Speech Communication 155 (2023) 102987

Contents lists available at ScienceDirect

Speech Communication

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

Comparing Levenshtein distance and dynamic time warping in predicting listeners' judgments of accent distance☆

Holly C. Lind-Combs [a,*], Tessa Bent [b], Rachael F. Holt [a], Cynthia G. Clopper [c], Emma Brown [b]

[a] Department of Speech and Hearing Science, The Ohio State University, United States
[b] Department of Speech, Language and Hearing Sciences, Indiana University, United States
[c] Department of Linguistics, The Ohio State University, United States

ARTICLE INFO

Keywords:
Perceptual accent rankings
Dynamic time warping
Levenshtein distance

ABSTRACT

Listeners attend to variation in segmental and prosodic cues when judging accent strength. The relative contributions of these cues to perceptions of accentedness in English remains open for investigation, although objective accent distance measures (such as Levenshtein distance) appear to be reliable tools for predicting perceptual distance. Levenshtein distance, however, only accounts for phonemic information in the signal. The purpose of the current study was to examine the relative contributions of phonemic (Levenshtein) and holistic acoustic (dynamic time warping) distances from the local accent to listeners' accent rankings for nine non-local native and nonnative accents. Listeners (n = 52) ranked talkers on perceived distance from the local accent (Midland American English) using a ladder task for three sentence-length stimuli. Phonemic and holistic acoustic distances between Midland American English and the other accents were quantified using both weighted and unweighted Levenshtein distance measures, and dynamic time warping (DTW). Results reveal that all three metrics contribute to perceived accent distance, with the weighted Levenshtein slightly outperforming the other measures. Moreover, the relative contribution of phonemic and holistic acoustic cues was driven by the speaker's accent. Both nonnative and non-local native accents were included in this study, and the benefits of considering both of these accent groups in studying phonemic and acoustic cues used by listeners is discussed.

1. Introduction

Speech signals provide indexical information about a speaker's gender, race, age, region of origin, and native language status, among others (Abercrombie, 1967; Bent and Holt, 2017). One indexical dimension to which listeners are highly sensitive is a speaker's status as a native or nonnative speaker of a language (Flege, 1984; Park, 2013). Listeners can recognize speakers as native or nonnative with samples as short as 30 ms (Flege, 1984), in monosyllables (Park, 2013), in languages to which listeners have no prior exposure (Major, 2007), and for stimuli that have been substantially altered (e.g., played backwards; Munro et al., 2010). Listeners are also highly sensitive to the strength of accents (e.g., Flege et al., 1995). Factors such as speaking rate (Bent et al., 2016) and type and number of segmental deviations from the listener's local accent (e.g., substitutions, adding and omitting phonemes; Gooskens and Heeringa, 2004) affect judgments of accent strength. Correlations between objective accent distance measures and perceptual judgments give insight into how these various cues influence perceived "accentedness" (Wieling et al., 2014a). However, studies that have attempted to reconcile objective distance measures with subjective listener perceptions have yielded inconsistent results (e.g., Anderson-Hsieh et al., 1992; Sereno et al., 2016, Gooskens and Heeringa, 2004), particularly across both non-local native (i.e., other dialects within the same language but distinct from the listener's own dialect) and nonnative accents (e.g., Bent et al., 2016, 2021). The present study assessed the relation between perceptual judgments of accent distance and both acoustic and phonemic measures of accent distance across native and nonnative accents.

☆ Author note: Statistical and experimental code, stimuli, and data are openly available at the project's Open Science Framework page https://osf.io/cnxgt/ within the "Perceptual Acoustic Distance Project" folder. Correspondence concerning this article should be addressed to Holly Lind-Combs, Department of Speech and Hearing Sciences, The Ohio State University, 110 Pressey Hall, 1070 Carmack Rd, Columbus, OH, 43,210, USA.

* Corresponding author.
E-mail address: lind-combs.1@buckeyemail.osu.edu (H.C. Lind-Combs).

https://doi.org/10.1016/j.specom.2023.102987
Received 11 October 2022; Received in revised form 10 August 2023; Accepted 20 September 2023
Available online 21 September 2023

1.1. Perceptual accent judgments

Perception of accentedness has been quantified in numerous ways, including rating scales, ladder tasks, and free classification. Likert scales – a type of rating scale – quantify listeners' judgments on a numeric scale with researcher-established qualitative labels (Munro and Derwing, 1995). Accent strength (Anderson-Hsieh et al., 1992; Riney et al., 2005), similarity to a listener's own accent (Gooskens and Heeringa, 2004), and native-likeness (Bartelds et al., 2020; Wieling et al., 2014b) have all been measured using Likert scales. Another less common rating scale that has been used in accent perception studies is a magnitude estimation scale (Brennan et al., 1975; Southwood and Flege, 1999), in which listeners are asked to rate accents compared to reference stimuli on perceived differences of magnitude. Southwood and Flege (1999) found that magnitude estimation – and not linear measurement – leaves ample room for response bias, as listeners seem to consistently compare the stimuli to the one previously heard (instead of to the reference stimulus they are provided).

The ladder task, introduced by Bradlow et al. (2010), is a more recently developed tool in which listeners organize speakers on a ladder structure, from closest to a dialect/language (usually the listener's native dialect/language) at the base, to farthest at the top rung. An advantage of this task is that the experimenter may establish a baseline (e.g., proximity to the local accent) without a label for the upper extreme. This feature is particularly useful for tasks in which both native and nonnative talkers are included, such that there is not an intuitive "end" to the scale (i.e., there is no objective farthest distance). An additional advantage is that the listener is provided with access to all stimuli throughout the task and can listen to each item multiple times with no restrictions on the order in which they are heard, which is helpful when drawing comparisons among numerous stimuli. Similar advantages are seen with the free classification task, in which participants group talkers in a two-dimensional matrix based on perceptual similarities (Clopper, 2008). However, the free classification task does not provide any sort of examiner-set baseline or predetermined number of groups.

Bent et al. (2016) compared the utility of the ladder task (in which listeners were asked to rank accents based on perceptual distance from Standard American English) to a free classification task (in which listeners were asked to group speakers by perceived region of origin). The ladder task was scored based on the average distance accents were placed from the baseline; in other words, how many "ladder rungs" away from Standard American English the accent was placed. Scoring of the free classification task used additive clustering analysis of a similarity matrix, in which each matrix cell represented the number of times specific pairs of accents were grouped together. The authors found complementary results between the tasks, in that listeners' sensitivity to nonnative accents was observed in both tasks, and listeners perceptually distinguished between non-local native and nonnative accents.

Both the ladder and free classification tasks allow researchers to identify how listeners organize stimuli, potentially revealing specific characteristics of the speech to which listeners are attending. While perceived distance could be extracted from a free classification task, this information is more implicit in participant responses, whereas indicating distance is explicitly required in completion of a ladder task. The current study utilized the ladder task, as the research question aimed to compare perceived distance with established objective distance measures.

1.2. Objective distance measures

Objective distance measures provide estimates of distance from a reference accent. The resulting measurements can then be compared to subjective accent distance judgments, with the goal of identifying the segmental or acoustic aspects of accented speech that contribute to perceptual judgments. Many researchers have begun using these objective distance measures in recent years (e.g., Gooskens and Heeringa, 2004; Wieling et al., 2014a, 2014b; Bartelds et al., 2020), but the nature of the relationship between objective and perceptual measures is complex and ripe for exploration.

1.2.1. Segmental measures

The Levenshtein Distance algorithm (Levenshtein, 1966) measures the distance between two sequences (such as phonemes). Its application to dialect distances began with Kessler's (1995) measurements of distance among Irish dialects. This measure finds an optimal phonemic alignment between target and reference stimuli and assigns an equal penalty to any phonemic deviation (i.e., substitution, omission, insertion of a phoneme) from the reference speaker (Kessler, 1995). In other words, the production of a word by one speaker is compared against the phonemic realization of that word by a comparison speaker. For each phoneme that differs, one point is assigned; points are summed across the word, with greater summed values reflecting a greater distance between the two productions of the stimulus. Therefore, the larger the Levenshtein distance, the more the target utterance deviates phonemically from the same utterance as produced by the reference speaker. The original Levenshtein algorithm operated on a binary system in which any change, regardless of the type, resulted in a single point penalty. Slight adaptations were made to this original algorithm by Gooskens and Heeringa (2004) by dividing the summed score by the total possible number of phonemes, to account for word length.

The original Levenshtein algorithm has also been adapted to better reflect how certain phonemic changes impact listeners' perceptions (Bent et al., 2021; Levy et al., 2019). These adaptations are based on theoretical assumptions from the literature that some phonemic changes impact listener perception more than others. Vieru et al. (2011) revealed that certain phonemic changes differentially predict accent identification by French listeners (e.g., French listeners use the phonemic change of /b/ → /v/ as a cue to identify a native Spanish speaker of French). Furthermore, the relative importance of these phonemic substitutions for accent identification was impacted by the speaker's native language. Flege (1984) found no evidence that a nonnative accent is more detectable in vowels versus consonants, but more recent results from Gao (2019) suggest that consonant changes may be more impactful than vowel changes in perceptions of nonnative accented speech. The Levenshtein adaptation proposed by Levy et al. (2019) follows Gao's (2019) results, with consonant substitutions carrying a heavier penalty than vowel substitutions. Levy et al. (2019) used their Levenshtein distance measure to broadly characterize phonemic differences among three speakers, with one speaker each representing Standard German, non-local native (Palatinate German), and nonnative (Korean-accented German) accents. They found that the non-local talker had greater distances from Standard German compared to the nonnative talker and showed the lowest intelligibility of the three talkers. Both the adapted Levenshtein distance measure proposed by Levy et al. (2019) and the traditional Levenshtein measure have been shown to predict accent perception (e.g., Bent et al., 2021 and Gooskens and Heeringa, 2004, respectively).

Additional adaptations and alternatives to the traditional Levenshtein measure have been proposed. Wieling et al. (2014a, 2014b) tested cognitively-based extensions and alternatives to the original Levenshtein distance including the pointwise mutual information (PMI). PMI-based Levenshtein distance attempts to account for listeners' language exposure. Using corpus data, this process compared pronunciations of the same word from various accents of English using Levenshtein distances calculated from logical segmental alignments of words (i.e., only aligning vowels with vowels; Wieling et al., 2014b). This method calculates the distance between two segments, based on the relative frequency of their alignment (Wieling et al., 2014b). More frequently "misaligned" segments may have a lesser impact on listeners in the context of judging "native-likeness," as these are misalignments that listeners hear frequently. For example, if [ɑ] → [o] is a more

frequently occurring substitution than [ʊ] → ∅ (a deletion of [ʊ]), the PMI score for the first change would be lower than for the second. Wieling et al. (2014b) found significant negative correlations between the PMI-based Levenshtein distance scores and perceptual ratings of "native-likeness," consistent with the prediction that more frequently misaligned segments have a smaller impact on listener judgments.

Another cognitively-based measure utilized by Wieling et al. (2014a) involves naive discriminative learning (NDL), based on a theory from human and animal behavioral learning (Rescorla and Wagner, 1972) that suggests that learners make predictions based on available cues. Depending on the outcome of these predictions, associated connection strengths between the cues and the predicted outcome are adjusted to improve the accuracy of future predictions. In this case, listeners make predictions about word meaning based on the sounds they hear. The cognitive basis of applying this theory to dialect distances is that typically listeners are most often exposed to talkers who sound similar to themselves, and this exposure shapes the association strengths between input and outcomes (i.e., between phonetic cues and word meaning). Corpus data were analyzed to determine the frequency with which the association between phonetic cues and word meanings (i.e., outcomes) are likely to be encountered. Summing the association strengths between the cues and outcomes results in activations for the accented speech. A difference score was calculated between the activations of each speaker and the "average" speaker (of American English, in this case, averaged across speakers within the study) to create the NDL-based pronunciation distance. NDL correlates highly with both the traditional Levenshtein measure and perceptual ratings of accentedness (Wieling et al., 2014a).

A limitation of both the PMI and NDL measures is their reliance on large amounts of speaker data. Because both measures account for the frequency with which certain segmental alignments occur, large amounts of data are required to ensure these frequencies appropriately reflect the speech of speakers outside of the dataset. For example, 395 transcribed speech samples of the *Please Call Stella* passage from the Speech Accent Archive[1] (Weinberger and Kunath, 2011) were used in Wieling et al. (2014a, 2014b) studies, representing native speakers from 99 different languages, to measure the frequency of possible segmental changes. The value of PMI and NDL as adaptations to the traditional Levenshtein is that they are grounded in cognition, accounting for listeners' language experiences and exposure, rather than assuming that all phonemic changes are (perceptually) created equal. Although (to these authors' knowledge) the PMI-based Levenshtein has not been directly compared to the traditional Levenshtein measure, Wieling et al. (2014a) demonstrated a strong correlation between NDL and the traditional Levenshtein measure, providing support for a cognitive basis for the traditional Levenshtein. Further, the traditional Levenshtein has greater feasibility, and may be better suited for use in experimentation than NDL.

1.2.2. Acoustic measures

A significant limitation of all of the previously mentioned measures of phonemic distance is their inability to account for acoustic variation beyond the phoneme level that may contribute to accent distance perception (e.g., Gooskens, 2005). This sort of acoustic variation includes both subphonemic (e.g., reduction, lengthening, etc.) and prosodic (e.g., stress, intonation, rhythm, etc.) information. The role of prosody in predicting accent perception in English has yielded mixed results. Munro (1995) revealed that untrained native English listeners could identify native versus nonnative speaker status from unintelligible, low-pass filtered speech, relying only on rhythm, stress, and intonation instead of segmental information. Likewise, Anderson-Hsieh et al. (1992) reported a stronger correlation between overall pronunciation ratings and impressionistic ratings of nonnative speakers' prosodic deviance, than pronunciation ratings and segmental or syllable-level changes. On the other hand, Sereno et al. (2016) found that segmental changes influenced accentedness, comprehensibility, and intelligibility ratings, while prosody influenced intelligibility ratings only. Sereno et al. (2016) and Anderson-Hsieh et al. (1992) both investigated the relative contributions of prosodic and segmental cues to listeners' judgments. Differences in methodologies among these studies (e.g., impressionistic judgments of prosody: Anderson-Hsieh et al., 1992 and Munro, 1995; versus more objective measures of prosodic differences such as F0 contours: Sereno et al., 2016) could at least partially explain these discrepant findings.

In studies of accent perception with a non-English target language, durational cues (German: Kolly et al., 2017), intonation (Norwegian: Holm, 2008; Italian: Vitale et al., 2014), and spectral cues (Thai: Wayland, 1997) have all been shown to play a significant role in predicting accent judgments. The majority of these studies used acoustic manipulations of speech stimuli to identify the relative contributions of prosodic versus segmental changes to accent perception. For example, Holm (2008) reported a significant reduction in ratings of accentedness when intonational manipulations (F0 slope and direction) were applied to stimuli in Norwegian, and Vitale et al. (2014) transplanted yes/no question and declarative sentence F0 trajectories between native and non-native speakers of Italian, finding that native prosody with non-native segments yielded lower accentedness ratings than the reverse. On the other hand, Vieru et al. (2011) revealed that prosody (quantified by rhythmic measures including proportion of vocalic intervals, pairwise variability index, and word-final schwa duration) only modestly contributed to accent classification by French speakers, in comparison to segmental contributions. These studies demonstrate that in addition to differing methodologies and outcome measures affecting findings, the target language under investigation is likely a significant predictor of the extent to which prosody influences accent judgments.

Investigations of the influence of prosody on listeners' within-language accent judgments (i.e., identifying dialect regions) show similarly mixed results (e.g., English: Alcorn et al. (2020), van Bezooijen and Gooskens (1999); Norwegian: Gooskens (2005); and Dutch: van Bezooijen and Gooskens (1999)). Alcorn et al. (2020) observed that while listeners use prosodic cues to make judgments about English speakers' region of origin within the United States (e.g., Southern, Midwestern), there is a greater detriment to listeners' accuracy when segmental information is removed than when prosodic information is removed. Prosodic cues similarly play only a minor role in within-language dialect identification in both British English and Dutch (van Bezooijen and Gooskens, 1999). The authors found a minimal effect on accuracy when prosodic cues were removed (monotonized speech), yet a large impact on accuracy when segmental content was removed (low-pass filtered speech). In contrast, prosodic cues play an integral role in distinguishing among Norwegian dialects: native listeners had much more difficulty identifying Norwegian dialects from monotonized recordings (i.e., intonation information removed) than when listening to original recordings (Gooskens, 2005). It is likely that phonemic and prosodic information are intertwined and are uniquely important factors considered by listeners in accent judgments, although their relative importance in accent perception may be language- or dialect-dependent.

To account for both the phonemic and non-phonemic (sub-phonemic, prosody, etc.) aspects of speech, dynamic time warping (DTW) has recently been used in accent perception research to quantify the acoustic distance between two speech signals. Instead of focusing only on the phonemes produced, this procedure optimizes alignment between numerical feature representations (e.g., Mel frequency cepstral coefficients: Bartelds et al., 2020; fundamental frequency: Gao, 2019), allowing for additional acoustic information to be captured beyond the phoneme level. From this alignment, the shortest path through a cost

[1] The Speech Accent Archive is a digitally available archive of more than 1000 transcriptions and audio files of speakers from various language backgrounds reading the *Please Call Stella* passage. Some demographic information for each speaker is also available (e.g., native language, gender, age, etc.).

matrix is calculated, ultimately resulting in a distance score that captures pertinent acoustic information in the signal. This automated procedure does not rely on listener judgments of phoneme production (i.e., transcription) and DTW therefore could be a faster and more reliable alternative to Levenshtein distance measures, which require manual phonemic transcription, while also accounting for differences in sub-phonemic or prosodic aspects of the speech signals. The acoustic information captured by DTW includes phonemic, subphonemic, and prosodic information, and will therefore be referred to as a holistic acoustic distance measure.

Bartelds et al. (2020) compared DTW output scores (of Mel frequency cepstral coefficients) with perceptual ratings of native-likeness on a 7-point Likert scale and with segmental deviations (PMI-based Levenshtein distance) of nonnative English speakers from 99 different L1 backgrounds. Results from a multiple regression model including the PMI-based Levenshtein distance measure, DTW, and number of mispronunciations (manually counted), revealed that both the Levenshtein distances and DTW scores significantly predicted perceptual judgments, while number of mispronunciations did not. The DTW variable contributed 5 % of the model variance. Although this contribution is modest, Bartelds et al. (2020) suggested that DTW captures some aspects of the signal beyond what phonemic measures (e.g., Levenshtein) capture. An additional experiment from their study included DTW analyses of MFCCs generated from a single speaker's repeated token but with prosodic manipulations (i.e., normal pronunciation, rising intonation, lengthened first syllable) and recorded with two different devices simultaneously. The purpose of this final experiment from Bartelds et al. (2020) was to both highlight the role of prosody in MFCC calculations, and to show that different recording devices of the same stimuli from the same speaker can have different DTW values. The authors posit that MFCC's sensitivity to differences in recording devices and procedures could explain the less-than-optimal performance of the DTW measure. In their experiment predicting human perceptual ratings of native-likeness from the acoustic and phonemic measures, Bartelds et al. (2020) used recordings from the Speech Accent Archive (Weinberger and Kunath, 2011), which they report could have impacted their DTW results given the wide range of recording conditions and quality. Speakers record and upload audio files from their own personal devices to the Speech Accent Archive (Weinberger and Kunath, 2011), eliminating the possibility of consistency in recording protocols. This recording sensitivity issue could be mitigated by using consistent recording conditions across stimuli, such as recording all stimuli using high-quality equipment in a sound-attenuated booth. In sum, this study shows that DTW may capture important acoustic information that segmental measures alone cannot, but suggests that consistency in recording conditions could improve its reliability. More recently, Bartelds et al. (2022) demonstrated that a self-supervised Transformer-based neural model of acoustic distance significantly predicted human accent perception – outperforming the segmental measure; however, the generalizability of the results were limited by the fact that the model required language-specific training for success, which is both costly and time-intensive.

1.3. Current study

The purpose of the current study was to examine the relative contributions of phonemic and holistic acoustic distance from the local accent to listeners' accent rankings for multiple non-local native and nonnative accents. The study compared the unweighted (original Levenshtein measure) and weighted (the adapted measure from Levy et al., 2019) Levenshtein distance measures against one another and analyzed how these phonemic measures compare to a holistic acoustic distance measure (DTW) in predicting perceptual accent judgments. Bartelds et al. (2020) also used a similar multi-method approach, preliminarily revealing promising results. The current study aimed to build upon that work by using: (1) both non-local native and nonnative accents; (2) a ladder task to allow for explicit perceptual comparison among accents by the listeners; (3) multiple sentence stimuli per talker, including a condition in which each talker produced a different sentence; and (4) shorter stimuli (1 sentence). The inclusion of non-local native and nonnative accents in the same perceptual ranking task represents a unique contribution of the current study, as the majority of research focuses on only one group (i.e., either native or nonnative accents, compared to a reference stimulus; see, for example: Flege et al., 1995; Wieling et al., 2014a, 2014b; Bradlow et al., 2010). The use of an interval scale in the ladder task on which accents can be compared provides a perceptual measure of accent distance (as opposed to strength), which may serve as a better theoretical match to the objective phonemic and holistic acoustic distance measures under investigation in the current study, as opposed to other perceptual measures (e.g., free classification, Likert scales of accent strength). Using multiple sentence stimuli per talker, one of which is different for each talker, allows for the examination of performance consistency of the distance measures across sentences with varied content. The presence of an interaction between sentences and one or both of the acoustic distance measures would suggest a dependence on linguistic content, which could speak to the usefulness of the distance measures. Finally, limiting the stimuli to one sentence allows for an acoustic or phonetic "first impression." Many studies (including Bartelds et al., 2020) use the *Please Call Stella* passage from the Speech Accent Archive (Weinberger and Kunath, 2011). The advantage of using this passage is in its representation of common English words and nearly the full spectrum of English consonants and vowels. However, the length of the passage increases the cognitive load, requiring the listener to hold more linguistic content in their working memory when making their perceptual judgments (cf. Alcorn et al., 2020). The inclusion of stimuli that vary in both length and content from the Stella passage can provide a more robust understanding of the contribution of segmental and acoustic cues on accent judgments. Further, by using stimuli recorded in sound-attenuated booths with consistent procedures and high-quality equipment, we mitigate the negative effects of recording inconsistencies on DTW outcomes noted by Bartelds et al. (2020).

2. Method

2.1. Participants

Fifty-two adult monolingual speakers of American English ($M = 21.2$ years; 16 males, 36 females) participated in this study. One additional participant was tested but excluded for exceeding the criterion set for time spent abroad (i.e., under 10 months). Listeners were recruited from Indiana University and the surrounding Bloomington, Indiana, community. Participants demonstrated hearing within normal limits by passing a pure-tone hearing screening at 25 dB HL at 250 Hz and at 20 dB HL for octave frequencies between and including 500 and 8000 Hz (re: ANSI, 2004). A majority listed Indiana as their home state ($n = 35$) with the remaining from the following states: Illinois ($n = 6$), Missouri ($n = 2$), New Jersey ($n = 2$), Florida ($n = 1$), Michigan ($n = 1$), Mississippi ($n = 1$), New York ($n = 1$), Ohio ($n = 1$), Pennsylvania ($n = 1$), and West Virginia ($n = 1$). Listeners were paid $10 for their participation. All research was approved by the local Institutional Review Board.

2.2. Stimuli

Stimuli were produced by 37 adult talkers ($M = 28$ years; 17 males, 20 females) from three native English accents (Midland American, Standard Southern British,[2] Scottish), six nonnative accents (Japanese-, Mandarin-, Korean-, Spanish-, French-, and German-accented), and one bilingual accent (Hindi-English). All nonnative speakers reported having

[2] Standard Southern British English will be referred to as "British" throughout the manuscript.

lived in the United States for no more than 4 years at the time of the recordings. Nonnative accents and native dialects were chosen to represent a variety of accent variation and geographic locations. Each accent was produced by four speakers (2 male and 2 female), except for British English (2 male, 1 female) and Scottish English (2 female). Speaker gender was nearly balanced to preclude gender from acting as a potential confounding variable in perceptual distance from Standard American English. Each talker contributed three sentences, two of which were the same across all talkers ("The cow was milked every day" and "Father forgot the bread") and one of which was unique for each talker (see Appendix A). The first two sentences were selected because they included phonemes that represent a variety of English consonants and vowels. The unique sentence was included with the goal of reducing listeners' reliance on specific sentence content when determining accent distance. The stimuli were selected from sentences in the Hearing in Noise Test-Children (HINT-C; Nilsson et al., 1994). Speech samples were either recorded at Indiana University's Speech Perception Lab, Ohio State University's Developmental Speech Research Lab, or obtained through the SpeechBox, ALLSSTAR Corpus from Northwestern University's Department of Linguistics (Bradlow, n.d.). All recordings were made in sound-attenuated booths with digital recorders and high-quality microphones. The RMS amplitude of all samples was normalized in Praat (Boersma, 2001).

Familiarity ratings were obtained for all of the non-local native and nonnative accents used in the study, except for the reference accent (Midland American English). The Midland accent was assumed to be highly familiar to all listeners because it is the local variety of the testing location. Listeners provided ratings of the amount of interaction with speakers from the various accent backgrounds on a scale of 0–10 with 0 indicating "never interacted" and 10 indicating "interact with daily." Results from the familiarity ratings were averaged across participants. An ANOVA was run to assess differences among accents in degree of familiarity, with accent as the predictor variable and familiarity rating as the outcome variable. The ANOVA was significant, indicating that familiarity ratings for certain accents were higher, on average, than others, $F(8, 459) = 6.76$, $p < .001$. Spanish- and Mandarin-accented English had the highest degree of familiarity (2.6 and 2.4, respectively), and Japanese- and Scottish-accented English represented the least familiar accents to listeners (0.6 and 0.3, respectively). Although there was a significant difference in familiarity of accents, all of the average familiarity ratings were at or below 2.6 (out of 10), indicating a low degree of experience with these accents.

2.3. Levenshtein distance measures

2.3.1. Unweighted

The unweighted Levenshtein distance is based on the traditional algorithm (Levenshtein, 1966), and compares a target stimulus to a reference, assigning one point per change, regardless of type (i.e., substitutions, omissions, additions, etc.) across all consonants and vowels (Kessler, 1995). The total number of penalties are summed at the word level, and then divided by the total possible number of phonemes (Gooskens and Heeringa, 2004). The total possible number of phonemes is not the number of phonemes produced by the reference speaker, but rather comes from the most logical alignment between the target and reference stimuli. If a non-local/nonnative speaker inserts a phoneme into the word, this increases the total possible number of phonemes by one. For example, if a nonnative speaker inserts a schwa in a target word 'milked' (e.g., /mɪlkət/) and the reference speaker does not (e.g., /mɪlkt/), the total possible number of phonemes increases from five to six. See Table 1 for a visualization of this alignment, and calculation of the total possible number of phonemes. Word level scores are summed across the sentence and divided by the total number of words in the sentence.

Table 1
Calculating total possible phonemes from most logical alignment using the unweighted Levenshtein distance.

Reference speaker	m	ɪ	l	k	∅	t	Total possible phonemes = 6
Non-local/nonnative speaker	|	|	|	|	|	|	
	m	ɪ	l	k	ə	t	

2.3.2. Weighted

An adapted version of the Levenshtein distance algorithm provides for more score variation depending on the type of error (see Table 2). This scoring method was adapted for intelligibility tasks by Levy et al. (2019) and is based on the assumption that not all phonemic changes carry equal weight in perception. As with the unweighted Levenshtein, the most logical alignment was used when comparing the reference and target stimuli using the weighted Levenshtein (see Table 1). Penalties were assigned per phoneme deviation (substitution, omission, or addition) as shown in Table 2, and then summed for each word. For example, substitution of one consonant for another received a 0.75-point penalty, while one vowel for another received a 0.5-point penalty. Unlike the unweighted Levenshtein measure, word-level scores in the weighted Levenshtein are not divided by the total possible number of phonemes in the target word, but instead summed at the word-level. Word length itself is not accounted for in this version. However, the word-length change penalty accounts for *changes* to word length. Both the unweighted and weighted Levenshtein are averaged at the sentence level. See Table 3 for example scoring for both unweighted and weighted Levenshtein.

To complete this analysis, the two sentences from each of the 37 speakers (the four Midland and all non-local speakers) used in the same sentence conditions were phonemically transcribed. In addition, all sentences for the unique sentence condition (Appendix A) were transcribed for all 37 speakers. Two research assistants trained in the use of IPA conventions independently transcribed each sentence, using broad (i.e., phonemic) transcription using a consistent set of symbols (e.g., the diphthong in the word 'same' always transcribed as /eɪ/ and not /ɛj/, if produced in a native-like manner). The transcriptions were compared, and any discrepancies between transcriptions were resolved by a third research assistant. In the case of ambiguities that could not be resolved among the three research assistants, the second author [TB] resolved disagreements. Therefore, the final version of each transcription was agreed upon by 3 to 4 researchers.

Each sentence was manually divided at word boundaries and compared to the local (Midland American English) dialect on the word level. The alignment of the reference and target phonemes was completed manually by research assistants, using the most logical alignment. For example, as seen in Table 1, all of the corresponding phonemes produced by the local native and non-local native/nonnative speaker were aligned, with the inserted schwa from the nonnative speaker aligned with a corresponding empty position in the local native speaker's production. All alignments were reviewed by the first author [HLC] to ensure use of the most logical and optimal alignment. The

Table 2
Weighted Levenshtein penalties for phonemic changes.

Change	Point Penalty
Vowel substituted by another vowel	0.5
Consonant substituted by another consonant	0.75
Insertion	1.0
Change to length*	1/log10(max(length(word1), length(word2)))
Other (deletions, vowel to consonant, consonant to vowel, etc.)	0.4

Note. *word1 = number of phonemes in the non-local accent speaker's production; word2 = number of phonemes for the local accent production.

Table 3
Examples of sentence-level score calculations using the weighted and unweighted Levenshtein algorithm.

	Unweighted Levenshtein				Weighted Levenshtein			
Target sentence	*Father forgot the bread*				*Father forgot the bread*			
Midland accent	fɑðɚ	fɔɹgɑt	ðə	bɹɛd	fɑðɚ	fɔɹgɑt	ðə	bɹɛd
French-accented English	fæðɚ	fɔɹgɑt	də	bɹɛd	fæðɚ	fɔɹgɑt	də	bɹɛd
Penalties	1/4	0/6	1/2	0/4	.5	0	.75	0
Levenshtein Score	.25	0	.5	0 = 0.188	.5	0	.75	0 = 0.313

transcription for each non-Midland speaker was compared to the transcription of each of the four Midland American English speakers included in this study. When a phoneme in one of the non-Midland speakers did not match any of the Midland speakers, a penalty from Table 2 was applied. Phonemes that matched at least one of the Midland speakers were not assigned a penalty (see Table 4). Not all productions of the stimuli were phonemically realized identically among the Midland speakers, and comparison of each stimulus to all four Midland speakers allowed for some flexibility in scoring to account for natural variability in production. Also note in Table 4 that the production of the target word "slept" by the Korean speaker received four penalties: two for deleting phonemes, one for a consonant change, and one for a change in word length. Higher penalties indicate larger phonemic differences between the Midland and non-Midland productions. No penalties were issued for changes in stress. A substitution of /ə/ for /ɚ/ was treated as a vowel change. Productions that lengthened consonants from one word to the next (e.g., 'bus stopped' was transcribed as [bʌsːtɑpt] for one Midland speaker) were calculated as though there were two separate consonants. For both unweighted and weighted Levenshtein measures, Midland speakers served as the comparison, and therefore all Midland speakers had scores of zero (as they could not receive penalties for phonemically deviating from themselves).

2.4. Dynamic time warping

Dynamic time warping computes the shortest distance through a cost matrix, typically generated from two vectors. In speech recognition, these vectors come from feature representations of the signals being compared. Mel frequency cepstral coefficients (MFCCs) are often used as feature representations in audio signal comparison, because they provide a more reliable representation of the phonetic content of the signal in automatic speech recognition compared to other feature representations (e.g., linear frequency cepstrum, among others; Davis and Mermelstein, 1980). MFCCs were calculated for 74 sentences (37 talkers x 2 sentences) for the same sentence conditions and 37 sentences for the unique sentence conditions, such that each Midland and non-local talker's production of each of the three sentences was measured. The process for calculating MFCCs in the current study was guided by the process outlined in detail by Bartelds et al. (2020), with a key difference being that the MFCCs in the current study were calculated over the entire sentence for each stimulus, instead of for each individual word from a paragraph. In the current study, MFCC calculations and subsequent DTW analyses were performed using a Python script (written by author HLC and shared in the OSF repository), incorporating the 'mfcc' function in the 'librosa' package (McFee et al., 2015).

Table 4
Examples of transcriptions, penalties and overall score using the weighted Levenshtein measure.

Sentence	The	baby	slept	all	night	Average
Midland 1	ðə	beibi	slɛpt	ɔl	naɪt	
Midland 2	ðə	beɪbi	sləpt	ɔl	naɪt	
Midland 3	ðə	beɪbi	slɛpd	ɔl	naɪt	
Midland 4	ðə	beɪbi	slɛpt	ɔl	naɪt	
Korean	ðə	beɪbi	sɛv	ɑl	naɪt	
Score	0	0	(0.4 + 0.4 + 0.75 + (1/LOG10(MAX((3), (5)))) = 2.981	0.5	0	(2.981 + 0.5)/5 = 0.696

To calculate the MFCCs, the signal was separated into 25-ms overlapping frames with a 10-ms step size, with 12 cepstral coefficients capturing the overall power of the signal in that window and accounting for variation in the signal intensity. The "0th" coefficient was included as a representation of the overall energy of each 25-ms frame, resulting in a total of 13 coefficients per frame. The first- and second-order derivatives were calculated for each of the 12 cepstral coefficients and energy representations for each 10-ms step, to account for the temporal changes between frames. This resulted in 39 coefficients for each 10-ms step. The concatenation of the vectors of 39 coefficients over the entire sample resulted in the MFCC feature representation for each speech stimulus item. The final step in obtaining the MFCCs was to normalize the coefficients, by applying a z-transform to each vector of MFCCs per windowed frame. Bartelds et al. (2020) demonstrated the importance of normalization of MFCCs in DTW to reduce the "noise" in the signal (i.e., less relevant acoustic information). Correlation of the acoustic distance measure with human judgments of accent distance increased in magnitude from $r = -0.27$ to $r = -0.71$ when the normalization procedure was used in processing the stimuli in Bartelds et al. (2020).

DTW between MFCC vectors from the non-Midland (non-local native and nonnative) accents and the reference stimuli (Midland, the local accent) was performed to provide a holistic acoustic-based distance measure. DTW scores were calculated as the shortest distance through a cost matrix consisting of MFCCs between a non-local native or nonnative accent (target) and reference (Midland) stimulus, with a higher DTW score representing a greater deviation from the reference stimulus. Each non-Midland stimulus was compared to the four reference stimuli (i.e., two male and two female Midland speakers), resulting in four acoustic distance scores per sentence. The lowest of the four scores was selected to represent the acoustic distance between that target speaker and the reference Midland dialect, to most closely match the procedure for the Levenshtein comparisons. Because three different sentences were used (two consistent, one unique), each speaker had three associated distance scores. Out of 111 possible comparisons, 107 of the lowest scores were between same-gender talker pairs.

Midland speakers received non-zero DTW scores (unlike for the Levenshtein scores, where all Midland speakers received zeros), as the MFCCs for each Midland stimulus was compared to the other three Midland stimuli. Following the same procedure as the non-Midland stimuli, the lowest of the three scores were selected to represent the acoustic distance between that target Midland speaker and the other three Midland dialects for each sentence stimulus. The choice to compare the Midland speakers to one another for the DTW measure and not the Levenshtein measure was based on the fact that there were substantial differences among DTW scores, while the Levenshtein differences were negligible (i.e., there were only occasional differences, such as flap (/ɾ/) for /d/).

2.5. Procedure

Participants were tested in one 60-minute session. Prior to the onset of testing, participants completed the consent process, hearing screening, and language background questionnaire. The experiment, conducted in a sound-attenuated booth, consisted of three ladder tasks with one-minute breaks between ladders. The two ladders in which each talker produced the same sentence were counterbalanced across participants and the ladder in which each talker produced a unique sentence was presented last. Verbal instructions were given to participants prior to the start of the experiment and on-screen instructions were provided at the start of each ladder. The experiment was created with custom software written in Python and run on a MacMini. Stimuli were presented binaurally through Sennheiser HD280 Pro-headphones at an average RMS amplitude of 65 dB SPL.

Each ladder had 20 rungs, with space for up to four speakers on each rung. While the choice of four spaces on each rung is somewhat arbitrary, imposing a limit is necessary to ensure that listeners rank some accents above/below others, thereby forcing them to make decisions about accent distance, while still allowing them to indicate that some accents are equally distant from the reference. A set of 37 rectangular icons appeared to the left of the ladder (an example of a starting ladder is displayed in the upper panel of Fig. 1). The bottom-most rung of the ladder was labeled "Standard American English" (i.e., the lay term for the Midland American English accent used by the local population). Participants clicked on one of the talker icons to hear that sentence. Participants were instructed to rank the talkers on the ladder according to how similar the talker sounded to the local accent, with talkers whose productions sounded most like the local accent placed near the bottom and those furthest from the local accent at the top (an example of a completed ladder is displayed in the bottom panel of Fig. 1). No model of the target dialect was provided to participants; rather, their knowledge and interpretation of the local dialect served as the reference stimulus. Providing more spaces than talkers allowed for variety in participants' representation of distance.

Midland American English speaker stimuli were included in the ladder task. The purpose of including Midland speakers as an accent variety in this study was to allow for the perceptual ranking of both native and nonnative accents of varying familiarity to the listeners. Although this Midland accent did not serve as the baseline in the ladder task, including the Midland variety added to the gradient of perceptual distances. Inclusion of this highly familiar variety served to situate the listener on a scale of perceptual distances and avoid lumping all varieties into a homogenous "other" group.

2.6. Data analysis

Statistical analyses were performed using R version 4.0.4 (R Core Team, 2021). Data and R code are available in the OSF repository. A linear mixed effects model was run to predict perceptual distance

Fig. 1. Examples of an empty ladder (top panel) and completed sample ladder (bottom panel).

(ladder rankings) from DTW, weighted and unweighted Levenshtein distance, and sentence condition (i.e., the 3-level factor of the sentence variable). Continuous variables were standardized (z-transformed), both to support convergence of the mixed effects models, and to allow for direct comparison of these variables' contributions to explaining variability in the outcome variable. A step-down approach was used to identify the random effect structure for the linear mixed effects model, beginning with the maximal design-driven random effects, including: a by-listener intercept, and by-listener slopes for each distance measure, the sentence variable, and their interactions; and a by-speaker intercept, and by-speaker slope for the sentence variable. This maximally design-driven model was over-specified, as expected, and resulted in a singular fit which did not converge. The model was then stepped down incrementally, removing one random effect term at a time (beginning with the random effects with the highest correlations with other variables in the random effects structure) until convergence was achieved. The model which ultimately achieved convergence included a by-listener random intercept and slopes for sentence and weighted Levenshtein, and a by-speaker random intercept and sentence slope. The lmerTest package was used in R to obtain *p*-values for fixed effects in the mixed effects model, and the criterion for significance was set at $p < .05$.

The variance inflation factor (VIF) function in R was used to identify the potential presence of unacceptable (multi)collinearity among the predictor variables. Using the heuristic that a VIF score greater than 10 indicates unacceptable collinearity, it was determined that the two Levenshtein predictor variables shared sufficient variance to inhibit interpretability (unweighted Levenshtein: 14.11; weighted Levenshtein: 13.52; DTW: 2.21). To address this collinearity, two separate models were constructed: one model with the unweighted Levenshtein measure and DTW, and the other with the weighted Levenshtein and DTW. The two models were then compared to determine which pair of acoustic distance measures resulted in the best-fitting model. The random effects structure for these two models was determined using the same methods described above, beginning with the maximal design-driven random effects (by-listener intercept, and by-listener slopes for each distance measure, the sentence variable, and their interactions; and by-speaker intercept and slope for the sentence variable). This model did not achieve convergence, and the models were stepped down by incrementally removing the variable or interaction with the highest degree of correlation to others in the random effects structure. The structures that achieved convergence were the same for the two models: both by-listener and by-speaker random intercepts and sentence slopes.

3. Results

Fig. 2 displays mean scores for each accent variety on the perceptual ladder task and the three objective distance measures. For the ladder task, scores represent the average ladder ranking by all listeners across

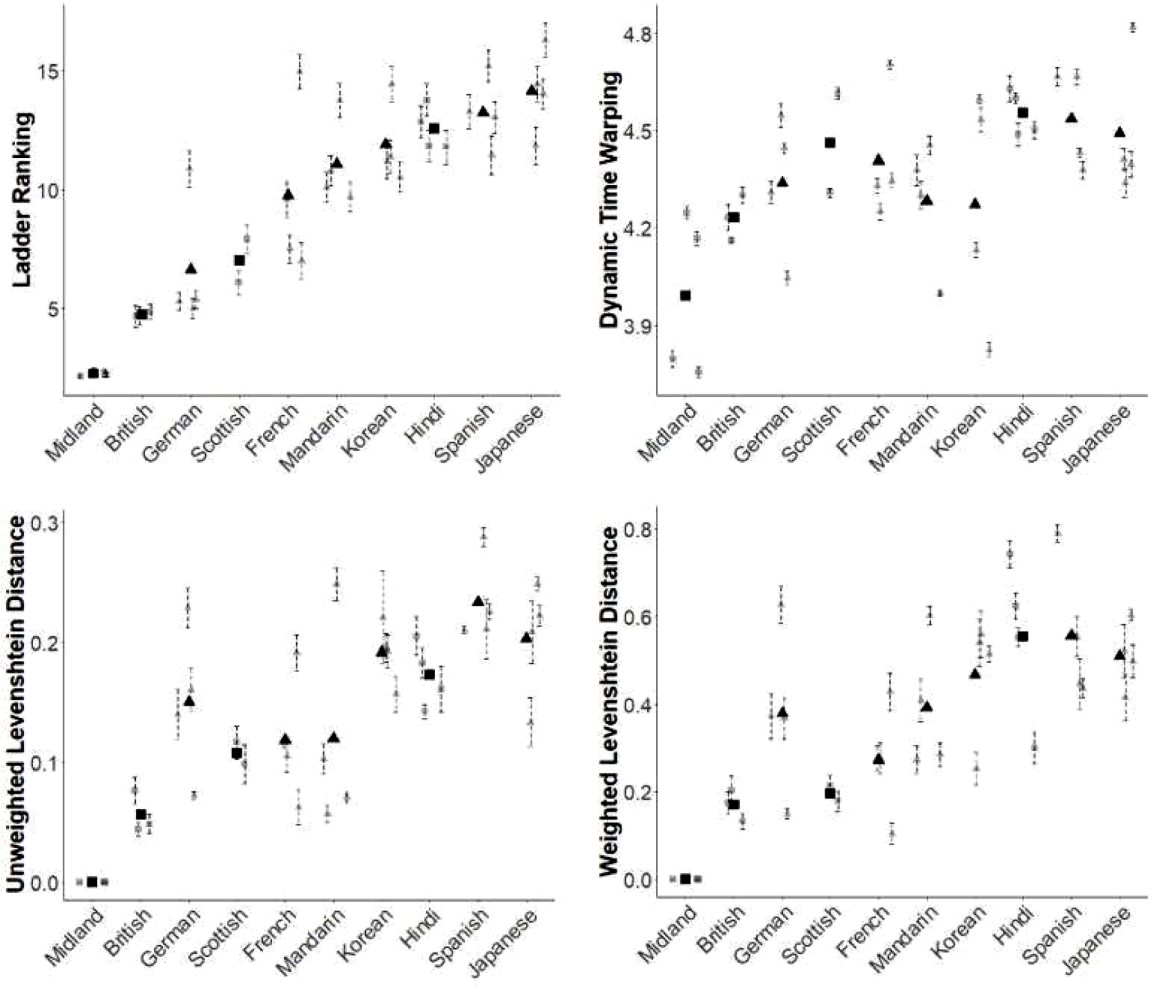

Fig. 2. Mean perceptual and objective accent distance measurements by accent for each speaker
Note. Vertical axes are measured in units unique to each distance measure. Horizontal axes represent accent groups. Means and error bars (+/- 1 standard error) are displayed for each speaker. Squares represent local and non-local native accents, and triangles represent nonnative accents. Large bold symbol for each accent represents the mean for all speakers of each accent; small gray symbols represent the average score per speaker.

all three sentences, per accent group. The DTW and Levenshtein distance scores represent the average score for each speaker across all three sentence stimuli. Lower scores on the ladder task represent greater perceived proximity to Midland American English. Similarly, lower scores on the Levenshtein and DTW distances represent greater proximity to the Midland speakers.

An inspection of Fig. 2 suggests a similar general pattern of accent distance across the four measures, with some variability present. An expected finding evident in the ladder ranking data (top left) is the tendency for increased variability among accent rankings for speakers with nonnative accents compared to non-local native ones. The Midland and British accents (and to a slightly lesser extent, the Scottish accent) show all of the speakers clustering around the group mean. The nonnative accents, on the other hand, tend to demonstrate more variable ladder rankings among speakers, with some well above or below the group mean. The error bars represent the within-speaker variability in ladder ranking for the three sentences, with the local and non-local native tending to demonstrate the least amount of variability in ladder rankings among the sentences. Thus, there is a trend for listeners, as a group, to rank the native accented speakers (both non-local and local) consistently with one another and show more variability in their rankings of the nonnative accents. Comparing among the objective measures of distance, the within-speaker variability tended to be small when measured using DTW; in other words, speakers' DTW scores tended to be relatively consistent across sentences. In contrast, both weighted and unweighted Levenshtein scores tended to be more variable across sentences for each speaker, as evidenced by the tendency for larger error bars around the speakers' mean scores.

3.1. Linear mixed effects models

Weighted and unweighted Levenshtein measures and DTW all independently predicted ladder task results (see Fig. 3). The simple correlations between the three distance measures and ladder rankings are all significant, though the relationship is stronger in both of the Levenshtein measures compared to the DTW measure, as visualized in Fig. 3. Results from the two linear mixed effects models are shown in Table 5. The sentence variable was not a significant predictor in either model, suggesting that the phonemic content of the sentences did not independently contribute to the ladder rankings by listeners. However, the interaction between sentence 2 and the unweighted Levenshtein measure in the first model was significant. Follow-up log-likelihood comparisons between nested models with and without the interaction between the sentence variable and unweighted Levenshtein distance revealed a significant interaction between the variables, $\chi^2(2) = 7.44$, $p = .024$. The interaction between the sentence variable and the weighted Levenshtein measure was not significant, nor was the interaction between DTW and sentence in either model. These findings suggest that the relative contribution of the unweighted Levenshtein measure to listeners' perceptual accent distance ratings may be at least partially dependent upon sentence content. In this case, the unweighted Levenshtein distance was a better fit to the ladder rankings for Sentence 1 than for Sentence 2.

Table 5
Results from two linear mixed effects models predicting ladder task results.

Model 1: Unweighted Levenshtein & DTW	Estimate	*t* value	*p* value
Intercept (Sentence 1)	9.24	16.74	< 0.001
Sentence 2	0.47	0.95	.347
Sentence 3	0.95	1.99	.051
Unweighted Levenshtein	1.34	3.27	.002
DTW	1.32	3.36	.002
Interaction: Unweighted Levenshtein by Sentence 2	−1.24	−2.16	.034
Interaction: Unweighted Levenshtein by Sentence 3	0.65	1.18	.245
Interaction: DTW by Sentence 2	−0.11	−0.22	.831
Interaction: DTW by Sentence 3	−0.45	−0.95	.347
Model 2: Weighted Levenshtein & DTW	**Estimate**	***t* value**	***p* value**
Intercept (Sentence 1)	9.38	17.87	< 0.001
Sentence 2	0.15	0.38	.745
Sentence 3	0.40	0.95	.347
Weighted Levenshtein	1.72	4.62	< 0.001
DTW	1.13	3.37	.002
Interaction: Weighted Levenshtein by Sentence 2	−0.64	−1.21	.230
Interaction: Weighted Levenshtein by Sentence 3	−0.51	−1.13	.263
Interaction: DTW by Sentence 2	−0.29	−0.59	.556
Interaction: DTW by Sentence 3	−0.31	−0.72	.478

The intercepts in both models 1 and 2 represent the group average ladder ranking for the first sentence. The coefficients for each predictor variable (unweighted Levenshtein and DTW in model 1, and weighted Levenshtein and DTW in model 2) represent the change in ladder ranking when the predictor variable increases by one standard deviation. The coefficients for both distance measures in each model were positive, indicating that, as expected, a 1-SD increase in the distance measure (i.e., getting farther from the reference Midland stimuli) results in an increase in ladder rankings. Although all of the distance measures were significant predictors of perceptual distance ratings, the unweighted and weighted Levenshtein's larger coefficient (as compared to DTW in each model) suggests that a 1-SD increase in this measure results in a larger impact on ladder rankings than a 1-SD increase in DTW. It is worth noting that the difference in coefficient size in model 1 between the unweighted Levenshtein and DTW predictors is minimal. Still, the weighted Levenshtein measure was a stronger predictor of perceptual

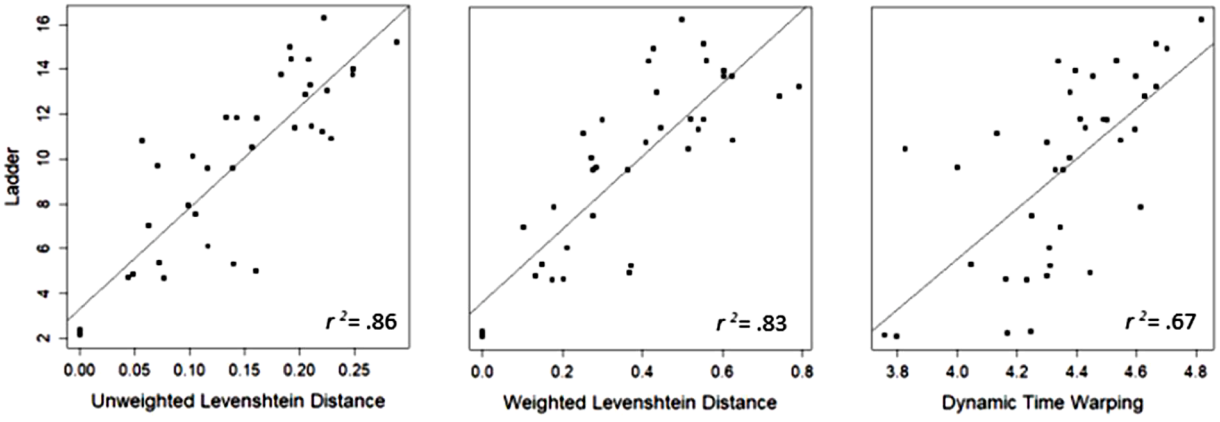

Fig. 3. Scatterplots displaying ladder task results correlated with unweighted Levenshtein distances (Left), weighted Levenshtein distances (Center), and dynamic time warping (Right)
Note. Individual points represent the average ladder rankings and distance values per speaker (averaged over the three sentences).

accent distance than DTW in model 2, indicating that the weighted Levenshtein measure was the strongest predictor of perceptual accent distance.

To address the research question of which measure serves as a better predictor of accent judgments, Akaike's Information Criterion (AIC) and log likelihood ratios for the two models were compared. Both AIC and log likelihoods are measures of model fit and can be used to compare models that are not nested, but that are modeled on the same data and for the same outcome variable (Akaike, 1974). AIC and log-likelihood values tend to covary, as they attempt to capture the same construct (that is, model fit), and models with lower AIC values tend to have higher log-likelihood ratios. As seen in Table 6, the model predicting ladder task results from the weighted Levenshtein measure and DTW had a lower AIC and higher, less-negative log-likelihood than the model using the unweighted Levenshtein measure and DTW as predictor variables. Although the difference between fit in these models is slight, the weighted Levenshtein and DTW is the best-fitting model given the data used in the present experiment.

4. Discussion

The purpose of the current study was to examine the relative contributions of phonemic (weighted and unweighted Levenshtein) and holistic acoustic (DTW) distances from the local accent to listeners' accent rankings for multiple non-local native and nonnative accents. Although these measures have been used in previous research, the current study is the first to compare their effectiveness in predicting accent judgments using a ladder task for short stimuli from both non-local native and nonnative talkers. Results from the current study suggest that the weighted Levenshtein distance measure is the strongest predictor of perceptual accent distance, although the differences in performance among the three distance measures were small.

4.1. Comparisons among the objective distance measures

4.1.1. Levenshtein distances

Both weighted and unweighted Levenshtein distances accounted for significant variability in accent rankings, with the weighted Levenshtein only slightly outperforming the unweighted. Given previous demonstrations of the impact of segmental deviations on perceptions of accent strength (Brennan et al., 1975; Derwing and Munro, 1997; Magen, 1998), the utility of Levenshtein measures as objective quantifications of segmental changes that can be consistently applied across studies is evident. The original, unweighted Levenshtein is highly correlated with perceptual judgments of accent distance (Gooskens and Heeringa, 2004; Wieling et al., 2014a). The weighted Levenshtein measure, introduced by Levy et al. (2019) was validated for use in predicting intelligibility by Bent et al. (2021). Although the weighted Levenshtein measure was designed to account for the assumed impact of varying types of pronunciation differences on intelligibility, there is limited evidence for a strong advantage of the weighted over the unweighted Levenshtein measure in this study. The current results indicate that the addition of deviation weights may not add enough explained variability in accent judgments over the unweighted Levenshtein measure to warrant its replacement. To our knowledge, no other studies have compared this weighted Levenshtein measure to the original, unweighted Levenshtein measure. However, Wieling et al. (2014a) compared the Native Discrimination Learning (NDL) Levenshtein measure – a cognitively based Levenshtein adaptation that takes into account listeners' familiarity with certain segmental deviations – to the traditional Levenshtein distance measure. The NDL Levenshtein adaptation and the original Levenshtein correlated very highly ($r = 0.89$). Therefore, the results from both Wieling et al. (2014a) and the current study suggest that the unweighted Levenshtein measure is likely sufficient for quantifying accent distance compared with various weighted measures for most purposes and situations.

Observing an advantage of the weighted Levenshtein measure over an unweighted counterpart could depend on the tasks used, the specific language under study, or the varieties included in the task. For example, (Pettersson et al., 2013), who proposed the specific assignments of weights which Levy et al. (2019) used, formulated these weights to normalize historic text to a more modern spelling for natural language processing of text in Swedish. Thus, this specific weighted Levenshtein metric could be better suited for text-based (versus speech-based) tasks. Consonant-to-consonant and vowel-to-vowel changes may differentially affect readers and listeners, based on the cognitive processes recruited for each task. How well these penalties predict perceptual accent judgments could also be related to the language under investigation. It is possible that the effect of certain deviations on intelligibility for a Swedish listener (or reader; Pettersson et al., 2013) would be different than the effect on German (Levy et al., 2019) or English (Bent et al., 2021) listeners. Further, consonants and vowels may differentially impact perceived accentedness among varieties, with consonant changes factoring in more heavily for nonnative varieties (Gao, 2019) and vowel changes better distinguishing among non-local native varieties (Clopper et al., 2005).

Although the current study compared two different Levenshtein calculations, there are many ways to calculate segmental changes that could possibly capture more of the variability in accent rankings. The impact of vowel versus consonant changes; use of phonetic-level transcription and scoring (i.e., inclusion of diacritics); or an alternative weighting system are all examples of possible ways in which segmental changes could be quantified and compared to perceptual accent rankings. For example, Vieru et al. (2011) revealed certain segmental changes could be used as cues to accent identification by French listeners, such as /b/ → /v/ indicating a native Spanish versus native Italian speaker of French. Though these changes were dependent upon speakers' native language, results demonstrate that generalizations regarding the relative importance of certain segmental changes could be made across varieties. In the current study, the relative difference in the predictive value of unweighted and weighted Levenshtein metrics to accent rankings was fairly minimal, suggesting that differences in Levenshtein calculations may be relatively minor. However, the way in which the segmental changes are calculated may differentially correlate with accent perception based on the type or length of stimuli (i.e., sentences versus words). Future work could investigate the utility of various segmental calculations in predicting stimuli of differing lengths.

4.1.2. Dynamic time warping

DTW was a significant predictor of ladder rankings, even when controlling for Levenshtein distance, indicating that holistic acoustic distance significantly explains variability in perceptual judgments of accent distance beyond phonemic distance alone. That being said, DTW contributed slightly less to perceptual accent rankings than did either Levenshtein measure. This result supports the findings of Bartelds et al. (2020) who similarly found a unique but modest contribution of DTW in explaining variability in accent judgments. Bartelds et al. (2020) suggested that recording inconsistencies may have influenced their findings. In the current study, all of the stimuli were recorded under similar high-quality conditions. The similar findings between Bartelds et al. (2020) study and the current study suggest that recording inconsistencies are not likely the primary source of the relatively small contribution of DTW measures to accent judgments.

There are several other possible explanations for the relatively

Table 6
AIC and Log-likelihood values for models predicting ladder results from unweighted and weighted Levenshtein and dynamic time warping.

	AIC	Log-likelihood
Unweighted Levenshtein + DTW	29,316.7	−14,636.3
Weighted Levenshtein + DTW	29,306.9	−14,631.5

modest independent contribution of DTW to variability in perceptual accent judgments: (1) its importance may be dependent on the language background of the speaker and/or the listener; and (2) it does not discriminate between linguistic and non-linguistic acoustic information. Fig. 2 shows that the holistic acoustic distance of certain accents (e.g., German, Scottish) played a larger role in listeners' perceptions of accent distance than other accents (e.g., Korean). The Levenshtein scores of the German and Scottish speakers were relatively high and low, respectively, compared to their ladder rankings, and the pattern of ladder rankings more closely matched the DTW scores for these two accents. This pattern of results suggests that the holistic acoustic distances between Midland American- and both Scottish- and German-accented English are driving listeners' perceptions of accent distance of these accents, more so than phonemic differences. On the other hand, Korean-accented speakers had a relatively low holistic acoustic distance (ranking only slightly above the British-accented speakers) and yet their ladder ranking seems to correlate more with their relatively high Levenshtein scores. For this accent, phonemic changes seem to contribute more heavily than holistic acoustic distance to listeners' decision to rank them as relatively farther from Midland American English. Thus, listeners' reliance on phonemic versus holistic acoustic distance to make their accent distance judgments may relate to the language background of the speaker. There are likely to be other factors that impact perceived distance that were not measured in this study, including attitudes about particular accents along dimensions related to solidarity and status.

Likewise, the utility of the DTW measure in explaining accent judgments may be dependent upon the listener's language background. Results from the current study suggest that native, monolingual American English listeners attend to cues that are not fully captured by phoneme deviations – such as prosodic or subphonemic cues – as evidenced by the significance of DTW in predicting judgments of accent distance. However, phonemic cues (i.e., Levenshtein distances) were more important than DTW in predicting ladder rankings for these monolingual American English listeners. This finding is consistent with previous research that found a relatively minimal benefit from prosodic information in distinguishing among English dialects (van Bezooijen and Gooskens, 1999; Alcorn et al., 2020) or when identifying the native language status of a French or English speaker (Grover et al., 1987; Vieru et al., 2011). On the other hand, prosodic information is important for distinguishing among Norwegian dialects (Gooskens, 2005). Norwegian listeners may therefore rely more heavily than American English listeners on prosodic cues in making accent strength or distance judgments. Further cross-linguistic support for this interpretation comes from Boula de Mareüil and Vieru-Dimulescu's (2006) study, which demonstrated a significant role of prosody in identifying both Spanish-accented Italian and Italian-accented Spanish. A comparison of the effectiveness of DTW in predicting accent judgments using the same task but with varied target languages (and listeners who are native speakers of those target languages) could provide insight into how DTW performs in predicting judgments as a function of talker and listener language background. For example, Bradlow et al. (2010) compared perceived distance from English (using ladder rankings) of 17 languages by listeners of 5 different native language backgrounds, revealing significant correlations among the ladder rankings based on native language background. Identifying the relative contributions of phonemic cues versus holistic acoustic cues in these perceptual distances from English as a function of listeners' native languages would reveal how native language background shapes which cues listeners attend to in making accent judgments.

Another possible explanation for the modest contribution of the DTW variable in predicting perceptual accent judgments could be related to extra, non-linguistic acoustic information captured by MFCCs that might obscure what listeners use to make accent judgments. MFCCs capture a global acoustic picture of the signal but are not able to differentiate between linguistic and non-linguistic content. Thus, this measure is likely capturing acoustic information that is not relevant to accent judgments. Bartelds et al. (2020) noted that the difficulty in generating computational representations of phonetic information is in the ability to capture only what is important, without superfluous acoustic information that may not contribute to accent judgments. Further, MFCCs are impacted by speaker-level variability, such as vocal tract anatomy. In fact, out of the 111 sentences analyzed (37 speakers, 3 sentences per speaker), the lowest DTW distance between the target and reference (Midland) speakers were speakers of the same gender in 107 instances, suggesting that vocal tract anatomy may play a substantial role in DTW scores. Even within genders, there is likely a fair amount of variability in vocal tract length (and anatomy in general) that could influence MFCC calculations. Including this sort of idiolectal information that varies from speaker to speaker in MFCC calculations represents a limitation of the DTW variable. Although DTW captures sub-phonemic, phonemic, and prosodic cues, the relative importance of each of these cues as well as the addition of other non-linguistic acoustic information captured by DTW make it difficult to discern what non-phonemic information is truly important in perceptual accent judgments.

One clear advantage DTW demonstrates over the phonemic measures is in its indifference to context. In the current study, DTW did not significantly interact with the sentence variable in either of the mixed effects models. Further, the relatively small error bars around the individual speaker means for the DTW scores compared to the Levenshtein scores (see Fig. 2) indicate that the DTW scores were less variable across sentences than the Levenshtein scores. This finding is expected, given that the Levenshtein measure is phoneme-based and therefore will vary depending upon which phonemes are present in a given sentence. Thus, DTW can provide an acoustic distance measure that is more impervious to sentence content than the Levenshtein distance, despite being outperformed by the Levenshtein distance in predicting accent distance judgments.

Levenshtein distances and DTW represent imperfect but perhaps complementary distance measures. One advantage of the Levenshtein distance is its ability to quantify changes in the speech signal (specifically, segmental changes) that may reflect how listeners perceptually weigh these changes. Limitations of this measure include the introduction of human bias when manually transcribing speech stimuli, and its inability to capture anything beyond the phoneme (at least in its instantiation in the current study). DTW (in this case, MFCCs), on the other hand, is an objective measure (and therefore, more resistant to human-level error in its calculation) and is able to capture a wide range of acoustic information. However, it captures only the general shape of the spectrum (Ryant et al., 2014), and weighs all of the acoustic information equally, which provides a poor representation of the cognitive underpinnings of accent perception. Human listeners take in the entire acoustic signal – not just the phonemes – but place more weight on linguistic information and are able to ignore irrelevant acoustic information when making accent judgments.

In summary, all three objective distance measures contributed to perceptual judgments of accent distance, as measured by the ladder task. Comparisons across models revealed the weighted Levenshtein distance as the best predictor of the perceptual accent distance rankings, although the differences were modest. However, these objective distance measures do not account for all of the variability in perceptions of accent distance. Fig. 3 demonstrates that although the distance measures perform fairly well in predicting accent judgments, there are clearly unaccounted for factors that contribute to listeners' accent rankings. Listeners could be using metalinguistic cues, social knowledge or assumptions, or information relating to intelligibility or comprehensibility to make accent judgments. These factors were outside of the scope of the current study, and therefore unaccounted for, but could have exerted influence on listeners' judgments.

The next steps in identifying the cues that contribute to listeners' perceptual judgments of both non-local native and nonnative accents include disentangling what non-phonemic cues (as captured by DTW) listeners are using that contribute to perceptual judgments beyond the

phoneme level (as captured by the Levenshtein distances). Objectively quantifying prosodic information (such as rhythm, F0 changes, and intonation, among others) is an important next step in determining the presence and relative importance of these cues in making perceptual judgments. Although DTW provides a good first step in capturing some of this information, this holistic acoustic distance measure casts a relatively wide net, making it difficult to draw conclusions about which cues are most important in accent judgments. Further, some acoustic information that was controlled for by the DTW measure in the current study – such as speaking rate – could also add to the understanding of how non-segmental information affects accent judgments. Pursuing self-trained neural models to predict accent distance (from Bartelds et al., 2022) is another potentially worthwhile future direction, if the cost- and time-related barriers to training these models could be addressed.

4.2. Non-local native versus nonnative accent rankings

The present study included a variety of non-local native and nonnative accents. A few studies have included both nonnative and native accents (Adank et al., 2009; Bent et al., 2016; Goslin et al., 2012, 2021; Floccia et al., 2009), yielding mixed results. Results from some studies have indicated that accent judgments are consistent with a native vs. nonnative distinction (i.e., all non-local native accents are rated as closer or less strong than nonnative; see Adank et al., 2009; Bent et al., 2016; Goslin et al., 2012), while others have demonstrated that listeners are not classifying accents based on native-status (Bent et al., 2021; Floccia et al., 2009; Levy et al., 2019). Adank et al. (2009) assessed reaction times of listeners to true/false questions presented in noise and in quiet with speakers of familiar and unfamiliar native and nonnative accents. Overall, a greater processing cost was seen for the nonnative than native accents. Floccia et al. (2009) investigated differences in adaptation to accent changes when the accents were nonnative versus non-local native, using reaction times to identify processing effects based on native language status. Diverging from Adank et al. (2009), they found a significant increase in reaction time when accent stimuli changed from baseline (local Plymouth English) to both non-local native and nonnative accents; still, the effect was stronger for the change to the nonnative accent than to the non-local native accent, suggesting some degree of processing differences of these two accent groups. It is worth noting that only one nonnative and one non-local native accent were included as stimuli in both Floccia et al.'s and Adank et al.'s studies, limiting the generalizability of this finding. In contrast, Bent et al. (2021) included one local native, three non-local native, and three nonnative accents in an intelligibility task (in quiet and noise), and reported that although performance by both children and adults was overall better for native and non-local native compared to nonnative stimuli, there was variability at the speaker-level. In other words, certain nonnative speakers' stimuli (e.g., German, Mandarin) yielded better intelligibility scores than non-local native stimuli. The results of the current study are consistent with Bent et al.'s (2021) findings in revealing variability in accent distance both within and between non-local native and nonnative accents, reflecting speaker-level variability. Although inclusion criteria for the non-native speakers of having lived in the United States for no more than 4 years attempted to control for accent strength, the overall strength of the accents (both non-local and nonnative) was not explicitly assessed. Certainly, different language learning profiles and residential histories contributed to the observed speaker-level variability in the current study.

Cristia et al. (2012) questioned the importance of the distinction between native and nonnative accents, particularly in how listeners interpret or consider these accents in their judgments of strength or distance. They challenged the notion that native dialects differ at the segmental level only, citing White et al. (2012) study demonstrating more prosodic similarity between Dutch and Standard Southern British English than between Standard Southern British and Glaswegian English. The inclusion of both non-local native and nonnative accents in the current study cannot speak to the question of perceptual distinction in processing of these accents. Although the present study included a relatively larger number of both accents and speakers from each accent (as compared to Floccia et al. (2009), for instance), suggestions of differences in strength, distance, or variability between non-local native and nonnative are limited by speaker-level variability, the small number of talkers included for each variety, and the relatively limited number of phonemes in various word-level positions. Still, investigating both non-local native and nonnative accents together in one study can provide a broader view of perception, as it provides a greater breadth of both acoustic and phonemic cues in the speech signal.

Future work could also include less-often studied native and nonnative accents (e.g., speakers from countries in Kachru's (2006) "outer circle," such as Pakistan or South Africa), to improve the generalizability of the findings of what contributes to perceptual judgments of accent distance. Including diverse and less often studied non-local native and nonnative accents would help expand the current understanding of what phonemic and non-phonemic cues are important for accent distance judgments.

Conclusion

Both phonemic and holistic acoustic distance cues are used by American English-speaking listeners when making judgments of accent distance for both non-local native and nonnative accents, with phonemic cues contributing more to accent rankings than holistic acoustic distance cues. The unweighted and weighted Levenshtein distances both significantly predicted accent distance judgments, with the weighted slightly outperforming the unweighted. The holistic acoustic distance measure is agnostic to the nature of the content it is analyzing (i.e., whether or not it is linguistically relevant), and may include extra, non-linguistic acoustic information that dampens its predictive performance. The significance of the phonemic versus the holistic acoustic distance measures in predicting perceptual accent judgments may be partially due to the native variety of the speaker. However, both phonemic and acoustic distance measures have limitations that are somewhat mitigated by using both to assess the potential cues used by listeners to make accent distance judgments. For the purposes of investigating cues used by listeners when making accent distance judgments, analyzing only non-local native or nonnative accents–as opposed to considering both of these accent groups–may be unnecessarily restrictive, when both accent groups provide significant information about how phonemic and acoustic cues are used in accent judgments.

Open practices statements

Experimental and statistical code, stimuli, and data are available at https://osf.io/cnxgt/. The experiments were not preregistered.

Declarations

Funding: This work was supported by the National Science Foundation (Award numbers: 1,941,691 (Bent) and 1,941,662 (Holt) and the Ohio State University Center for Cognitive and Brain Sciences (Lind-Combs).

Conflicts of interest/Competing interests: The authors have no relevant financial or non-financial interests to disclose.

Ethics approval: The research was approved by the Institutional Review Boards at Indiana University.

Consent to participate: Informed consent was obtained from all individual participants included in the study.

Availability of data and materials: Data are available at https://osf.io/cnxgt/.

Code availability: Statistical and experimental code are available at https://osf.io/cnxgt/.

CRediT authorship contribution statement

Holly C. Lind-Combs: Software, Formal analysis, Data curation, Writing – original draft, Visualization, Funding acquisition. **Tessa Bent:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Rachael F. Holt:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Project administration, Funding acquisition. **Cynthia G. Clopper:** Formal analysis, Writing – review & editing. **Emma Brown:** Methodology, Investigation.

Declaration of Competing Interest

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.

Data availability

All data, code, and stimuli have been uploaded to the OSF repository. The link to the OSF page is included in the title page of the manuscript.

Acknowledgments

We are grateful to all the participants who took part in our studies. We would like to acknowledge the significant contributions of Joe Moder, for providing consultation in the development of Python scripts for calculating MFCCs and DTW; and Lindsey Altum, Megan Hancock, Yi Liu, Ali Stallons, and Amy Warrington for their work on transcribing stimuli. We also gratefully acknowledge funding from the National Science Foundation (Award Nos: 1941691 [Bent], 1941662 [Holt]) and The Ohio State University Center for Cognitive and Brain Sciences (Lind-Combs).

Appendix A

Accent	Gender	Sentence
British	Female	The clown has a funny face.
	Male	The boy fell from the window
		A lady went to the store.
French	Male	The dishcloth is soaking wet.
		The oven door is open.
	Female	They had some chocolate pudding.
		The bus stopped suddenly.
German	Female	She's paying for her bread.
		The dinner plate was hot.
	Male	He broke his leg again.
		The lady wore a coat.
Hindi	Female	The baby has blue eyes.
		They're shopping for school clothes.
	Male	They have two empty bottles.
		The kitchen window is clean.
Japanese	Male	They are coming for dinner.
		The table has three legs.
	Female	A child ripped open the bag.
		The sun melted the snow.
Korean	Female	The baby slept all night.
		There was a bad train wreck.
	Male	The puppy played with the ball.
		The old woman was at home.
Mandarin	Male	They're watching the train go by.
		The woman cleaned her house.
	Female	The oven was too hot.
		A girl came into the room.
Scottish	Female	They had a wonderful day.
		They finished dinner on time.
Spanish	Male	The big boy kicked the ball.
		A dog was eating some meat.
	Female	He's washing his face with soap.
		They are drinking coffee.

References

Abercrombie, D., 1967. Elements of General Phonetics. Aldine Pub. Co. http://books.google.com/books?id=SvVYAAAAMAAJ.

Adank, P., Evans, B.G., Stuart-Smith, J., Scott, S.K., 2009. Comprehension of familiar and unfamiliar native accents under adverse listening conditions. J. Exp. Psychol. Hum. Percept. Perform. 35 (2), 520–529. https://doi.org/10.1037/a0013552.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6), 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Alcorn, S., Meemann, K., Clopper, C., Smiljanic, R., 2020. Acoustic cues and linguistic experience as factors in regional dialect classification. J. Acoust. Soc. Am. 147, 657–670. https://doi.org/10.1121/10.0000551.

Anderson-Hsieh, J., Johnson, R., Koehler, K., 1992. The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. Lang. Learn. 42 (4), 529–555. https://doi.org/10.1111/j.1467-1770.1992.tb01043.x.

ANSI, 2004. S3.21-2004, Methods for Manual Pure-Tone Threshold Audiometry. American National Standards Institute, New York.

Bartelds, M., Richter, C., Liberman, M., Wieling, M., 2020. A new acoustic-based pronunciation distance measure. Front. Artif. Intell. 3 https://doi.org/10.3389/frai.2020.00039.

Bartelds, M., de Vries, W., Sanal, F., Richter, C., Liberman, M., Wieling, M., 2022. Neural representations for modeling variation in speech. J. Phon. 92, 101137 https://doi.org/10.1016/j.wocn.2022.101137.

Bent, T., Atagi, E., Akbik, A., Bonifield, E., 2016. Classification of regional dialects, international dialects, and nonnative accents. J. Phon. 58, 104–117. https://doi.org/10.1016/j.wocn.2016.08.004.

Bent, T., Holt, R.F., 2017. Representation of speech variability. WIREs Cognit. Sci. 8 (4), e1434. https://doi.org/10.1002/wcs.1434.

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. J. Acoust. Soc. Am. 150 (6), 4103–4117. https://doi.org/10.1121/10.0008930.
Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot Int. 5 (9/10), 341–345.
Boula de Mareüil, P., Vieru-Dimulescu, B., 2006. The contribution of prosody to the perception of foreign accent. Phonetica 63 (4), 247–267. https://doi.org/10.1159/000097308.
Bradlow, A.R. (n.d.). SpeechBox. Retrieved in 2018 from https://speechbox.linguistics.northwestern.edu/.
Bradlow, A., Clopper, C., Smiljanic, R., Walter, M.A., 2010. A perceptual phonetic similarity space for languages: evidence from five native language listener groups. Speech Commun. 52 (11–12), 930–942. https://doi.org/10.1016/j.specom.2010.06.003.
Brennan, E.M., Ryan, E.B., Dawson, W.E., 1975. Scaling of apparent accentedness by magnitude estimation and sensory modality matching. J. Psycholinguist. Res. 4 (1), 27–36. https://doi.org/10.1007/BF01066988.
Clopper, C.G., Pisoni, D.B., de Jong, K., 2005. Acoustic characteristics of the vowel systems of six regional varieties of American English. J. Acoust. Soc. Am. 118 (3 Pt 1), 1661–1676. https://doi.org/10.1121/1.2000774.
Clopper, C.G., 2008. Auditory free classification: methods and analysis. Behav. Res. Methods 40 (2), 575–581. https://doi.org/10.3758/brm.40.2.575.
Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A., Floccia, C., 2012. Linguistic processing of accented speech across the lifespan. Front. Psychol. 3, 479. https://doi.org/10.3389/fpsyg.2012.00479.
Davis, S., Mermelstein, P., 1980. Comparison of para measurement representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. 28 (4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420.
Derwing, T.M., Munro, M.J., 1997. Accent, intelligibility, and comprehensibility: evidence from four l1s. Stud. Second Lang. Acquis. 19 (1), 1–16.
Flege, J.E., 1984. The detection of French accent by American listeners. J. Acoust. Soc. Am. 76 (3), 692–707. https://doi.org/10.1121/1.391256.
Flege, J.E., Munro, M.J., MacKay, I.R.A., 1995. Factors affecting strength of perceived foreign accent in a second language. J. Acoust. Soc. Am. 97 (5), 3125–3134. https://doi.org/10.1121/1.413041.
Floccia, C., Butler, J., Goslin, J., Ellis, L., 2009. Regional and foreign accent processing in English: can listeners adapt? J. Psycholinguist. Res. 38 (4), 379–412. https://doi.org/10.1007/s10936-008-9097-8.
Gao, Z., 2019. Weighing Phonetic Patterns in Nonnative English Speech. George Mason University [Doctoral dissertation].
Gooskens, C., 2005. How well can Norwegians identify their dialects? Nord. J. Linguist. 28, 37–60. https://doi.org/10.1017/S0332586505001319.
Gooskens, C., Heeringa, W., 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. Lang. Var. Change 16 (03). https://doi.org/10.1017/S0954394504163023.
Goslin, J., Duffy, H., Floccia, C., 2012. An ERP investigation of regional and foreign accent processing. Brain Lang. 122 (2), 92–102. https://doi.org/10.1016/j.bandl.2012.04.017.
Grover, C., Jamieson, D.G., Dobrovolsky, M.B., 1987. Intonation in English, French and German: perception and production. Lang. Speech 30, 277–296.
Holm, S., 2008. Intonational and Durational Contributions to the Perception of Foreign-Accented Norwegian: An experimental Phonetic Investigation. Norwegian University of Science and Technology [Doctoral dissertation].
Kachru, B.B., 2006. The English language in the outer circle. World Engl. 3, 241–255.
Kessler, B., 1995. Computational dialectology in Irish Gaelic. In: Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics. https://aclanthology.org/E95-1009.
Kolly, M.J., Boula de Mareüil, P., Leemann, A., Dellwo, V., 2017. Listeners use temporal information to identify French- and English-accented speech. Speech Commun. 86, 121–134. https://doi.org/10.1016/j.specom.2016.11.006.
Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. 10 (8), 707–710.
Levy, H., Konieczny, L., Hanulíková, A., 2019. Processing of unfamiliar accents in monolingual and bilingual children: effects of type and amount of accent experience. J. Child Lang. 46 (2), 368–392. https://doi.org/10.1017/S030500091800051X.
Magen, H.S., 1998. The perception of foreign-accented speech. J. Phon. 26 (4), 381–400. https://doi.org/10.1006/jpho.1998.0081.
Major, R.C., 2007. Identifying a foreign accent in an unfamiliar language. Stud. Second Lang. Acquis. 29, 539–556.
McFee, B., Raffel, C., Liang, D., Ellis, D.P.W., McVicar, M., Battenberg, E., Nieto, O., 2015. librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, pp. 18–25.
Munro, M.J., 1995. Nonsegmental factors in foreign accent: ratings of filtered speech. Stud. Second Lang. Acquis. 17 (1), 17–34.
Munro, M.J., Derwing, T.M., 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Lang. Learn. 45 (1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x.
Munro, M.J., Derwing, T.M., Burgess, C.S., 2010. Detection of nonnative speaker status from content-masked speech. Speech Commun. 52 (7), 626–637. https://doi.org/10.1016/j.specom.2010.02.013.
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. J. Acoust. Soc. Am. 95 (2), 1085–1099. https://doi.org/10.1121/1.408469.
Park, H., 2013. Detecting foreign accent in monosyllables: the role of L1 phonotactics. J. Phon. 41, 78–87. https://doi.org/10.1016/j.wocn.2012.11.001.
Pettersson, E., Megyesi, B., Nivre, J., 2013. Normalisation of historical text using context-sensitive weighted levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Linköping University Electronic Press, Sweden, Oslo, Norway, pp. 163–179.
R. Core Team, 2021. R: A Language and Environment For Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL. https://www.R-project.org/.
Rescorla, R., Wagner, A., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Classical Conditioning II: Current Research and Theory, 2. Vol..
Riney, T.J., Takagi, N., Inutsuka, K., 2005. Phonetic parameters and perceptual judgments of accent in English by American and Japanese listeners. TESOL Q. 39 (3), 441–466. https://doi.org/10.2307/3588489.
Ryant, N., Slaney, M., Liberman, M., Shriberg, E., Yuan, J., 2014. Highly accurate Mandarin tone classification in the absence of pitch information. Speech Prosody 2014, 673–677. https://doi.org/10.21437/SpeechProsody.2014-123.
Sereno, J., Lammers, L., Jongman, A., 2016. The relative contribution of segments and intonation to the perception of foreign-accented speech. Appl. Psycholinguist. 37 (2), 303–322. https://doi.org/10.1017/S0142716414000575.
Southwood, M.H., Flege, J.E., 1999. Scaling foreign accent: direct magnitude estimation versus interval scaling. Clin. Linguist. Phon. 13 (5), 335–349. https://doi.org/10.1080/026992099299013.
van Bezooijen, R., Gooskens, C., 1999. Identification of language varieties: the contribution of different linguistic levels. J. Lang. Soc. Psychol. 18 (1), 31–48. https://doi.org/10.1177/0261927X99018001003.
Vieru, B., Boula de Mareüil, P., Adda-Decker, M., 2011. Identification and characterisation of non-native French accents. Speech Commun. 53, 292–310.
Vitale, M., Mareüil, P.B.D., Meo, A.D., 2014. An acoustic-perceptual approach to the prosody of Chinese and native speakers of Italian based on yes/no questions. Speech Prosody 2014, 648–652. https://doi.org/10.21437/SpeechProsody.2014-118.
Wayland, R., 1997. Non-native production of Thai: acoustic measurements and accentedness ratings. Appl. Linguist. 18, 345–373.
Weinberger, S.H., Kunath, S.A., 2011. The Speech Accent Archive: Towards a Typology of English Accents. Brill, pp. 265–281. https://doi.org/10.1163/9789401206884_014.
White, L., Mattys, S.L., Wiget, L., 2012. Language categorization by adults is based on sensitivity to durational cues, not rhythm class. J. Mem. Lang. 66 (4), 665–679. https://doi.org/10.1016/j.jml.2011.12.010.
Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., Baayen, R.H., 2014a. A cognitively grounded measure of pronunciation distance. PLoS One 9 (1), e75734. https://doi.org/10.1371/journal.pone.0075734.
Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., Nerbonne, J., 2014b. Measuring foreign accent strength in English: validating Levenshtein distance as a measure. Lang. Dyn. Change 4 (2), 253–269. https://doi.org/10.1163/22105832-00402001.