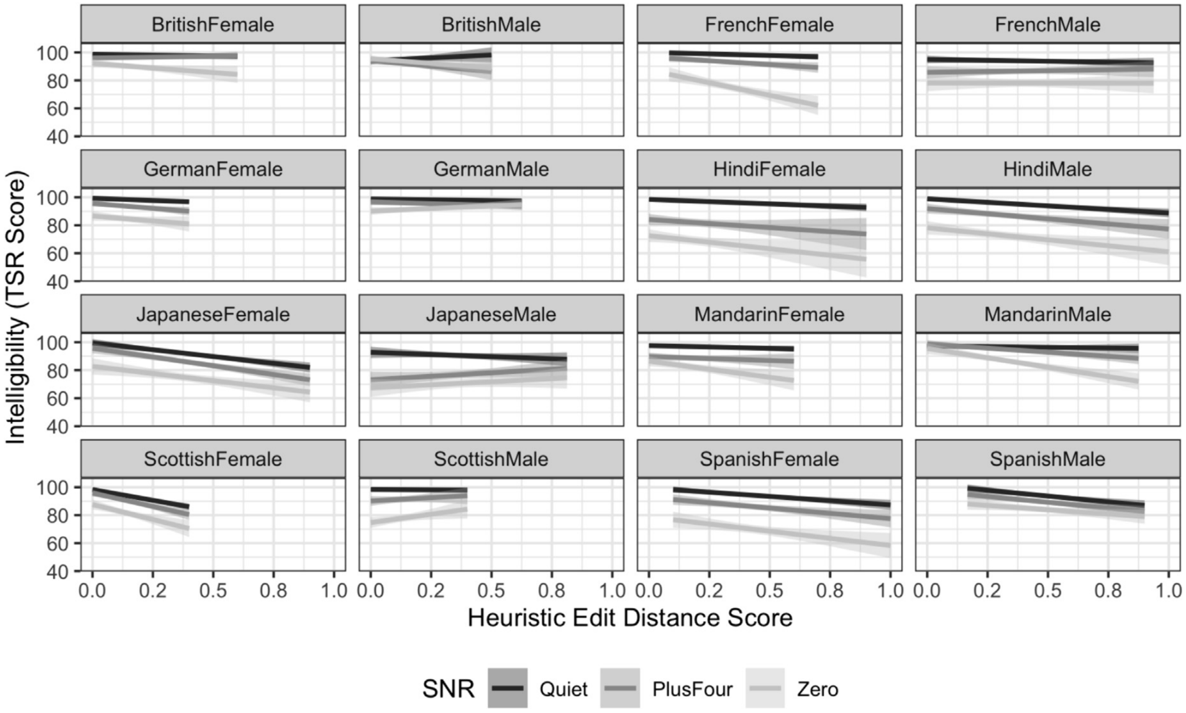

Fig. 5. Intelligibility (TSR) scores by heuristic edit distance score for each nonlocal talker as a function of listening condition. Standard error displayed by shaded bands.

extract and categorize differences within the two transcription strings. Manual processing of the full output was conducted by two research assistants to ensure the reliability of the results, resulting in an average agreement of 97% between the script and a human rater. Remaining differences in categorization were discussed and agreed upon by the two raters. Pronunciation differences were assigned to one of the following six categories: consonant substitutions, vowel substitutions, deletions, insertions, consonant for vowel substitutions, and vowel for consonant substitutions. The overall distributions across all speakers (Fig. 6) and for the individual talkers (Fig. 7) are shown below.

Again, simple correlations were calculated between the two intelligibility metrics and the two categories of pronunciation differences with the largest proportion of differences. Speaker-level averages for intelligibility at each SNR level and the pronunciation difference frequency were used to calculate the correlation coefficients corresponding to the correlations displayed in the scatterplots in Fig. 8. The analysis revealed moderate negative correlations among intelligibility and the number of consonant and vowel differences. Moderate negative correlations, approaching or reaching significance, were observed between percent words correct and the frequency of vowel substitutions in quiet, $r(16) = -0.56$, $p < 0.05$, at + 4 dB SNR, $r(16) = -0.44$, $p = 0.065$, and at 0 dB SNR, $r(16) = -0.52$, $p < 0.05$, as well as consonant substitutions in quiet, $r(16) = -0.66$, $p < 0.05$, at + 4 dB SNR, $r(16) = -0.54$, $p < 0.05$, and at 0 dB SNR, $r(16) = -0.58$, $p < 0.05$. Moderate negative correlations, again approaching or reaching significance, were also observed between TSR score and the frequency of vowel substitutions in quiet, $r(16) = -0.51$, $p < 0.05$, at + 4 dB SNR, $r(16) = -0.40$, $p = 0.10$, and at 0 dB SNR, $r(16) = -0.49$, $p < 0.05$ as well as consonant substitutions in quiet, $r(16) = -0.54$, $p < 0.05$, at + 4 dB SNR, $r(16)= -0.45$, $p = 0.061$, and at 0 dB SNR, $r(16)= -0.54$, $p < 0.05$. The relation between intelligibility and substitutions tended to be slightly stronger for consonant substitutions than vowel substitutions.

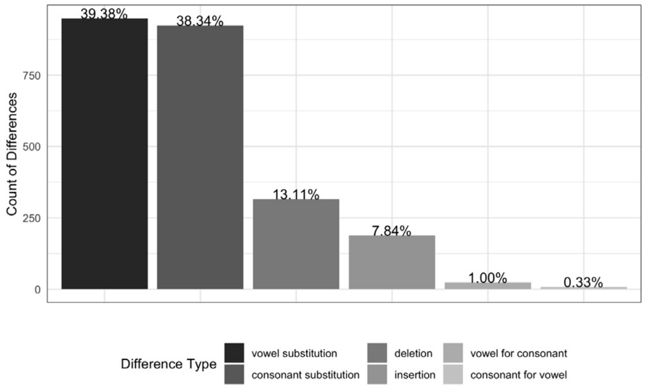

Fig. 6. Frequency of each category of pronunciation differences across all talkers. The percentage for each category reflects how often that pronunciation difference occurs out of the total number of pronunciation differences.

4. Discussion

Talkers with unfamiliar accents tend to be less intelligible than those with local, familiar accents, but the extent to which listeners have difficulty understanding unfamiliar accents can vary widely. Understanding why specific accents or talkers cause varied amounts of challenge for word recognition success is important for speech perception theories as well as practical applications related to pedagogy, experimental

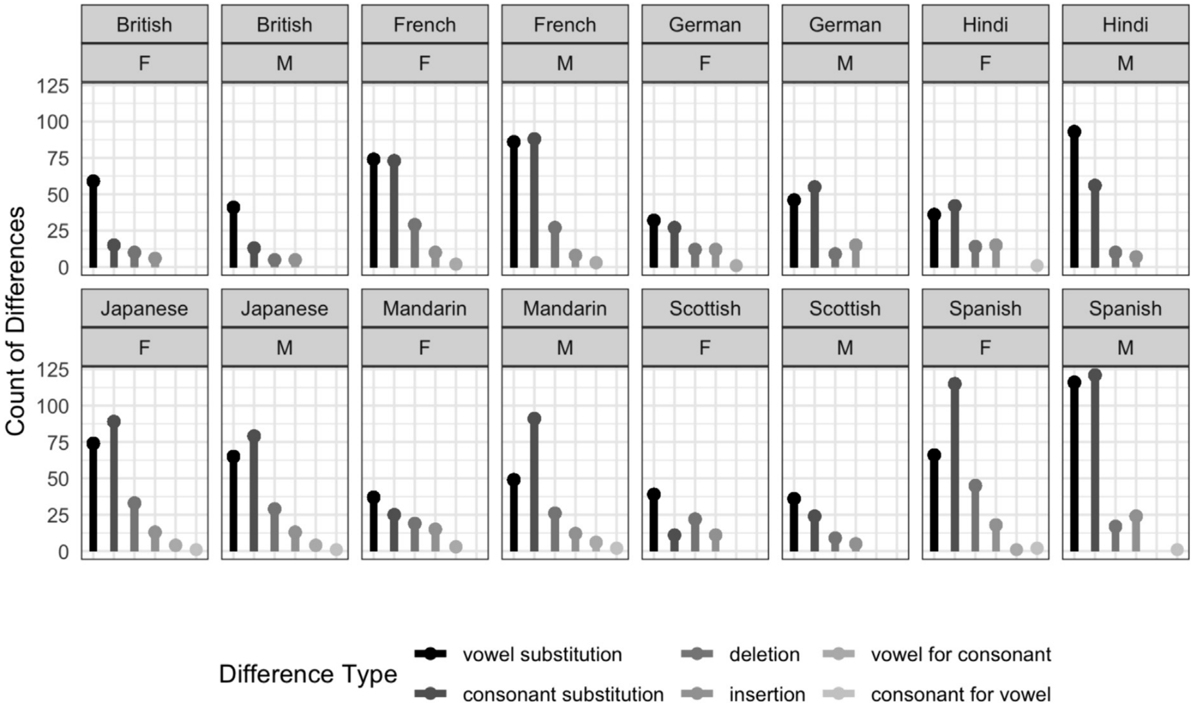

Fig. 7. Lollipop plots showing the frequency of each pronunciation difference category for each talker.

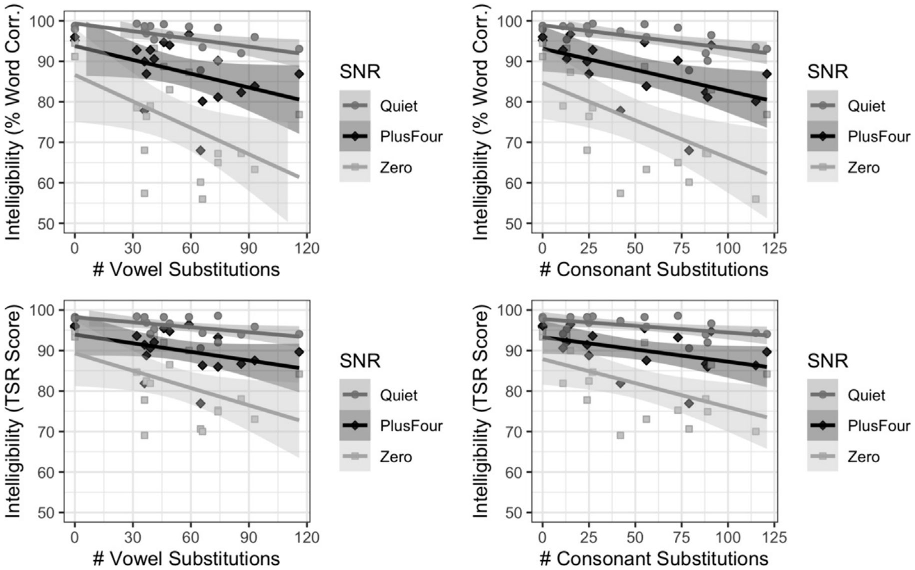

Fig. 8. Scatterplots showing the speaker-level relations between percent words correct score (top) and TSR score (bottom) and the frequency of vowel substitutions (left) and consonant substitutions (right) as a function of SNR.

design, and clinical practice. However, not only does the field not agree upon methods for a priori determination of which talkers are likely to be more or less intelligible, some investigators have lamented that “as every phonetician or phonologist knows, it is very difficult, if not impossible, to quantify the amount of foreign or regional accent within an utterance” (Floccia et al., 2009, p. 402). Since this time, the field has moved toward incorporating quantitative methods for capturing distances from the local, familiar accent and L2 or regional accents (e.g., Bartelds et al., 2020, 2022). This study uses metrics that have been developed in the field of dialectometry (for a review see Wieling & Nerbonne, 2015). Incorporating these quantifiable metrics for acoustic–phonetic distance among accent or dialect varieties into studies with objective word recognition data is an essential step for predicting intelligibility across talkers.

To address these issues, this study investigated intelligibility in quiet and two noise-added conditions across nine accents, both L1 and L2 varieties, with two talkers representing each accent. The primary goal was to examine the explanatory value of two edit distance variants, which measure phonetic differences from a reference accent, for intelligibility scores across listening environments using accessible, recently developed web applications for the edit distance calculations. Further, the impact of two scoring methods for listener transcriptions on the relation between these distance metrics and intelligibility results was investigated.

The edit distance metrics included two automatically scored metrics – weighted and heuristic. The weighted version assigns penalties based on the frequency of alignments across the stimulus set with higher penalties for less frequent alignments. The heuristic metric gives equal penalties to all differences except for consonant-for-vowel and vowel-for-consonant substitutions and is not sensitive to frequency of substitutions. Although these methods result in slightly different edit distance scores depending on the types of differences between the target and reference accent, the two metrics are strongly, positively correlated with one another. Furthermore, both edit distance metrics significantly predicted variance in intelligibility in the models, but the heuristic metric explained more of the variance in both models. Thus, these edit distance metrics are both capturing important cues for intelligibility; while the heuristic slightly outperformed the weighted one for these materials, both account for a significant amount of variance in human intelligibility. There are advantages to weighted models in that they capture distances that are more "linguistically sensible" (Wieling et al., 2014, p. 260). It remains possible that the weighted distance metric would outperform the heuristic one with a larger dataset or with different materials. Similar findings were observed in the exploratory correlational analyses, which compared the edit distances (weighted and heuristic) and intelligibility scores (percent words correct and Token Sort Ratio) at the talker level, divided by listening condition.

Future studies should continue to compare different types of pronunciation distance measurement to determine whether other variants of edit distance measurement can better predict intelligibility variation among talkers and accents than the edit distance measures employed here. A range of variables could be independently evaluated in terms of their predictive value. For instance, one limitation of our study was that our weighted variant normalized for word length, but our heuristic one did not. Future studies should test these parameters independently. There are also multiple ways to calculate weightings, which could be tested in relation to intelligibility across listening environments. Exploring the interactions among weightings, word length normalization, as well as other variants such as multiple sequence alignment procedures (List, 2012) in relation to intelligibility specifically may be beneficial. Thus, this study is not an exhaustive investigation of all edit distance variants and their relation to intelligibility scores. It represents an initial step at examining the predictive value of two edit distance measures for intelligibility using web-based tools. Both within LED-A (used here for the weighted variant) and other available software packages (e.g., LingPy, List & Forkel, 2024), there are many edit distant variants that could be tested.

The effectiveness of signal-based intelligibility models for predicting intelligibility across listening conditions could also be a particularly fruitful future direction. These metrics have shown substantial promise in their relation to human perception, particularly with rating data, such as accentedness and comprehensibility ratings (e.g., Bartelds et al., 2022; Saito et al., 2023). Because there is not a deterministic relationship between rating data, such as accentedness or comprehensibility, and intelligibility, it is possible that different distance metrics could explain most variance across different types of perceptual measures. Indeed, the correlation coefficients observed in our study between intelligibility scores and edit distances were slightly lower in magnitude to previous studies that have compared accent strength ratings data completed by human raters with Levenshtein distances (e.g., Wieling, Bloem, et al., 2014) as well as relations between machine-based algorithms measuring pronunciation distance (Saito et al., 2023 and references therein) and human rating data. Most of these studies had correlation coefficients between 0.6 – 0.9, whereas ours fell primarily in the 0.5 – 0.6 range. It may be that pronunciation distances, whether based on automated machine-based computation or measures requiring substantial human input, are more highly correlated with rating data, such as ratings of accent strength (Wieling, Bloem, et al., 2014), comprehensibility (Saito et al., 2023), fluency (Cucchiarini et al., 2000, 2002), or proficiency (Kang & Johnson, 2018). The differences in strength of these relations may be explained by the differences in types of perceptual data being assessed. For example, even speech that has relatively high accentedness can still also be highly intelligible (Munro & Derwing, 1999) so that there are cases in which a talker's production diverges from the local standard would receive an elevated pronunciation distance score and likely a higher score on a measure like comprehensibility or accent strength but listeners could still accurately recognize the word. More work that examines the relation between pronunciation distance metrics (edit distance or computational acoustic) and intelligibility could help to explain why these relations may be stronger with rating data than intelligibility data.

Determining the extent to which intelligibility (and perceived accentedness) are impacted by prosodic variation is an important next step. Although the transposition approaches have attempted to address this issue, the introduction of unnaturalness by the synthetic manipulations has limited the conclusions that can be drawn. Signal-based intelligibility models provide one way of incorporating information below and beyond the phoneme level. Other approaches could have separable, quantifiable metrics for different phonetic and prosodic distances from the home accent that would allow for the determination of the contribution of these different types of variation to intelligibility and other perceptual measures (e.g., Saito et al., 2023). The edit distance metrics used here also did not incorporate information smaller than a phoneme. That is, we used broad transcription with only diacritics that were essential for the scoring methods (i.e., the tie diacritic for diphthongs and the syllabic diacritic). The edit distance methods allow for the incorporation of any diacritic. These phonetic changes are given smaller penalties (e.g., 0.5 rather than 1.0 for a phoneme change in the heuristic model). The task of fully narrowly transcribing a large set of recordings would add substantially to the time and can decrease interrater reliability (Shriberg & Lof, 1991). A follow-up study from this one could investigate whether adding diacritics for a subset of the data used here increases the amount of intelligibility variability accounted for.

When averaged across listening conditions and for the two talkers representing each accent, there were relations between edit distance scores and intelligibility for seven of the eight accents. However, the ways in which the Levenshtein scores were related to intelligibility scores differed across talkers when separated by listening condition in ways that were not clearly tied to accent variety. There were several distinct patterns. In one pattern, listeners showed decrements in intelligibility with increasing distance scores across all three listening conditions. This pattern was clearly observed for the both Hindi talkers, the Japanese female talker, the Scottish female talker, and the Spanish female talker. In contrast, other talkers showed patterns in which listeners appeared to be relatively resistant to increasing edit distances for the easier listening conditions with steeper slopes for the hardest listening condition (e.g., British female, French female, Mandarin male). The remaining talkers showed a variety of other patterns including ones in which there did not appear to be a strong impact of edit distance scores (e.g., French male) or even some in which there was a positive relation between increasing edit distances and intelligibility scores for some listening conditions (Japanese male and Scottish male). These results suggest that although overall edit distances are significant contributors to intelligibility, the scores will need to be considered in terms of the listening conditions and other features of the talker's production patterns. That is, the intelligibility of a particular sentence by a particular talker cannot be estimated purely from their edit distance score, but likely will need to take into account factors such as rhythm, stress, speaking rate, and intonation as well. These patterns across edit distance scores and intelligibility did not clearly pattern with specific accent varieties nor whether the talker was an L1, L2, or bilingual speaker.

Even with multiple edit distance metrics that account for multiple levels of phonetic and phonological variation, there is still the need to account for many other linguistic factors that could interact with pronunciation differences to impact intelligibility. For example, some pronunciation differences may result in lexical confusions whereas others would not. For example, a vowel substitution of [oʊ] for /ɑ/ in the word "ball" results in a different real word whereas the same substitution in "shot" results in a nonword. If this substitution is in the sentence "the puppy played with the ball," the sentence level semantics also do not necessarily help the listener with mapping the production to the appropriate word since playing with a bowl, while perhaps less frequent, is certainly plausible. Conversely real-world knowledge and plausibility could help listeners overcome some types of phonetic differences in certain items. None of the metrics described above account for these types of semantic factors that certainly can contribute to intelligibility differences across items. That said, these scores may be used at the talker level or sentence level to help select stimuli for experiments where specific intelligibility levels are targeted.

In addition to relating intelligibility to the edit distance metrics, which was our central focus in the paper, we also conducted an analysis of the types of production differences found in our dataset, separating the differences by the types that are typically included in edit distance measures (e.g., substitutions, insertions, and deletions). This preliminary analysis showed that vowel and consonant substitutions were by far the most common types of differences from the local variety. This broad pattern held for nearly all talkers. However, whether vowel or consonant substitutions were more prevalent as well as the prevalence of the other types of production differences varied substantially across talkers. The production differences for L1 talkers tended to be primarily in vowels, which is consistent with literature showing substantial vowel differences for regional varieties within and across countries (Clopper et al., 2005; Blackwood Ximenes et al., 2017). In contrast, the L2 speakers tended to have more production differences in consonants. Our preliminary analysis correlating these production difference patterns to intelligibility suggests a slightly tighter negative relationship between consonant than vowel production differences and intelligibility. This pattern is consistent with research showing the prioritization of consonants over vowels during lexical processing in tasks such as lexical decision (Nazzi & Cutler, 2019). Because L1 talkers are less likely to display consonant substitutions across varieties, perhaps that specifically contributes to listeners' maintenance of relatively high levels of intelligibility at various levels of noise for L1 talkers relative to L2 talkers. Furthermore, listeners may have more experience with linking multiple vowel variants to specific lexical items and therefore are more adept at recognizing words that include vowel differences from the local standard.

This analysis also allowed for a preliminary comparison of the production patterns between the two talkers representing each accent. For some accents, the two talkers representing the variety (British, French, German, Japanese, Scottish) had very similar production difference patterns, but in the other cases, each of the two talkers representing the accent variety (Hindi, Mandarin, Spanish) were quite distinct in their production difference patterns. Of course, these analyses are only one way of capturing pronunciation differences across talkers and accents. Future work could continue these investigations to provide even more detailed views of production patterns and associated intelligibility patterns. For example, analyses could be conducted on error patterns at the word position level as there is some evidence that pronunciation differences or errors impact intelligibility differentially depending on the position in the word (e.g., Bent et al., 2007; Kim & Gurevich, 2023). Other analyses could investigate patterns by phoneme, specific substitution pattern, or type of word within which the error occurred (e.g., function vs. content word). These types of analyses should build on the general findings here regarding the relationship between broad phonetic distance and intelligibility as well as our initial investigation into pronunciation difference types.

We incorporated two talkers for each accent in an initial attempt to extract away from talker-level effects and focus on accent-level effects, but specific talkers selected here to represent each accent did not all show consistent patterns in either their relations between edit distances and intelligibility nor their difference pattern distribution profiles. Even larger datasets and talker samples will be needed to delineate among the many factors that contribute to intelligibility across accents, including the incorporation of multiple talkers of the same gender from each accent background. Further, our listeners were from a relatively homogeneous sample of monolingual American English speakers. Changing the characteristics of the listeners could also impact the relations among accent, edit distances, and intelligibility outcomes. Future studies should also include other social or sociolinguistic factors, such as listener familiarity with accents or language attitudes, to determine their impact on intelligibility across accent varieties.

Previous studies that have employed accent distance metrics have primarily focused on rating data (e.g., accent strength judgements) with fewer investigating intelligibility as we have done here. The utility of these metrics for predicting other types of perceptual effects should also be investigated. That is, intelligibility and accent strength ratings are just two measures of the types of information listeners extract from speech (Baese-Berk et al., 2023). Future work should also assess how pronunciation distance may impact other aspects of perceptual processing, such as listening effort and memory. For example, there is evidence that even fully intelligible L2 speech can still incur a processing cost above that of more familiar L1 varieties (McLaughlin & Van Engen, 2020). It is possible that the relation between edit distance measures and listening effort could be even stronger than between edit distances and intelligibility. That is, a sentence that has a higher edit distance score but for which listeners can successfully recover the linguistic message may still require more effort than one in which the edit distance score is closer to zero (i.e., aligns more closely with the local variety).

In addition to evaluating two edit distance metrics, we also investigated how the method for scoring listeners' responses would impact results. There was a strong, positive correlation between intelligibility scores based on percent words correct (PWC) compared with the token sort ratio (TSR) scores from Bosker (2021). Overall, there were few differences in the results when the data were scored using the two methods. That is, the same significant main effects and interactions were observed with the two scoring methods. Both models also accounted for roughly the same amount of variance. Therefore, the very substantial time savings with the TSR method suggests that adopting this scoring methodology will likely not have very substantial impacts on the main results from a study. Particularly as we are moving as a field toward large datasets, tools such as these will be essential. These tools should continue to be evaluated, however, with different types of speech materials and listeners to ensure that the results continue to align with the more traditional, “gold standard” measures of intelligibility.

5. Conclusion

This study supports the use of edit distance metrics to capture how differences in segmental productions across both unfamiliar L1 and L2 accents can impact intelligibility. The results suggest that relatively simple metrics (i.e., automatically calculated scores that are not adjusted for frequency of pronunciations within a large corpus) may be sufficient to capture the impact of phonetic differences across accent varieties on intelligibility. Furthermore, the findings support the use of an automated method (i.e., the Token Sort Ratio from Bosker [2021]) for scoring orthographic transcriptions that are commonly used in intelligibility studies. Although the pronunciation distance metrics employed here significantly predicted intelligibility across a range of accents and several listening conditions, a full accounting of intelligibility across accents will also require additional metrics at other levels of linguistics structure including phonological, syntactic, and semantic variables.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Tessa Bent: Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Malachi Henry:** Writing – review & editing, Writing – original draft, Validation, Formal analysis. **Rachael F. Holtb:** Writing – review & editing, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Holly Lind-Combs:** Writing – review & editing, Software, Formal analysis.

Acknowledgements

We would like to thank our research assistants, Jessica Bell, Payton Bastie, Journie Dickerson, Audrey Kunath, and Alondra Rodriguez, programming assistance from Jill Henry, as well as our funding from the National Science Foundation (Award Numbers: 1941691 and 1941662).

Appendix A. Summary results of the selected full mixed-effects beta regression model

Characteristic	Factor/Level	*Beta*	95% *CI*	*p*-value
HeuristicDistance	—	0.87	0.46, 1.3	<0.001
SNR	Quiet	—	—	—
	PlusFour	0.02	−0.04, 0.07	0.5
	Zero	−0.03	−0.09, 0.03	0.3
Accent	Midland	—	—	—
	British	−0.02	−0.13, 0.10	0.8
	German	−0.09	−0.21, 0.04	0.2
	Scottish	−0.01	−0.13, 0.11	0.8
	French	0.15	−0.03, 0.33	0.11
	Mandarin	−0.07	−0.19, 0.06	0.3
	Spanish	−0.29	−0.50, −0.08	0.007
	Hindi	0.42	0.28, 0.57	<0.001
	Japanese	0.01	−0.17, 0.19	>0.9
HeuristicDistance by SNR	Heuristic-Distance * PlusFour	−0.31	−0.76, 0.15	0.2
	Heuristic-Distance * Zero	−0.15	−0.58, 0.28	0.5
Heuristic DistancebyAccent	Heuristic-Distance * Accent	—	—	—
	Heuristic-Distance * British	−0.68	−1.3, −0.06	0.032
	Heuristic-Distance * German	−0.89	−1.5, −0.27	0.005
	Heuristic-Distance * Scottish	−0.38	−1.1, 0.37	0.3
	Heuristic-Distance * French	−0.98	−1.6, −0.42	<0.001
	Heuristic-Distance * Mandarin	−0.66	−1.2, −0.12	0.016
	Heuristic-Distance * Spanish	−0.02	−0.55, 0.52	>0.9
	Heuristic-Distance * Hindi	−1.1	−1.6, −0.55	<0.001
	Heuristic-Distance * Japanese	—	−0.07, 0.21	—
SNR by Accent	PlusFour * British	0.07	−0.24, 0.04	0.3
	Zero * British	−0.1	−0.11, 0.21	0.2
	PlusFour * German	0.05	0.02, 0.33	0.5
	Zero * German	0.18	−0.23, 0.08	0.026
	PlusFour * Scottish	−0.07	0.34, 0.63	0.4
	Zero * Scottish	0.48	−0.30, 0.09	<0.001
	PlusFour * French	−0.11	0.07, 0.45	0.3
	Zero * French	0.26	−0.19, 0.12	0.007
	PlusFour * Mandarin	−0.04	0.00, 0.31	0.7
	Zero * Mandarin	0.16	−0.25, 0.20	0.05
	PlusFour * Spanish	−0.03	0.36, 0.79	0.8
	Zero * Spanish	0.57	−0.56, −0.27	<0.001
	PlusFour * Hindi	−0.42	0.28, 0.55	<0.001
	Zero * Hindi	0.42	−0.48, −0.12	<0.001
	PlusFour * Japanese	−0.3	0.31, 0.65	0.001
	Zero * Japanese	0.48	−1.0, 0.47	<0.001
Heuristic Distance by SNRbyAccent	Heuristic-Distance * PlusFour * British	−0.28	0.00, 1.5	0.5
	Heuristic-Distance * Zero * British	0.75	−0.46, 1.1	0.049
	Heuristic-Distance * PlusFour * German	0.34	−0.72, 0.83	0.4
	Heuristic-Distance * Zero * German	0.05	−1.4, 0.59	0.9
	Heuristic-Distance * PlusFour * Scottish	−0.39	−1.2, 0.58	0.4
	Heuristic-Distance * Zero * Scottish	−0.33	−0.63, 0.75	0.5
	Heuristic-Distance * PlusFour * French	0.06	−0.04, 1.3	0.9
	Heuristic-Distance * Zero * French	0.62	−0.62, 0.79	0.065
	Heuristic-Distance * PlusFour * Mandarin	0.08	−0.12, 1.2	0.8
	Heuristic-Distance * Zero * Mandarin	0.55	−0.71, 0.63	0.11
	Heuristic-Distance * PlusFour * Spanish	−0.04	−0.71, 0.56	>0.9
	Heuristic-Distance * Zero * Spanish	−0.07	−0.74, 0.59	0.8
	Heuristic-Distance * PlusFour * Hindi	−0.08	−0.10, 1.1	0.8
	Heuristic-Distance * Zero * Hindi	0.52	0.08, 0.13	0.1
	Heuristic-Distance * PlusFour * Japanese	—	0.15, 0.28	—
	Heuristic-Distance * Zero * Japanese	—	0.12, 0.24	—

Model Syntax: TSR Score ∼ UW Edit Distance*Accent*SNR + (1 | SNR) + (1 + Accent + UW Edit Distance | Sentence) ++ (1 + UW Edit Distance | Participant).

Note: Due to rank deficiency in the model matrix, the table does not include interactions between the Japanese accent and Edit Distance, nor the three-way interaction between the Japanese accent, Edit Distance, and SNR as this level of the interaction was dropped by the model due to redundancy.

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance, 35*(2), 520–529. https://doi.org/10.1037/a0013552.

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic & Physiological Optics: The Journal of the British College of Ophthalmic Opticians (Optometrists), 34*(5), 502–508. https://doi.org/10.1111/opo.12131.

Babel, M. (2022). Adaptation to social-linguistic associations in audio-visual speech. *Brain Sciences, 12*(7). https://doi.org/10.3390/brainsci12070845 7.

Baese-Berk, M. M., Levi, S. V., & Van Engen, K. J. (2023). Intelligibility as a measure of speech perception: Current approaches, challenges, and recommendations. *The Journal of the Acoustical Society of America, 153*(1), 68–76. https://doi.org/10.1121/10.0016806.

Bailey, D. J., Speights Atkins, M., Mishra, I., Li, S., Luan, Y., & Seals, C. (2022). An automated tool for comparing phonetic transcriptions. *Clinical Linguistics & Phonetics, 36*(6), 495–514.

Bamford, J., & Wilson, I. (1979). Methodological considerations and practical aspects of the BKB sentence lists. In J. Bench & J. Bamford (Eds.), *Speech-hearing Tests and the Spoken Language of Hearing-impaired Children* (pp. 148–187). Academic.

Bartelds, M., de Vries, W., Sanal, F., Richter, C., Liberman, M., & Wieling, M. (2022). Neural representations for modeling variation in speech. *Journal of Phonetics, 92*. https://doi.org/10.1016/j.wocn.2022.101137 101137.

Bartelds, M., Richter, C., Liberman, M., & Wieling, M. (2020). A new acoustic-based pronunciation distance measure. *Frontiers in Artificial Intelligence, 3*, 39.

Beechey, T. (2022). Is speech intelligibility what speech intelligibility tests test? *The Journal of the Acoustical Society of America, 152*(3), 1573–1585. https://doi.org/10.1121/10.0013896.

Beijering, K., Gooskens, C., & Heeringa, W. (2008). Predicting intelligibility and perceived linguistic distances by means of the Levenshtein algorithm. *Linguistics in the Netherlands, 25*(1), 13–24.

Bent, T. (2014). Children's perception of foreign-accented words. *Journal of Child Language, 41*(6), 1334–1355. https://doi.org/10.1017/S0305000913000457.

Bent, T., Baese-Berk, M., Borrie, S. A., & McKee, M. (2016). Individual differences in the perception of regional, nonnative, and disordered speech varieties. *The Journal of the Acoustical Society of America, 140*(5), 3775–3786.

Bent, T., Bradlow, A. R., & Smith, B. L. (2007). Segmental errors in different word positions and their effects on intelligibility of non-native speech: All's well that begins well. In M. Munro & O.-.-S. Bohn (Eds.), *Language Experience in Second Language Speech Learning: In honor of James Emil Flege* (pp. 331–347). Amsterdam: John Benjamins.

Bent, T., & Holt, R. F. (2018). Shhh... I need quiet! Children's understanding of American, British, and Japanese-accented English speakers. *Language and Speech, 61*(4), 657–673. https://doi.org/10.1177/0023830918754598.

Bent, T. & Holt, R. F. (2023). Predicting intelligibility from pronunciation distance metrics. *International Congress of Phonetics Sciences*. Prague, Czech Republic.

Bent, T., Holt, R. F., Van Engen, K. J., Jamsek, I. A., Arzbecker, L. J., Liang, L., & Brown, E. (2021). How pronunciation distance impacts word recognition in children and adults. *Journal of the Acoustical Society of America, 150*(6), 4103–4117. https://doi.org/10.1121/10.0008930.

Bent, T., Lind-Combs, H., Holt, R. F., & Clopper, C. (2023). Perception of regional and nonnative accents: a comparison of museum laboratory and online data collection. *Linguistics Vanguard*. https://doi.org/10.1515/lingvan-2021-0157.

Blackwood Ximenes, A., Shaw, J. A., & Carignan, C. (2017). A comparison of acoustic and articulatory methods for analyzing vowel differences across dialects: data from American and Australian English. *The Journal of the Acoustical Society of America, 142*(1), 363–377. https://doi.org/10.1121/1.4991346.

Bosker, H. R. (2021). Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies. *Behavior Research Methods, 53*(5), 1945–1953. https://doi.org/10.3758/s13428-021-01542-4.

Case, J., Seyfarth, S., & Levi, S. V. (2018). Does implicit voice learning improve spoken language processing? implications for clinical practice. *Journal of Speech, Language, and Hearing Research, 61*(5), 1251–1260. https://doi.org/10.1044/2018_JSLHR-L-17-0298.

Clopper, C. G., & Bradlow, A. R. (2008). Perception of dialect variation in noise: Intelligibility and classification. *Language and Speech, 51*(3), 175–198. https://doi.org/10.1177/0023830908098539.

Clopper, C. G., Pisoni, D. B., & De Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical Society of America, 118*(3), 1661–1676. https://doi.org/10.1121/1.2000774.

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*, 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.

Floccia, C., Butler, J., Goslin, J., & Ellis, L. (2009). Regional and foreign accent processing in English: Can listeners adapt? *Journal of Psycholinguistic Research, 38*(4), 379–412. https://doi.org/10.1007/s10936-008-9097-8.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance, 32*(5), 1276–1293. https://doi.org/10.1037/0096-1523.32.5.1276.

Gooskens, C., & van Heuven, V. J. (2020). How well can intelligibility of closely related languages in Europe be predicted by linguistic and non-linguistic variables? *Linguistic Approaches to Bilingualism, 10*(3), 351–379. https://doi.org/10.1075/lab.17084.goo.

Gooskens, C., & Schneider, C. (2019). Linguistic and non-linguistic factors affecting intelligibility across closely related varieties in Pentecost Island. *Vanuatu. Dialectologia: Revista Electrònica*, 61–85.

Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language, 122*(2), 92–102. https://doi.org/10.1016/j.bandl.2012.04.017.

Hamann, A. H. and J. D. (2023). *systemfit: Estimating Systems of Simultaneous Equations* (1.1-30). https://cran.r-project.org/web/packages/systemfit/index.html.

Heald, S., & Nusbaum, H. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience, 8* https://www.frontiersin.org/articles/10.3389/fnsys.2014.00035.

Heeringa, W., Van Heuven, V., and Van de Velde, H. (2023). LED-A: Levenshtein Edit Distance App [computer program]. Retrieved 8 December 2023 from https://www.led-a.org/.

Huckvale, M. (2007). ACCDIST: An Accent Similarity Metric for Accent Recognition and Diagnosis. In C. Müller (Ed.), *Speaker Classification II: Selected Projects* (pp. 258–275). Springer. https://doi.org/10.1007/978-3-540-74122-0_20.

Jenkins, J. (2000). *The Phonology of English as an International Language*. OUP Oxford.

Jurado-Bravo, M. Á. (2021). Exploring the use of levenshtein distances to calculate the intelligibility of foreign-accented speech. *Cognitive Sociolinguistics Revisited, 48*, 153.

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

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

Kent, R. D., Miolo, G., & Bloedel, S. (1994). The intelligibility of children's speech: A review of evaluation procedures. *American Journal of Speech-Language Pathology, 3*, 81–95.

Kim, H., & Gurevich, N. (2023). Positional asymmetries in consonant production and intelligibility in dysarthric speech. *Clinical Linguistics & Phonetics, 37*(2), 125–142. https://doi.org/10.1080/02699206.2021.2019312.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122* (2), 148–203. https://doi.org/10.1037/a0038695.

Levi, S. V. (2015). Talker familiarity and spoken word recognition in school-age children. *Journal of Child Language, 42*(4), 843–872. https://doi.org/10.1017/S0305000914000506.

Levy, H., Konieczny, L., & Hanulíková, A. (2019). Processing of unfamiliar accents in monolingual and bilingual children: Effects of type and amount of accent experience. *Journal of Child Language, 46*(2), 368–392.

Lind-Combs, H. C., Bent, T., Holt, R. F., Clopper, C. G., & Brown, E. (2023). Comparing Levenshtein distance and dynamic time warping in predicting listeners' judgments of accent distance. *Speech Communication, 155*. https://doi.org/10.1016/j.specom.2023.102987 102987.

List, J.M. (2012). Multiple sequence alignment in historical linguistics. In Proceedings of ConSOLE (Vol. 19, pp. 241-260).

List, J.M. & Forkel, R. (2024). LingPy. A Python library for historical linguistics. Version 2.6.13. URL: https://lingpy.org, DOI: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy. With contributions by Simon Greenhill, Tiago Tresoldi, Christoph Rzymski, Gereon Kaiping, Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel, Patrick Elmer, Arne Rubehn. Passau: University of Passau.

Martinez, A. M. C., Spille, C., Roßbach, J., Kollmeier, B., & Meyer, B. T. (2022). Prediction of speech intelligibility with DNN-based performance measures. *Computer Speech & Language, 74* 101329.

Mazerolle, M. J. (2023). *AICcmodavg: Model Selection and Multimodel Inference Based on (Q)AIC(c)* (2.3-2). https://cran.r-project.org/web/packages/AICcmodavg/index.html.

McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupil response for accurately recognized accented speech. *The Journal of the Acoustical Society of America, 147* (2), EL151–EL156. https://doi.org/10.1121/10.0000718.

Munro, M. J. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition, 20*, 139–154.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of 2nd-language learners. *Language Learning, 45*(1), 73–97.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*, 285–310. https://doi.org/10.1111/0023-8333.49.S1.8.

Nagle, C. L., & Huensch, A. (2020). Expanding the scope of L2 intelligibility research: Intelligibility, comprehensibility, and accentedness in L2 Spanish. *Journal of Second Language Pronunciation, 6*(3), 329–351. https://doi.org/10.1075/jslp.20009.nag.

Nagle, C. L., Huensch, A., & Zárate-Sández, G. (2023). Exploring phonetic predictors of intelligibility, comprehensibility, and foreign accent in L2 Spanish speech. *The Modern Language Journal, 107*(1), 202–221. https://doi.org/10.1111/modl.12827.

Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics. Compass, 3* (1), 175–198. https://doi.org/10.1111/j.1749-818X.2008.00114.x.

Nilsson, M., Soli, S. D., & Gelnett, D. J. (1996). *Development of the Hearing in Noise Test for Children (HINT-C)*. House Ear Institute.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America, 95*(2), 1085–1099.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Hochenberger, R., Sogo, H., Kastman, E., & Lindelov, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics, 2*(2), 33–52. https://doi.org/10.1146/annurev-linguist-030514-125050.

Pinet, M., Iverson, P., & Huckvale, M. (2011). Second-language experience and speech-in-noise recognition: Effects of talker-listener accent similarity. *Journal of the Acoustical Society of America, 130*(3), 1653–1662. https://doi.org/10.1121/1.3613698.

Pongkittiphan, T., Minematsu, N., Makino, T., Saito, D., & Hirose, K. (2015). Automatic prediction of intelligibility of English words spoken with Japanese accents-comparative study of features and models used for prediction. In SLaTE (pp. 19-22).

R Core Team. (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/.

Rogers, C. L., Dalby, J., & Nishi, K. (2004). Effects of noise and proficiency on intelligibility of Chinese-accented English. *Language and Speech, 47*, 139–154. https://doi.org/10.1177/00238309040470020201.

Saito, K., Macmillan, K., Kachlicka, M., Kunihara, T., & Minematsu, N. (2023). Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Studies in Second Language Acquisition, 45*(1), 234–263. https://doi.org/10.1017/S0272263122000080.

San, N., Paraskevopoulos, G., Arora, A., He, X., Kaur, P., Adams, O., & Jurafsky, D. (2024). Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. In Proceedings of SIGTYP2024, pages 100-112.

Sanders, N. C., & Chin, S. B. (2009). Phonological distance measures. *Journal of Quantitative Linguistics, 16*(1), 96–114. https://doi.org/10.1080/09296170802514138.

Seifert, M., Morgan, L., Gibbin, S., & Wren, Y. (2020). An alternative approach to measuring reliability of transcription in Children's speech samples: Extending the concept of near functional equivalence. *Folia Phoniatrica Et Logopaedica, 72*(2), 84–91. https://doi.org/10.1159/000502324.

Sereno, J., Lammers, L., & Jongman, A. (2016). The relative contribution of segments and intonation to the perception of foreign-accented speech. *Applied Psycholinguistics, 37*(2), 303–322. https://doi.org/10.1017/S0142716414000575.

Shi, T., Kasahara, S., Pongkittiphan, T., Minematsu, N., Saito, D., & Hirose, K. (2015). A measure of phonetic similarity to quantify pronunciation variation by using ASR technology. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. ISBN 978-0-85261-941-4. Paper number 0432.1-5 retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0432.pdf.

Shriberg, L. D., & Lof, G. L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics & Phonetics, 5*(3), 225–279. https://doi.org/10.3109/02699209108986113.

Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., Loiselle, L. M., Oakes, T., & Cook, S. (2012). Development and validation of the AzBio sentence lists. *Ear and Hearing, 33*(1), 112–117. https://doi.org/10.1097/AUD.0b013e31822c2549.

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics, 25*(1), 1–24. https://doi.org/10.1006/Jpho.1996.0031.

Van Engen, K. J., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2014). Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. *Journal of Speech, Language, and Hearing Research, 57*(5), 1908–1918.

Wieling, M., & Nerbonne, J. (2015). Advances in dialectometry. *Annual Review of Linguistics, 1*(1), 243–264. https://doi.org/10.1146/annurev-linguist-030514-124930.

Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., & Nerbonne, J. (2014a). Measuring foreign accent strength in English. *Language Dynamics and Change, 4*(2), 253–269.

Wieling, M., Margaretha, E., & Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics, 40*(2), 307–314.

Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., & Baayen, R. H. (2014b). A cognitively grounded measure of pronunciation distance. *PloS One, 9*(1), e75734.

Wilson, E. O., & Spaulding, T. J. (2010). Effects of noise and speech intelligibility on listener comprehension and processing time of Korean-accented English. *Journal of Speech Language and Hearing Research, 53*(6), 1543–1554. https://doi.org/10.1044/1092-4388(2010/09-0100).

Winters, S., & O'Brien, M. G. (2013). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication, 55*(3), 486–507. https://doi.org/10.1016/j.specom.2012.12.006.

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.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2.

Yoho, S. E., & Borrie, S. A. (2018). Combining degradations: The effect of background noise on intelligibility of disordered speech. *J Acoust Soc Am, 143*(1), 281. https://doi.org/10.1121/1.5021254.