Journal of Phonetics 99 (2023) 101255

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

Perception and production of Mandarin-Accented English: The effect of degree of Accentedness on the Interlanguage Speech Intelligibility Benefit for Listeners (ISIB-L) and Talkers (ISIB-T)

Sheyenne Fishero *, Joan A. Sereno, Allard Jongman

Department of Linguistics, University of Kansas, 1541 Lilac Ln Room 427, Lawrence, KS 66045, USA

ARTICLE INFO

Article history:
Received 5 July 2022
Received in revised form 9 May 2023
Accepted 24 May 2023
Available online xxxx

Keywords:
Interlanguage
Intelligibility
Accentedness
L2 proficiency
L2 learners
Nonnative speakers
Nonnative listeners

ABSTRACT

Previous research on the Interlanguage Speech Intelligibility Benefit (ISIB) indicates nonnative listeners may have an advantage at understanding nonnative speech of talkers with the same first language (L1) due to shared interlanguage knowledge. The present study offers a comprehensive analysis of various factors that may modulate this advantage, including the proficiency of both the listeners and the talkers, the mapping of phonemes between the L1 and second language (L2), and the acoustic properties of the phones. Accuracy scores on a lexical decision task were used to investigate both native English listeners' and native Mandarin learners' of English perception of native English and Mandarin-accented English speech. Results show clear ISIB-L and ISIB-T effects and demonstrate the dynamic nature of ISIB effects, with both being modulated by speaker and listener proficiency. More striking ISIB effects typically occur at the most extreme ends of accentedness. Additionally, an advantage for common-phoneme over unique-phoneme words in nonnative speech was observed. While nonnative productions of common-phoneme words are more accurate than those of unique-phoneme words, for the most accented productions, nonnative listeners are faster to respond to these unique, often mispronounced, productions.

The nonnative listener advantage at perceiving nonnative speech depends on various factors, including listener proficiency, speaker proficiency, phoneme characteristics, and the acoustics of specific speech tokens.

1. Introduction

Production of second language (L2) speech sounds, even those that are similar to sounds in a speaker's first language (L1), rarely begins (or even finishes) at a nativelike level. Instead, L2 speech typically retains an accent throughout an adult learner's life. For example, previous research has found that productions of French [u] from native English adult learners of French, who had lived in Paris for 12 years, did not have nativelike F2 values (Flege, 1987). Adult learners of an L2 do not begin their acquisition journey with a blank slate, but instead, they initially refer to their native language when learning an L2. As they begin to accumulate knowledge about their new language, learners develop an interlanguage that contains information from both the L1 and L2 (Selinker, 1972). The non-native speech that results may also be more difficult than native speech for native listeners to understand. For example, previous research has found that native Dutch listeners required a 3 dB better signal-to-noise ratio to reach the speech reception threshold for English-accented Dutch speech compared to native Dutch speech, demonstrating that foreign-accented speech often results in increased difficulty in comprehensibility for native listeners (Van Wijngaarden, 2001).

Whether this increased difficulty in comprehending accented speech holds true for non-native listeners is still highly debated. According to the Interlanguage Speech Intelligibility Benefit (ISIB), a talker and listener who share the same first and second language have an improved ability to understand each other when communicating in their second language because both interlocutors have a shared interlanguage (Bent & Bradlow, 2003). The ISIB phenomenon has been identified across a variety of different languages (Bent & Bradlow, 2003; Han, Choi, Lim, & Lee, 2011a; Hayes-Harb, Smith, Bent, & Bradlow, 2008; Imai, Walley, & Flege, 2005; Koo, 2018; Li & Mok, 2015; Pinet, Iverson, & Huckvale, 2011; Sereno, McCall, Jongman, Dijkstra, & van

* Corresponding author.
E-mail address: sfishero@ku.edu (S. Fishero).

https://doi.org/10.1016/j.wocn.2023.101255
0095-4470/

Heuven, 2002; So & Attina, 2014; Stibbard & Lee, 2006; Xie & Fowler, 2013). However, several studies have found little or no evidence of a non-native speech intelligibility benefit (Algethami et al., 2011; Chen, 2015; Munro, Derwing, & Morton, 2006). Understanding the role of both the listener and the talker in the manifestation of the ISIB will offer greater insight into the circumstances that allow an ISIB to arise and a better grasp of foreign-accented speech perception and production.

1.1. Interlanguage speech intelligibility benefit

1.1.1. Matched and mismatched ISIB effects

In Bent and Bradlow (2003), English sentences produced by high- and low-proficiency native Chinese, high- and low-proficiency native Korean, and native English speakers were played to native English, native Chinese, native Korean, and a mixed group of L1 listeners to transcribe what they heard. A matched ISIB, that is, a benefit in communication between talker and listener when they share the same L1 (Bent and Bradlow (2003) defined a benefit as situations in which one language group performed equally well or better than another group), was found for non-native listeners, who found the high-proficiency speaker who shared their L1 equally intelligible as the native English speaker. A matched ISIB was also found for Korean listeners of low-proficiency Korean speech, which they found as intelligible as native English speech. A mismatched ISIB, an advantage in L2 communication between talker and listener without the same L1, was also found when listening to high-proficiency non-native English.

While Bent and Bradlow (2003) offered evidence of an ISIB, it also introduced a number of issues. This study defined "benefit" to mean one group performs as well as another group, but this definition of the term does not always capture an advantage. The offline nature of the task also brings into question whether tasks that allow more time for responses will be more likely to demonstrate evidence of an ISIB due to the reduced stress they place on nonnative listeners, who typically have lower processing speed abilities when processing their L2, while time-sensitive tasks (such as RT measures) may exacerbate native listeners' advantage over nonnative listeners when listening to native speech (McDonald, 2006). This means that when ISIB effects are found in on-line time-sensitive tasks, they may offer stronger support for the ISIB phenomenon because they indicate the strength of the nonnative language advantage surpasses that of the native processing speed advantage. Additionally, if no ISIB effects are found from accuracy scores due to ceiling effects, reaction time measures that are available from online tasks may offer another means by which to compare groups. McLaughlin and Van Engen (2020) found that even when accuracy may be at ceiling for L1 listeners of intelligible L2-accented speech, they still have greater cognitive load compared to when listening to native-accented speech, which could potentially manifest itself in reaction time measures. Fine-grained methodology and online tasks may be able to identify the more subtle characteristics of the perception of foreign-accented speech that accuracy scores alone may not always capture.

In order to identify the role of phonological similarity of languages in the occurrence of a mismatched ISIB, a follow-up study examined high- and low-proficiency Arabic- and Korean-accented English and native English speech perception by L1 Korean, L1 Saudi Arabian Arabic, L1 English, and a mixed L1 group of listeners (Stibbard & Lee, 2006). These languages were chosen because they are more distinct than Chinese and Korean in terms of syllable structure constraints. Korean and Chinese do not allow coda-final consonant clusters, a trait that could similarly impact the way in which speakers of these languages produce and perceive English speech, so it is plausible this similarity in phonological structure between both L1 languages may have contributed to the emergence of a mismatched ISIB in Bent and Bradlow (2003). The term "benefit" was defined only as situations in which one group performed better than another group, not just when two groups performed similarly.

Results based on keyword transcription accuracy scores showed that Korean speakers were more intelligible to native Korean listeners than low-proficiency Arabic speakers, and high-proficiency Korean speakers were the most intelligible to Korean listeners, thus demonstrating a matched ISIB. However, they found no indication of a mismatched ISIB. The results of Stibbard and Lee (2006) likely differed from Bent and Bradlow (2003) in finding no evidence of a mismatched ISIB because of their decision to choose two unrelated L1 languages, Korean and Arabic, as well as their methodological changes, such as removing the speaker-blocked design. While Stibbard and Lee (2006) provide additional evidence of the presence of an ISIB, their study demonstrated that speakers and listeners must share an L1 in order to show improved intelligibility. Both of these studies relied on sentence-level processing to test for an ISIB, which might have been affected by other contextual variables such as sentential prosody or semantics. These properties may have given native English speakers an advantage in the task and minimized ISIB effects that may otherwise be present. Furthermore, no detailed categorization of listener proficiency was included in either study, which could potentially impact whether an ISIB emerges.

1.1.2. ISIB-L and ISIB-T

The ISIB may not reflect a single phenomenon, but instead, may reflect two separate phenomena (Hayes-Harb et al., 2008). ISIB-L (ISIB-Listener) is defined as a non-native speech comprehension advantage for non-native listeners over native listeners, while ISIB-T (ISIB-Talker) is defined as a non-native listener comprehension advantage for non-native talker speech over native talker speech (Hayes-Harb et al., 2008). This differs (i.e., the separate contribution of nonnative over native listeners and nonnative talkers over native talkers) from the matched and mismatched ISIB distinctions found in Bent and Bradlow (2003) and Stibbard and Lee (2008) that focused solely on whether the interlocutors shared an L1. The Hayes-Harb et al. (2008) study involved high- and low-proficiency Mandarin learners of English and native English speakers recording English minimal pairs that differed in word-final stop voicing. Listeners included high- and low-proficiency native Mandarin learners of English, as well as native English listeners. Researchers avoided sentence-level processing effects by using a forced-choice identification task to distinguish between voiced and voiceless word-final stops in Mandarin-accented English. Word identification accuracy scores were

analyzed in order to test for ISIB-L and ISIB-T effects. Both listeners and speakers were categorized as low- or high-proficiency by means of native English listeners' accentedness judgment scores of participants' speech.

Results indicated that low-proficiency native Mandarin listeners were more accurate than native English listeners in identifying the stop voicing of low-proficiency native Mandarin speakers, thus demonstrating an ISIB-L for low-proficiency listeners. Results do not give any evidence of an ISIB-T. Furthermore, results demonstrated that identifying the proficiency level of listeners and not just speakers is important because the presence of an ISIB effect differed depending on whether listeners were of high or low proficiency, although more fine-grained gradient distinctions in proficiency were not analyzed. The researchers posit that the reason an ISIB-L was found only for low-proficiency interlocutors may result from less variability in L2 English exposure for less-experienced low-proficiency language learners. This greater similarity in interlanguage may be what causes an ISIB. Because an ISIB-L was found independently from an ISIB-T, this study also provides evidence that these are two phenomena that should be separately analyzed.

The acoustic characteristics of foreign-accented speech that may drive ISIB effects were also measured. While no significant differences arose for any of the absolute measures (absolute and relative duration of the vowel preceding the final stop, as well as the absolute and relative final stop voicing, burst, and closure durations were measured) between talker types, native English talkers had a greater proportion of voicing in their final voiced stop productions and a smaller proportion of voicing in their final voiceless stop productions compared to native Mandarin talkers. Additionally, analyses were conducted on low-proficiency native Mandarin speakers' tokens that had the greatest difference in accuracy between native English and low-proficiency listeners (with low-proficiency participants being more accurate). The acoustic measurements of these stimuli showed that for the accurate low-proficiency listeners, voiceless stops had a longer voicing duration and a greater proportion of voicing, and voiced stops had a lower proportion of voicing and a shorter voicing duration. This offers evidence that low-proficiency listeners may not use acoustic cues in the same way as native English listeners when listening to low-proficiency Mandarin-accented English.

1.1.3. Phoneme presence in L1 and L2

Previous research compared accuracy scores and reaction time measures in a lexical decision task for stimuli with phonemes common to both English and Dutch and for stimuli containing phonemes unique to English for native English and native Dutch listeners of Dutch-accented English (Sereno et al., 2002). Results demonstrated that Dutch listeners had faster reaction times when listening to Dutch-accented English than native English (which would be defined as an ISIB-T using the Hayes-Harb et al. (2008) definition), while, for ISIB-L, Dutch listeners were equally accurate when listening to Dutch-accented and native English. Native English listeners were less accurate in their lexical decision scores when responding to stimuli spoken by Dutch-accented speakers containing phonemes unique to English compared to stimuli containing phonemes common to both Dutch and English. Furthermore, Dutch listeners were equally accurate and fast when responding to Dutch-accented unique-phoneme and common-phoneme stimuli, but slower for native English unique-phoneme stimuli compared to common-phoneme stimuli.

Acoustically examining common and unique contrasts, Han, Choi, Lim, and Lee (2011b) compared acoustic properties of vowels (vowel duration, F1, and F2 measures) in Korean-accented English to identify the underlying acoustic properties that may explain ISIB effects. Productions of English minimal pairs contrasting in vowel by [ɪ] vs [i] or [æ] vs [ɛ] were compared for native English, low-proficiency native Korean, and high-proficiency native Korean learners of English. Results showed that vowels not present in Korean ([ɪ] and [æ]) were produced with longer durations and F1 and F2 values closer to [i] and [ɛ], which are present in Korean, when produced by native Korean speakers compared to native English speakers. These studies further demonstrate the possible influence of phoneme inventories in the L1 and L2 on the presence of ISIB effects.

The predicted behavior of participants when faced with common- vs unique-phoneme words may vary depending on the theory. Some previous research indicates that common-phoneme words are easier to perceive than unique-phoneme words (Sereno et al., 2002), while unique-phoneme words are produced in a less nativelike way (Han et al., 2011b), indicating a potential perceptual advantage for common-phoneme words. On the other hand, some researchers argue that L2 phonemes that are very similar to L1 phonemes may be very difficult to produce in a nativelike way if listeners cannot perceive their differences across the L1 and L2, and that the ability to perceive is based on what the sound can be mapped onto in the L1, not necessarily just whether the sound is present in the L1 and L2 (Flege, 1995; Flege & Bohn, 2021). According to this theoretical framework (SLM-r), the unique-phoneme words may actually be produced in a more nativelike way and easier to perceive than the common-phoneme words because common-phoneme words' similarity to Mandarin phonemes may hinder Mandarin learners from perceiving their differences in English and Mandarin. Additionally, SLM-r predicts that the type of input a learner receives may play a role in what happens during L2 phonetic category creation. Exposure to foreign-accented input may be driving the creation of non-nativelike categories in the L2 for learners, indicating a dynamic relationship between talker and listener accentedness. Nonnative interlocutors with a shared L1 who never created a separate L2 category for sounds that are similar in the L1 and L2 may have similarly merged their phonetic representations into a new category with features of both the L1 and L2. This phenomenon can potentially explain any intelligibility benefit found for nonnative interlocutors with a shared L1 and L2 as representing this merged phonetic category of certain speech sounds that both interlocutors may share.

1.1.4. Foreign accent adaptation

Previous research on perceptual adaptation to foreign-accented speech has indicated that listeners are able to adapt to foreign-accented speech when exposed to a single talker over the course of an experiment, as well as when exposed to multiple talkers in training and then tested on a novel talker of the same language background (Bradlow & Bent, 2008).

Some research has even found accent-independent adaptation where training with multiple speakers with varying L1 backgrounds led to adaptation to a novel speaker with a different L1 from training (Baese-Berk, 2009). People seem to adapt to foreign-accented speech very quickly, sometimes in as little as one minute (Clarke & Garrett, 2004; Xie et al., 2018). These results indicate that the speech perception system seems flexible and adaptive to the input it receives. While foreign-accent adaptation may benefit overall perception of foreign-accented speech, it is not clear whether L1 and L2 listeners differentially benefit from this adaptation. Furthermore, previous research has found that the degree of adaptation that occurred depended on the accentedness of the talker (Bradlow & Bent, 2008).

Both the theoretical predictions of SLM-r and previous research on foreign accent adaptation point to a dynamic relationship between the degree of accentedness of the talkers and listeners in foreign-accented speech perception. This indicates the importance of measuring accentedness for both talkers and listeners in an equitable and gradient way when investigating the perception of foreign-accented speech. The present study does this by obtaining degree of accentedness ratings of the speech of both the talkers and listeners from native English judges.

Overall, the perception of L2 speech by native and nonnative listeners is influenced by several factors. While Bent and Bradlow (2003) found evidence for an Interlanguage Speech Intelligibility Benefit in which L2 learners who share the same first language have an improved ability to understand each other in their L2, additional studies found weaker and more nuanced evidence for an intelligibility advantage of nonnative speech by nonnative listeners (Algethami et al., 2011; Chen, 2015; Munro et al., 2006; Stibbard & Lee, 2006). Previous studies have identified an ISIB-L for listeners, an ISIB-T for talkers, an influence of task, as well as a contributing influence of phoneme overlap between the L1 and L2. While the studies discussed above offer great insight into the complexities of L2 speech perception research, several unanswered questions remain that the present study addresses.

1.2. The present study

The present study investigates the perception of native[1] and L2 speech by native and nonnative listeners, specifically examining proficiency of both the speakers and the listeners. Native English, weakly accented native Mandarin, and strongly accented native Mandarin listeners of English participated in a lexical decision task, listening to native English, weakly accented native Mandarin, and strongly accented native Mandarin speech in order to directly test for ISIB effects in both accuracy and reaction time measures. Detailed accentedness measures were gathered for both the listeners and the talkers, allowing for a more gradient evaluation of the ISIB effects. These accentedness scores, which index the perceived degree of foreign accent of a speaker, were used as a proxy for phonetic and phonological proficiency. The use of accentedness scores to measure both talker and listener proficiency allows for a more equitable and gradient measure of proficiency for both talkers and listeners.

The methodology of the current study aims to improve several aspects of previous studies. Degree of accentedness was measured equitably for both the talkers and listeners in the form of a native English listener accentedness judgment task. Ceiling effects of speaker accentedness were avoided by recruiting both strongly and weakly accented speakers, reaction times were collected, stimulus presentation was randomized, and top-down sentence-level processing strategies were mitigated by using a lexical decision task. Additionally, stimuli were equally divided between words that contain phonemes that occur in both English and Mandarin ("common-phoneme" words) and words that contain phonemes that only occur in English ("unique-phoneme" words). Finally, analyses of the acoustic properties of five sets of unique- and common-phoneme minimal pairs were also conducted on the speech of all speakers of the lexical decision task. Such comparisons may reveal what specific acoustic properties may drive foreign accentedness, and consequently, ISIB effects.

Thus, the following research questions are addressed by this study. Is there a native English listener and speaker advantage? Is there evidence of an ISIB-L or an ISIB-T for Mandarin-accented English? Does the degree of accentedness of both the talker and listener impact the presence of an ISIB in a gradient way? Will the ISIB effect be stronger for words that contain phonemes that only occur in English than for words with phonemes that occur in both English and Mandarin? Finally, are there acoustic differences among common- and unique-phoneme words that influence intelligibility?

We hypothesize:

(a) higher accuracy scores and faster reaction times for native English listeners compared to native Mandarin listeners hearing native English speech (native English advantage)
(b) higher accuracy scores and faster reaction times for native English speech compared to Mandarin-accented English for native English listeners (native English speaker advantage)
(c) an ISIB-L, meaning higher accuracy and faster reaction times for native Mandarin listeners compared to native English listeners hearing Mandarin-accented English
(d) an ISIB-T, meaning higher accuracy and faster reaction times for Mandarin-accented English compared to native English for native Mandarin listeners
(e) gradience in ISIB effects dependent on both listener and talker accentedness
(f) greater intelligibility and more nativelike acoustic properties for words containing phonemes that occur in English and Mandarin compared to words containing phonemes that only occur in English

The present study, using both accuracy and reaction time measures in a lexical decision task, explored the various factors that affect L2 speech perception, by examining both speaker and listener accentedness to test for ISIB effects. Degree of accentedness, used to measure proficiency, was obtained for all listener and speaker participants in an equitable way that makes it possible to compare the gradient relationship between the accentedness of listener and speaker.

[1] Our use of the term native speaker/listener was determined by information from the language background questionnaire. Specifically, a native speaker of English (Mandarin) was someone who answered affirmatively to the question "Are you a native speaker of English (Mandarin), and listed English (Mandarin) as one of the languages they spoke from 0-5 years of age. It should be noted that this designation as native did not make any assumptions about accentedness, which was evaluated independently.

Further, acoustic analyses were conducted to identify differences in pronunciation of common and unique phonemes between different language backgrounds.

2. Methods

2.1. Stimulus creation

The present experiment involved a lexical decision task with native English and Mandarin-accented English speech (collecting accuracy scores and reaction time measures).

2.1.1. Lexical decision task stimuli

120 high-frequency (defined as having at least 50 instances per million in a written English corpus, COCA; Davies, 2008) monosyllabic stimuli were recorded, divided equally among unique-phoneme words, common-phoneme words, unique-phoneme nonwords, and common-phoneme nonwords (see Appendix A). Common-phoneme words were defined as those that contain only phonemes that occur in both English and Mandarin (/p/, /t/, /k/, /f/, /m/, /n/, /s/, /l/, /w/, /j/, /i/, /u/, /ŋ/, and /ʌ/ (Lee & Zee, 2003)). Unique-phoneme words were defined as words that contained at least one phoneme that occurs in English but not in Mandarin (/b/, /d/, /g/, /v/, /h/, /ä/, /θ/, /z/, /ɪ/, and /æ/ (Lee & Zee, 2003)) (Sereno et al., 2002).

The words were all monosyllabic and high-frequency English words in order to increase the likelihood that Mandarin listeners would be familiar with all words used in the experiment. The unique-phoneme word and common-phoneme word lists were also controlled for word frequency (COCA; Davies, 2008) and phonological neighborhood density (IPhOD; Vaden, Halpin, & Hickok, 2009). In addition, all stimuli were controlled for number of phonemes and legality of syllable structure in Mandarin, and the unique-phoneme word and unique-phoneme nonword lists were controlled for number of unique segments, defined as the number of phonemes in a token not found in standard Mandarin.

2.1.2. Speaker Accentedness stimuli

An additional 10 high-frequency monosyllabic English words, including both common-phoneme and unique-phoneme words, were also recorded for a subsequent accentedness judgment task to obtain degree of accentedness scores for all speakers (see Appendix B). These scores were used to select which two native English, two strongly accented and two weakly accented speakers' stimuli would be used for the Lexical Decision Task to ensure a wide range of accentedness in the experiment.

2.1.3. Speakers

Speakers included three native speakers (one male; mean age 25.7) of a Midwestern dialect of English, three native Mandarin speakers (two males; mean age 27) with a perceived weak accent as initially judged by the first author, and four native Mandarin speakers (two males; mean age 23.3) with a perceived strong accent as initially judged by the first author. These speakers were all University of Kansas (KU) students who received $10 compensation for their participation.

2.1.4. Elicitation of stimuli

All recordings were done in an anechoic chamber at KU. The stimuli were recorded with a Marantz PMD 671 solid-state recorder using an ElectroVoice N/D 767a microphone. The recordings were digitized with a sampling rate of 22,050 Hz.

During the recording session, stimuli were shown to speakers via PowerPoint slides. Each stimulus was elicited twice. Some nonwords were elicited a third time because the participant incorrectly pronounced them during the initial recording (18/1300; 1.4%). The first token spoken by each speaker was used unless it contained an interruption, mispronunciation, or other imperfection such as lip smacking (162/1300; 12.5%, including the nonwords recorded a third time discussed above). The PowerPoint automatically proceeded at a rate of 2.5 seconds per slide for words and four seconds per slide for nonwords, with built-in 10-second breaks after roughly every 50 tokens. The word and nonword elicitations were blocked and were split up by a 25-second instruction slide that contained information about how to pronounce the nonwords. Nonwords contained a real word in parentheses that could be used to aid in the pronunciation of the nonword. For example, a slide with a nonword like 'doov' contained the word 'move' in parentheses to indicate that 'doov' should rhyme with 'move.' The initial token in the PowerPoint, as well as the initial token after each break, were fillers included to avoid prosodic word list effects. Recordings were segmented and RMS normalized using Praat (Boersma & Weenink, 2021).

Afterwards, participants were instructed to complete a detailed language background questionnaire asking about their language experience, exposure to Mandarin-accented English and native English, years spent in the United States and elsewhere, and other biographical information. Native Mandarin listeners were also asked about their training to improve their English accent, their frequency of exposure to native English speech, and the native language of their English teachers growing up.

2.2. Speaker Accentedness ratings

2.2.1. Raters

The productions of the subset of 10 stimuli used for identifying speaker accentedness (10 stimuli * 10 speakers = 100 stimuli) were presented to 5 native English raters (two males; mean age 19.6) of a Midwestern dialect of English. They were undergraduate KU students recruited from an introductory linguistics course and received extra course credit for their participation. All participants lived their entire lives in the United States. Native English raters identified the stimuli and judged their degree of accentedness to identify the strongest and weakest accented speakers. These ratings were used to select two native English speakers, two weakly accented Mandarin speakers, and two strongly accented Mandarin speakers for the lexical decision task described below.

2.2.2. Procedure

Participants were tested in the University of Kansas Phonetics and Psycholinguistics Laboratory (KUPPL). The stimuli were presented over headphones using Paradigm experimental software (Perception Research Systems, 2007), beginning

with a practice block of five monosyllabic English words not used in the experiment spoken by native English, weakly accented native Mandarin, and strongly accented native Mandarin speakers to ensure participants initially heard a wide range of accents. This was done to encourage full usage of the rating scale during the experiment.

Participants were first presented with a stimulus and instructed to identify the stimulus they heard from a set of four options, three of which were phonologically similar to the target word but contained at least one different phoneme. This identification task included stimuli selected to represent plausible misidentifications, and the four options remained constant for each word, regardless of speaker. For example, for the target word *moon,* the four options were *moon, mood, moan,* and *noon.* The first token of each of the 10 English words spoken by each speaker was presented to the judges unless it contained an interruption, mispronunciation, or other imperfection such as lip smacking (24/100; 24%).

Next, participants were instructed to select a number on a 1–5 Likert scale on the screen corresponding to the perceived degree of accentedness of each stimulus. On this scale, 1 was labeled as representing a speaker with "little to no foreign accent", and 5 was labeled as "strong foreign accent". Both tasks were self-paced, and the experiment did not proceed until a response was recorded. Mean ratings over all correctly identified stimuli from each speaker were collected in order to determine which speakers would be used in the lexical decision task described below.[2]

2.2.3. Speaker Accentedness rating results

The 120 stimuli of the male and female native English speakers with the lowest degree of accentedness rating (1.10 and 1.11, respectively), the male and female native Mandarin speakers with the lowest degree of accentedness rating (1.89 and 1.91, respectively), and the male and female native Mandarin speakers with the highest degree of accentedness rating (2.76 and 3.76, respectively) were selected as the stimuli for the Lexical Decision Task, resulting in six total speakers and 720 total stimuli.

Mean scores indicate the native English speakers' accentedness judgment scores (M = 1.11, range 1–2) were expectedly lower than those of the weakly accented Mandarin speakers (M = 1.90, range 1–5), and those of the strongly accented Mandarin speakers (M = 3.24, range 1–5). The strongly accented Mandarin speakers' accentedness judgment scores were also higher than those of the weakly accented Mandarin speakers.

2.3. Lexical decision experiment

A lexical decision task was administered to collect accuracy and reaction time measures of the native English, weakly accented native Mandarin, and strongly accented native Mandarin listeners to native English speech, weakly Mandarin-accented English speech, or strongly Mandarin-accented English speech.

[2] Mean ratings were compared per target word, and ratings generally hovered close together between a mean of 2.03 and 2.67, with mean ratings of target word *young* falling higher at 3.05.

2.3.1. Participants

36 native English (16 males, mean age 20.8) and 36 native Mandarin (15 males, mean age 28.4) speakers were tested. They were recruited from introductory linguistics classes at KU, flyers, and the Prolific recruitment system (https://www.prolific.co) (Palan & Schitter, 2018). All received either extra course credit or $10 for their participation. None of the native English speakers selected had ever studied Mandarin. All of the native Mandarin listeners were currently living in the U.S. at the time of testing (range 0.5 years–43 years, with an average of 15 years). 12 native English and 12 native Mandarin participants were tested in the KUPPL laboratory at KU, and 24 native English and 24 native Mandarin participants were tested online using Gorilla Experiment Builder. This was due to necessary changes in protocol arising from COVID-19 in-person testing restrictions.

2.3.2. Materials

The 120 words and nonwords elicited from the six speakers as described above were used for the Lexical Decision Task. The experiment was created using Paradigm experimental software (Perception Research Systems (2007), 2007) and Gorilla Experiment Builder (https://www.gorilla.sc) (Anwyl-Irvine et al., 2019). Over the course of the experiment, each listener heard all 120 stimuli once, 20 from each speaker. The stimuli were divided so that each listener heard five unique-phoneme words, five common-phoneme words, five unique-phoneme nonwords, and five common-phoneme nonwords from each of the six speakers. Six different combinations of the 120 stimuli were used to produce a full set of 720 stimuli. Listeners heard the tokens in a random order.

2.3.3. Procedure

The stimuli were presented with an intertrial interval of three seconds. Participants were first presented with a practice block containing ten practice stimuli not used in the lexical decision task. They were instructed to either press a corresponding button on a button box (if in the lab) or press the "A" or "L" keys on their computer keyboard (if online) to indicate as quickly and as accurately as possible if the stimulus they heard was a word or nonword. Participants were instructed to keep the index finger of each hand on the two buttons. Word and nonword button locations were counterbalanced across participants to reduce handedness effects. After the practice block, the lexical decision task began, which presented participants with 120 word and nonword stimuli.

After the lexical decision task, the participants heard the 60 English word stimuli again and were asked to click on the word they heard from a set of four choices, three of which were phonologically similar to the target word but contained at least one different phoneme. These stimuli were selected to represent plausible misidentifications, and the four options remained constant for each word, regardless of speaker. This task was used to ensure participants perceived and responded to the intended word or nonword during the lexical decision task. The task was self-paced, but participants were only allowed to listen to each stimulus once.

2.3.4. Analyses

Accuracy scores and reaction time measures for words only were analyzed. Words incorrectly identified as nonwords (403 /4320; 9.3%) were considered errors. Responses with reaction times that were 2 standard deviations above or below the mean of each participant (165/4320; 3.8%) were also considered errors. Finally, misidentifications during the four-alternative forced-choice identification task (427/4320; 9.9%) were also classified as errors. The overall error rate was 23.0% (995/4320). No error reaction time measures were included in the reaction time analysis. Further results are discussed in the main Results section below.

All listeners from the lexical decision task also produced the same 10 monosyllabic English words as the initial speakers did during stimulus creation to provide stimuli for native English judgments of accentedness to obtain listener degree of accentedness measures that could be used to identify possible interactions between degree of accentedness of speakers and listeners in whether an ISIB effect was found.

Those who completed the experiment in KUPPL recorded these ten words twice each using the same procedures and equipment mentioned under Stimulus Creation. Those who completed the experiment online were instructed to record these ten words twice each in a quiet room using a microphone. Thirty-five participants used a built-in laptop microphone, eleven used a headset with a built-in microphone, one used an Amazon Echo, and one used an external microphone. These participants read the wordlist in list form on their computer screen using online-voice-recorder.com to record their speech (123apps LLC, 2021). The PowerPoint and online wordlist both had filler words at the beginning and end of the list to reduce prosodic wordlist effects. Recordings were segmented and RMS normalized using Praat (Boersma & Weenink, 2021).

2.4. Listener Accentedness ratings

2.4.1. Raters

Native English judges rated the degree of accentedness of the lexical decision task participants based off their 10 monosyllabic English word productions. One set of five native English speakers (three males; mean age 19.6) of a Midwestern dialect of English recruited from introductory linguistics courses at KU rated the degree of accentedness of those tested in the laboratory, and one set of five native English speakers (three males; mean age 46.4) recruited from Prolific rated the degree of accentedness of those tested online. Stimuli recorded in-person in an anechoic chamber were not rated by the same raters as stimuli recorded remotely by participants to avoid differences in ratings due to quality of recordings.

2.4.2. Stimuli

Stimuli included the 720 recordings from the lexical decision task participants (36 English talkers * 10 words and 36 Mandarin talkers * 10 words). The first token of each of the 10 English words was used for each speaker unless the recording contained an interruption, mispronunciation, or other imperfection in the sound file (82/720; 11.4%).

2.4.3. Procedures

Procedures for listener accentedness judgments were the same as described above for the initial degree of accentedness judgment tasks of the speakers, with participants instructed to identify the stimulus they heard from a set of four options. Then, participants rated the perceived degree of accentedness of each stimulus on a 1–5 Likert scale, with 1 representing a speaker with “little to no foreign accent”, and 5 representing a speaker with a “strong foreign accent”.

The rating scores of misidentified tokens were removed from the analysis (319/3600; 8.9%). Then, mean accentedness judgment rating scores were calculated for each listener. Mean accentedness scores for native English listeners were 1.48 and ranged from 1.11 to 2.03. Mean accentendess scores for native Mandarin listeners were 2.63 and ranged from 1.26 to 4.07.

3. Results

Three mixed-effects logistic regression models were conducted using Speaker Accentedness scores and Listener Accentedness scores as continuous variables (rather than binary variables, Speaker L1 and Listener L1). The first model compared overall differences between accuracy for all speakers and listeners (ISIB effects), the second model included only native Mandarin speech responses (to test for ISIB-L effects), and the third model included only native Mandarin listener responses (to test for ISIB-T effects). The mixed-effects logistic regression models allowed for the coding of accentedness as a continuous variable, justified by the overlap of accentedness scores between native English listeners (range 1.11–2.03) and native Mandarin listeners (range 1.26–4.07), indicating a more nuanced relationship between individual participants and phonetic and phonological proficiency rather than simply native language.

3.1. Accuracy

3.1.1. Overall ISIB effects

In the first model, the dependent variable was Accuracy scores. The fixed effects included Listener Accentedness, Speaker Accentedness, Uniqueness (common-phoneme word vs. unique-phoneme word; reference = common-phoneme word), and their interactions. The model included random intercepts for subject and item[3]. Table 1 below contains the best fitting model results.

The negative simple effect of Listener Accentedness indicated that as the accentedness of listeners increased, listener accuracy decreased for common-phoneme words. The lack of interaction between Uniqueness and Listener Accentedness indicates that this pattern remained the same for unique-phoneme words.

The significant interaction between Listener Accentedness and Speaker Accentedness indicated that while more native-like speech (lower Speaker Accentedness score) had similar accuracy regardless of Listener Accentedness, more strongly accented speech (higher Speaker Accentedness score) demonstrated a larger increase in accuracy as Listener

[3] Full model syntax: Accuracy ~ Listener Accentedness * Speaker Accentedness * Uniqueness + (1 | Subject) + (1 | Item).

Table 1
Mixed-effects logistic regression analysis with best fit for all listeners' accuracy scores for all real word speech tokens for the lexical decision task.

Effect	Estimate	Std. Error	*t*	*p*
(Intercept)	1.640	0.144	11.426	<0.001
Listener Accentedness	−0.164	0.066	−2.498	0.013
Speaker Accentedness	−0.292	0.059	−4.929	<0.001
Uniqueness (Unique)	−0.398	0.192	−2.074	0.038
Listener Accentedness * Speaker Accentedness	0.223	0.046	4.809	<0.001
Speaker Accentedness * Uniqueness (Unique)	−0.588	0.082	−7.132	<0.001

Accentedness increased for common-phoneme words (see Fig. 1). The lack of three-way interaction between Uniqueness, Listener Accentedness, and Speaker Accentedness, indicates that this pattern remained the same for unique-phoneme and common-phoneme words. This finding shows that more strongly accented native Mandarin listeners had a greater advantage over more nativelike English listeners at identifying Mandarin-accented English compared to native English speech, which was similarly intelligible to both native English and Mandarin-accented listeners.

The negative simple effect of Uniqueness indicated that accuracy scores for common-phoneme words were significantly higher than for unique-phoneme words. The lack of interaction between Uniqueness and Listener Accentedness indicates that this effect of Uniqueness is similar across Listener Accentedness scores. The negative simple effect of Speaker Accentedness indicated that as the accentedness of speakers increased, listener accuracy decreased for common-phoneme words. The Speaker Accentedness by Uniqueness interaction indicated that unique-phoneme words were more influenced by Speaker Accentedness than common-phoneme words. The lack of three-way interaction between Uniqueness, Speaker Accentedness, and Listener Accentedness indicates that this effect is similar across Listener Accentedness scores. As shown in Fig. 2, the difference between common-phoneme words and unique-phoneme words widened as Speaker Accentedness increased, with more nativelike speech having a smaller difference between common-phoneme and unique-phoneme word accuracy and more strongly accented speech having a greater difference between common-phoneme and unique-phoneme word accuracy.

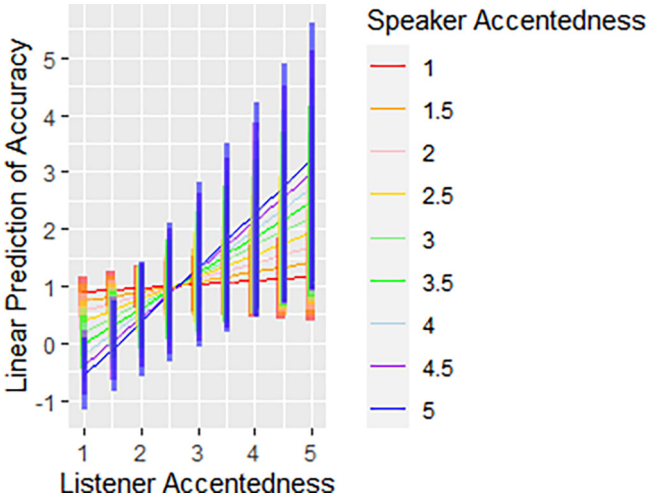

Fig. 1. Estimated marginal accuracy means of Listener Accentedness and Speaker Accentedness for all listeners (native English and L2 Mandarin) with error bars representing the standard errors of the estimated parameters.

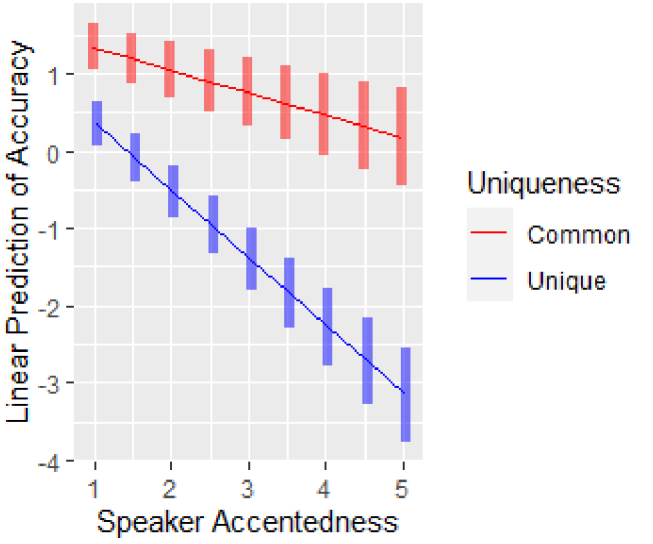

Fig. 2. Estimated marginal accuracy means of common-phoneme and unique-phoneme words by speaker accentedness with error bars representing the standard errors of the estimated parameters.

3.1.2. ISIB-L effects

An additional model was run for only the native Mandarin L2 accented speech in order to identify any ISIB-L effects. The fixed effects included Listener Accentedness, Speaker Accentedness, Uniqueness (common-phoneme word vs unique-phoneme word; reference = common-phoneme word), and their interactions. The model included random intercepts for subject and item.[4] Table 2 below contains the best fitting model results.

The significant interaction between Listener Accentedness and Speaker Accentedness indicated that the slope of Speaker Accentedness changed with Listener Accentedness for common-phoneme words (see Fig. 3). The lack of three-way interaction between Listener Accentedness, Uniqueness, and Speaker Accentedness, indicates that this pattern remained the same for unique-phoneme words. While more nativelike speech (lower Speaker Accentedness score) had similar accuracy regardless of Listener Accentedness, more strongly accented speech (higher Speaker Accentedness score) demonstrated a larger increase in accuracy as Listener Accentedness increased. This indicates that more strongly

[4] Full model syntax: Accuracy for Native Mandarin L2 Accented Speech Tokens ~ Listener Accentedness * Speaker Accentedness * Uniqueness + (1 | Subject) + (1 | Item).

Table 2
Mixed-effects logistic regression analysis with best fit for all listeners' accuracy scores for Mandarin-accented real word speech tokens for the lexical decision task.

Effect	Estimate	Std. Error	*t*	*p*
(Intercept)	1.628	0.168	9.704	<0.001
Speaker Accentedness	−0.482	0.089	−5.437	<0.001
Listener Accentedness	−0.032	0.070	−0.448	0.654
Uniqueness (Unique)	−0.752	0.226	−3.325	<0.001
Speaker Accentedness * Listener Accentedness	0.174	0.069	2.534	0.011
Speaker Accentedness * Uniqueness (Unique)	−0.521	0.121	−4.317	<0.001

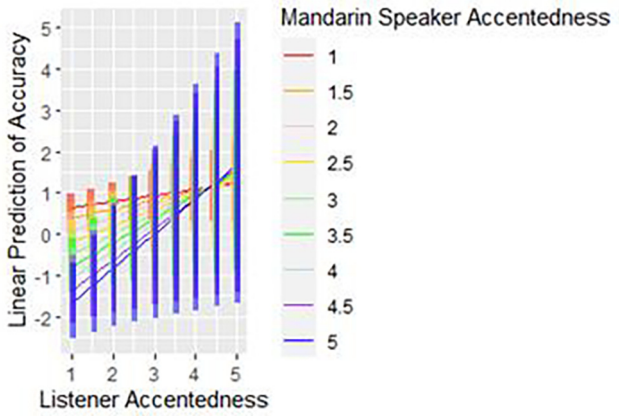

Fig. 3. Estimated marginal accuracy means of Speaker Accentedness and Listener Accentedness for only native Mandarin L2 speech tokens (ISIB-L) with error bars representing the standard errors of the estimated parameters.

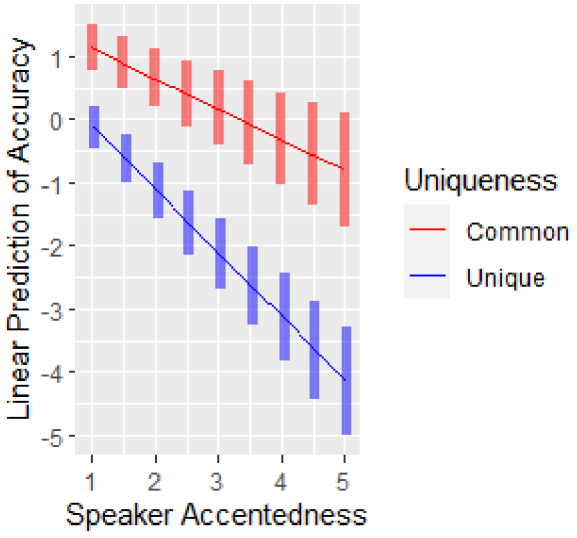

Fig. 4. Estimated marginal accuracy means of common-phoneme and unique-phoneme words by Speaker Accentedness for only native Mandarin speech tokens with error bars representing the standard errors of the estimated parameters.

Mandarin-accented listeners had an advantage over more nativelike English listeners in understanding strongly accented L2 Mandarin - English speech.

The negative simple effect of Uniqueness indicated that accuracy scores for common-phoneme words were significantly higher than accuracy scores for unique-phoneme words. The negative simple effect of Speaker Accentedness indicated that as the accentedness of native Mandarin speakers increased, listener accuracy decreased for common-phoneme words. The Speaker Accentedness by Uniqueness interaction indicated that unique-phoneme words were more influenced by Speaker Accentedness than common-phoneme words (see Fig. 4). This shows that the difference in accuracy scores for common-phoneme and unique-phoneme words is greater for more strongly accented speech compared to more weakly accented speech. The lack of interaction between Listener Accentedness and Uniqueness, as well as the lack of three-way interaction between Listener Accentedness, Uniqueness, and Speaker Accentedness indicates that this relationship between common-phoneme and unique-phoneme words does not differ depending on Listener Accentedness.

3.1.3. ISIB-T effects

An additional mixed-effects logistic regression model was run only for L2 listeners in order to identify any ISIB-T effects. The fixed effects included Speaker Accentedness, Listener Accentedness, Uniqueness (common-phoneme word vs unique-phoneme word; reference = common-phoneme word), and their interactions. The model included random intercepts for subject and item[5]. Table 3 below contains the best fitting model results.

The negative simple effect of Listener Accentedness indicated that as the accentedness of native Mandarin listeners increased, listeners' accuracy decreased for common-phoneme words. The lack of interaction between Uniqueness and Listener Accentedness indicates that this pattern of Listener Accentedness remained the same for unique-phoneme and common-phoneme words.

The significant interaction between Listener Accentedness and Speaker Accentedness indicated that the slope of Speaker Accentedness changed with Listener Accentedness for common-phoneme words (see Fig. 5). The lack of three-way interaction between Uniqueness, Listener Accentedness, and Speaker Accentedness, indicates that this pattern remained the same for unique-phoneme and common-phoneme words. While more nativelike speech (lower Speaker Accentedness score) had similar accuracy regardless of Listener Accentedness, more strongly accented speech (higher Speaker Accentedness score) demonstrated a larger increase in accuracy as Listener Accentedness increased. This showed that native English speech was similarly intelligible for both less accented and more strongly accented Mandarin listeners,

[5] Full model syntax: Accuracy of L2 Listeners ~ Listener Accentedness * Speaker Accentedness * Uniqueness + (1 | Subject) + (1 | Item).

Table 3
Mixed-effects logistic regression analysis with best fit for native Mandarin listeners' accuracy scores for all real word speech tokens for the lexical decision task.

Effect	Estimate	Std. Error	*t*	*p*
(Intercept)	1.574	0.148	10.563	<0.001
Listener Accentedness	−0.255	0.080	−3.186	0.001
Speaker Accentedness	−0.145	0.083	−1.739	0.082
Uniqueness (Unique)	−0.356	0.198	−1.795	0.073
Listener Accentedness * Speaker Accentedness	0.212	0.066	3.223	0.001
Speaker Accentedness * Uniqueness	−0.646	0.115	−5.613	<0.001

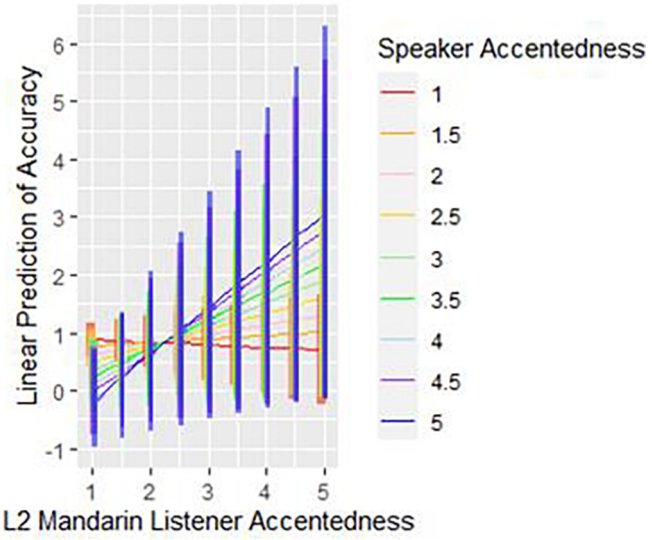

Fig. 5. Estimated marginal accuracy means of Speaker Accentedness by Listener Accentedness for only native Mandarin L2 listeners (ISIB-T) with error bars representing the standard errors of the estimated parameters.

whereas, especially for highly accented Mandarin English speech, more strongly accented native Mandarin listeners had an advantage over less-accented nativelike listeners. More strongly accented native Mandarin listeners had an advantage at understanding more strongly accented speech compared to native speech.

The Speaker Accentedness by Uniqueness interaction indicated that unique-phoneme words were more influenced by Speaker Accentedness than common-phoneme words (see Fig. 6). This signaled that the difference in accuracy scores for common-phoneme and unique-phoneme words is greater for more strongly accented speech compared to more weakly accented speech. The lack of three-way interaction between Uniqueness, Speaker Accentedness, and Listener Accentedness indicates that this pattern was not influenced by Listener Accentedness.

3.2. Reaction times

Mixed-effects linear regression models were conducted on the participants' reaction times. The data were fitted into models using the lmer() function of the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015). Models were backwards fitted using the step() function of the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) in R (R Core Team (2021), 2021). The dependent variable Reaction Time was continuous and included log-transformed values of the reaction time minus the duration of each speech token. Only words classified as correct for the accuracy measure were included in the reaction time models. Three models were run on the data. The first model compared overall differences between reaction times for native and nonnative speakers and listeners, while the other two tested for ISIB-L and ISIB-T effects.[6]

3.2.1. Overall ISIB effects for reaction times

In the first model, the dependent variable was Reaction Time. The fixed effects included Listener L1 (English vs. Mandarin; reference = English), Speaker L1 (English vs. Mandarin; reference = English), Uniqueness (common-phoneme word vs. unique-phoneme word; reference = common-phoneme word), and their interactions. The model included random intercepts for subject and item.[7] No fixed effects remained after backwards fitting the model using the step() function of the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team (2021), 2021). This indicates that no statistically significant amount of variance in reaction times could be explained by knowing the L1 of the speaker, the L1 of the listener, or the uniqueness status of the word.[8,9]

3.2.2. ISIB-L effects for reaction times

A second model was run to identify whether ISIB-L effects were found. The dependent variable was reaction time for only L2 speech tokens. The fixed effects included Speaker Accentedness, Listener L1 (English vs Mandarin; reference = Eng lish), Uniqueness (common-phoneme word vs unique-phoneme word; reference = common-phoneme word), and their interactions. The model included random intercepts for subject and item[10]. The resulting model failed to meet the assumption of homoscedasticity even when power-transformed by the optimal lambda value using the boxcox () function of the EnvStats package (Millard, 2013) of the statistics software R (R Core Team (2021), 2021). A bootstrapping method was therefore used to ensure the output of the model is stable despite not meeting the statistical assumptions of linear regression models. The bootmer () function of the lme4 package

[6] The first five utterances heard by each listener from each native Mandarin speaker were compared to the last five utterances heard by each listener from each native Mandarin speaker to identify whether any rapid accent adaptation occurred. The mean accuracy values decreased from the first five (M = 77.6%, SD = 41.7%) to the last five (M = 69.3%, SD = 46.1%), and mean reaction time values increased slightly from the first five (M = 1105 ms, SD = 427 ms) to the last five (M = 1108 ms, SD = 451 ms), indicating that no strong degree of accent adaptation seemed to occur that would allow participants to adapt to and find more intelligible the speakers over the course of the experiment.

[7] Full model syntax: Reaction Time ~ Listener L1 * Speaker L1 * Uniqueness + (1 | Subject) + (1 | Item).

[8] In order to ensure there was no bias in only assessing reaction times of correct responses, an additional mixed-effects linear regression model was run with a dependent variable of log-transformed reaction time for incorrect responses with the same fixed effects as listed in section 3.2.1 (Listener L1, Speaker L1, Uniqueness, and their interactions) and the same random intercepts (subject and item). Again, no fixed effects remained after backwards fitting the model.

[9] In order to ensure there were no differences in the pattern of reaction times for participants tested in-person and online, an additional mixed-effects linear regression model was run with a dependent variable of log-transformed reaction time for correct responses with fixed effects of Listener L1, Speaker L1, Uniqueness, Online status (online vs in person), and their interactions, along with the same random intercepts (subject and item). Again, no fixed effects remained after backwards fitting the model.

[10] Full model syntax: Reaction Time for L2 Speech Tokens ~ Listener L1 * Speaker Accentedness * Uniqueness + (1 | Subject) + (1 | Item).

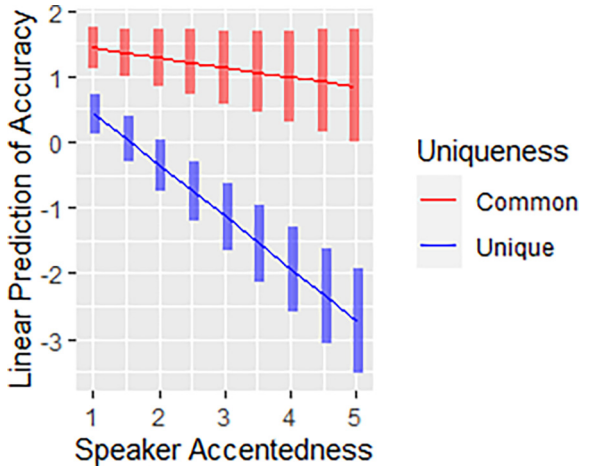

Fig. 6. Estimated marginal accuracy means of common-phoneme and unique-phoneme words by Speaker Accentedness for only native Mandarin listeners with error bars representing the standard errors of the estimated parameters.

Table 4
Bootstrapping values for all listeners' reaction time scores for Mandarin-accented real word speech tokens for the lexical decision task.

Effect	*t*	SE	*p*
(Intercept)	2.684	0.026	<0.05
Speaker Accentedness	0.021	0.011	>0.05
Listener L1	−0.019	0.034	>0.05
Uniqueness	0.022	0.017	>0.05
Speaker Accentedness * Listener L1	0.048	0.015	<0.05
Speaker Accentedness * Uniqueness	0.068	0.017	<0.05
Listener L1 * Uniqueness	-0.011	0.017	>0.05
Speaker Accentedness * Listener L1 * Uniqueness	-0.087	0.024	<0.05

(Bates et al., 2015) was used, and confidence intervals were created using the boot.ci() function from the boot package (Canty & Ripley, 2021) to resample the most complex version of the model 2000 times. Table 4 contains the best-fitting model results.

While the interaction between Speaker Accentedness and Listener L1 indicated that native Mandarin listeners were more impacted by Speaker Accentedness compared to native English listeners in reaction time for common-phoneme words, and the interaction between Speaker Accentedness and Uniqueness showed that unique-phoneme words were more greatly impacted by Speaker Accentedness, the three-way interaction between Speaker Accentedness, Listener L1, and Uniqueness showed the more complex relation among these factors.

The three-way interaction indicated that native English and native Mandarin listeners differ in how common-phoneme versus unique-phoneme words' reaction times are impacted by speaker proficiency. As shown in Fig. 7, while native English listeners were faster at responding to common-phoneme words compared to unique-phoneme words, native Mandarin listeners showed the opposite pattern, with faster reaction times for unique-phoneme words compared to common-phoneme words. Moreover, the unique-phoneme and common-phoneme word reaction times diverged to a greater extent for native English listeners compared to native Mandarin listeners when listening to more strongly accented speech. Native Mandarin listeners showed less divergence in reaction times when listening to unique-phoneme and common-phoneme words as speaker degree of accentedness increased.

3.2.3. ISIB-T effects for reaction times

A third model was run to identify whether ISIB-T effects were found. The dependent variable was reaction times for only L2 listeners. The fixed effects included Speaker L1 (English vs. Mandarin; reference = English), Listener Accentedness, Uniqueness (common-phoneme word vs unique-phoneme word; reference = common-phoneme word), and their interactions. The model included random intercepts for subject and item.[11] No fixed effects remained after backwards fitting the model using the step() function of the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team (2021), 2021), indicating that no statistically significant amount of variance in reaction times could be explained by knowing the L1 of the speaker, accentedness of the listener, or the uniqueness status of the word.

3.3. Type of input analysis

In addition to the regression models created to analyze the accuracy scores and reaction times of participants for the lexical decision task, we explored the role of the type and extent of English that the nonnative listeners were exposed to. Based on Language Background Questionnaire data, a native English input score for Mandarin listeners was calculated based on their native English exposure in daily life. All participants were living in the U.S. at the time of testing. A Mandarin-accented English input score was also calculated based on how often Mandarin participants heard Mandarin-accented English in their daily life.

First, Pearson correlations were conducted between Native English and Mandarin-accented English Input scores and accuracy for native English and Mandarin-accented English speech. The correlation between native Mandarin listeners' Mandarin-accented English Input scores and Accuracy when listening to Mandarin-accented English was not significant ($r(34) = -0.023$, $p = .894$). Moreover, the correlation between native Mandarin listeners' Native English Input scores and Accuracy when listening to native English input was also not significant ($r(34) = 0.251$, $p = .140$). These results suggest that Mandarin participants' accuracy in lexical decision was not affected by type and extent of English input.[12]

Second, we assessed the relationship between input score and accentedness by comparing input between the least accented and most accented halves of native Mandarin listeners. An independent samples t-test comparing Native English Input scores for the least accented half of native Mandarin listeners (M = 21.1) and the most accented half of native Mandarin listeners (M = 8.4) indicated the least accented Mandarin listeners had significantly higher Native English Input

[11] Full model syntax: Reaction Time for L2 Listeners ~ Listener Accentedness * Speaker L1 * Uniqueness + (1 | Subject) + (1 | Item).

[12] It is also possible that using self-report questionnaire information to measure English input does not produce reliable input scores. In order to test this, additional correlations were run between years spent in the US and accuracy when listening to native English speech, age of arrival in the US and accuracy when listening to native English speech, years spent in the US and accuracy when listening to Mandarin-accented speech, and age of arrival in the US and accuracy when listening to Mandarin-accented speech, none of which were significant.

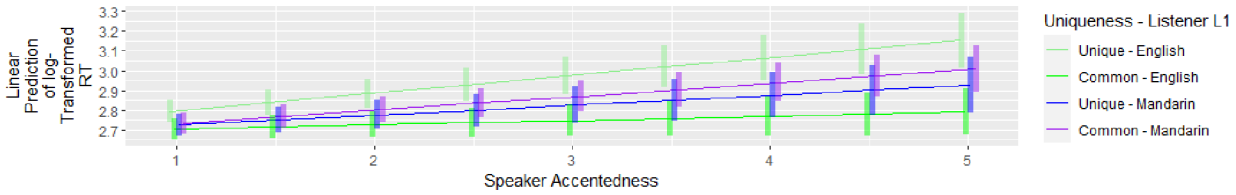

Fig. 7. Estimated marginal reaction time means of common-phoneme and unique-phoneme words for native English and native Mandarin listeners by Speaker Accentedness (y-axis values represent predicted log-transformed reaction times) with error bars representing the standard errors of the estimated parameters.

scores compared to the most accented Mandarin listeners, t (34) = 3.87, p <.001. This higher Native English Input score suggests the least accented Mandarin listeners had more exposure to native English than the most accented Mandarin listeners. An independent samples t-test comparing Mandarin-accented English Input scores for the least accented half of native Mandarin listeners (M = 3.4) and the most accented half of native Mandarin listeners (M = 4.1) indicated the less accented Mandarin listeners did not have a significantly different Mandarin-accented English Input score compared to the more accented Mandarin listeners, t (34) = −1.53, p =.136. This means the least accented Mandarin listeners heard a similar amount of Mandarin-accented English speech on a regular basis as the most accented Mandarin listeners.

4. Acoustic analysis

Acoustic analyses were conducted on five near-minimal pairs of speech stimuli used in the lexical decision task. These pairs included the productions of *to* vs *do, peace* vs *beat, weak* vs *league, tough* vs *love*, and *seat* vs *sit.* Each pair included one common- and one unique-phoneme stimulus, respectively.

4.1. Stimulus measures

The productions were analyzed from the two Native English and four Mandarin-accented English speakers used in the lexical decision task. Absolute and relative VOT, voicing, and burst durations were compared for word-initial voiceless and voiced consonants in *to* and *do* and *piece* and *beat*. Absolute and relative closure voicing duration, closure duration, burst duration, and pre-stop vowel duration measures were also collected for word-final voiceless and voiced stops in *weak* and *league*. Absolute and relative pre-fricative vowel duration and voicing duration of the final voiceless and voiced fricatives in *tough* and *love* were also collected.[13] Finally, F1 and F2 values of the vowels in *seat* and *sit* were measured.

4.2. Findings

Three words (*league, love,* and *sit)* were found to differ most drastically in acoustic realization among native and nonnative speakers, showing differences in final consonant voicing and vowel quality that resulted in non-nativelike productions from the native Mandarin speakers. These three tokens were all classified as unique-phoneme words, indicating that some difficulty may arise when producing less familiar sounds in an L2, a finding similar to that of Han et al. (2011b). The native Mandarin speakers devoiced the final [g] in *league* and produced the vowels in *love* and *sit* with nonnative formant frequency values. As a result of the non-nativelike productions of these tokens, listeners found it difficult to perceive these words as intended, as demonstrated by very poor accuracy scores on the lexical decision task for these three tokens (52.78%, 0%, and 8.33%, respectively). The acoustic analyses add further evidence that words containing phonemes unique to an L2 may be harder for nonnative speakers to produce in a native-like way, and thus may result in perceptual difficulties during the perception of L2 speech. For more detailed results of the acoustic analyses conducted, see Appendix C.

5. Discussion

The current study aimed to provide a more detailed understanding of the various factors that may impact the Interlanguage Speech Intelligibility Benefit (ISIB) and the relationship between native and nonnative speech perception. Crucially, accentedness scores for both the listeners and speakers were collected so that a continuous and complete evaluation of listener and speaker effects could be made. The present study used lexical decision to identify whether evidence of ISIB-L and ISIB-T effects could be found in Mandarin learners of English. We included native speakers of English, as well as native speakers of Mandarin with a weak or strong accent in English. These male and female speakers produced 120 words and nonwords so that half contained phonemes unique to English and half contained phonemes present in both English and Mandarin. Thirty-six native English and 36 native Mandarin listeners completed the lexical decision task. Participants also produced stimuli that were presented to native English judges to rate the phonetic and phonological proficiency of all speakers and listeners in the experiment. This proficiency measure created an accentedness score used in the data analysis to assess the role of speaker and listener proficiency in finding ISIB effects. Thus, native English to weakly accented native Mandarin to strongly accented native Mandarin listeners of English participated in a lexical decision task of native English to weakly accented native Mandarin to strongly accented native Mandarin speech in order to directly test for ISIB effects in accuracy and reaction time measures across both the native-nonnative listener and speaker continua.

Given the range of listener and speaker proficiencies, the goal of the present research was to identify whether native Mandarin listeners would be better than native English listeners at understanding Mandarin-accented English (ISIB-L), as

[13] Relative VOT, voicing, burst, closure, and vowel durations were defined as the relevant measure relative to word duration. Relative closure voicing duration was defined as closure voicing duration relative to the duration of the closure. Relative fricative voicing duration was defined as fricative voicing duration relative to the duration of the fricative.

well as whether native Mandarin listeners would be better at understanding Mandarin-accented English speech compared to native English speech (ISIB-T). Furthermore, whether a phoneme is present in the L1 and L2 (common phoneme) or just the L2 (unique phoneme), as well as the type of English input that participants were exposed to, were studied to identify their contribution to ISIB effects. Additionally, acoustic analyses were conducted to identify what may drive the difficulty in comprehension of Mandarin-accented English.

As expected, native English listeners were more accurate compared to native Mandarin listeners when listening to native English speech, and native English listeners were more accurate when listening to native English speech compared to Mandarin-accented English speech. It is not surprising that native English listeners prefer native English speech and that native English listeners are better than nonnative listeners at understanding native English speech. These results mirror those of earlier researchers, who also found that native English listeners were more accurate than nonnative listeners at native English speech (Xie and Fowler, 2013), and that native English listeners found native English speech more intelligible than nonnative speech (Bent & Bradlow, 2003; Hayes-Harb et al., 2008; Koo, 2018; Sereno et al., 2002; Xie & Fowler, 2013). The native English advantage found in the present study is robustly observed throughout the literature.

While the clear native interlocutor effect is not unexpected, the current study also revealed robust nonnative listener and speaker effects. In terms of ISIB-L effects, which is a benefit for native Mandarin listeners over native English listeners for Mandarin-accented speech, the present data do show higher accuracy scores for native Mandarin listeners compared to native English listeners for Mandarin-accented English speech (Fig. 3). Moreover, this ISIB-L effect is modulated by listener proficiency, from native English listeners to the most accented native Mandarin listeners. Our results additionally show that the ISIB-L is also affected by speaker proficiency. While more nativelike speech had similar accuracy regardless of the accentedness of the listeners, more strongly accented speech (typically indicative of native Mandarin speakers) showed a greater increase in accuracy as the accentedness of the listener increased (typically indicative of native Mandarin listeners). This indicates that less nativelike L2 listeners find strongly accented speech easier to understand than more nativelike L2 listeners, which offers evidence in favor of gradient ISIB-L effects modulated by both speaker and listener accentedness. This finding can lend support to the SLM-r in that the less nativelike L2 listeners may be those who are more likely to have merged L1 and L2 categories, which may give them an advantage over more nativelike listeners with distinct L1 and L2 categories at perceiving foreign-accented speech containing properties of both L1 and L2 categories.

In terms of ISIB-T effects, which is a nonnative listener advantage for nonnative speech over native speech, the present results do show higher accuracy scores for Mandarin-accented English speech compared to native English speech for Mandarin listeners (Fig. 5). Specifically, this ISIB-T effect is modulated by speaker proficiency, from native English to the most Mandarin-accented speech. Our results additionally show that the ISIB-T is also affected by listener proficiency. While perception of nativelike speech was less influenced by listener accentedness, more strongly accented speech showed a greater increase in accuracy as the accentedness of the listener increased. Therefore, strongly accented listeners as compared to less strongly accented listeners are more likely to experience ISIB-T effects, showing greater accuracy for strongly accented speech compared to native and less accented English speech, thus giving evidence in favor of gradient ISIB-T effects modulated by both listener and speaker accentedness. This finding again lends support to the SLM-r in that the less nativelike L2 listeners may have merged L1 and L2 categories that more closely match the more strongly accented speech signal.

Results show clear ISIB-L and ISIB-T effects and demonstrate the dynamic nature of ISIB effects, with both being modulated by speaker and listener proficiency with more striking effects typically occurring at the most extreme ends of accentedness. This could potentially explain some of the less robust results that have been reported in previous literature. Due to the influence of speaker and listener proficiency in finding ISIB effects, previous studies finding no effects may not have chosen either speakers or listeners with higher degrees of accentedness and/or lower proficiency levels.

Our data further shows that the native interlocutor advantage is part of a continuum. Recall that all listeners, both native English and native Mandarin listeners, were assessed for accentedness. Some of the native Mandarin listeners had accentedness scores in the native English accentedness range, with a continuum of accentedness values found across listeners. Interestingly, this also means that some of the native English listeners had accentedness scores in the native Mandarin accentedness range. Our data show that listeners rated as less accented (those more likely to be native English speakers) were more likely to be more accurate compared to listeners rated as more strongly accented, and speakers rated as less accented (those more likely to be native English speakers) were also more likely to be correctly perceived. Noteworthy is the fact that the native and nonnative accentedness ranges overlap, highlighting the continuous nature of these effects.

During the present study, participants completed a lexical decision task, and both accuracy and reaction time data were collected and analyzed. The accuracy data provided the strongest evidence for ISIB effects. While the use of on-line measures such as reaction time may sometimes reveal additional effects that do not show up in accuracy data, the present reaction time results showed few effects as compared to the accuracy data, possibly due to greater variability. The slower processing speed, and thus slower reaction times, expected of nonnative listeners compared to native listeners (McDonald, 2006) may reduce any ISIB effects that may be found since finding a nonnative interlocutor advantage over native interlocutors in reaction times would require significantly faster responses by these nonnative interlocutors. Additionally, reaction times may index multiple aspects of speech perception, including cognitive effort or a speed-accuracy tradeoff. The present study's results hint at a potential speed-accuracy tradeoff for the perception of unique-phoneme words produced by nonnative speakers. Accuracy scores for these words were lower than for common-phoneme words, but nonnative listeners had faster reaction times for these productions.

The present results also show an advantage for words containing phonemes that occur in English and Mandarin com-

pared to words containing phonemes that only occur in English. This difference in comprehension ability between common-phoneme and unique-phoneme words spoken by native Mandarin speakers is indicated by the Speaker Accentedness by Uniqueness interaction in all of the accuracy models. These results indicate that common-phoneme words spoken by speakers with high degrees of accentedness (indicative of native Mandarin speakers) were more accurately perceived compared to unique-phoneme words, and that unique-phoneme words were more negatively impacted by a stronger speaker accentedness than common-phoneme words. These results do not lend support for the prediction of SLM-r that perceptually distinct categories may have an advantage in L2 category learning compared to more perceptually similar categories in the L1 and L2 that may cause difficulty in creating separate categories for the L1 and L2. Additionally, we find that the speech tokens produced in the least nativelike way, and consequently perceived largely incorrectly, were unique-phoneme words. These acoustic analyses add further evidence that words containing phonemes unique to an L2 may be harder for nonnative speakers to produce in a nativelike way, and thus may result in perceptual difficulties during the perception of L2 speech.

While this general preference for common-phoneme words over unique-phoneme words shows up in the accuracy data, this pattern is more complex when the reaction time data are also considered, with sizeable response time differences between native and nonnative listeners. A greater divergence between unique-phoneme and common-phoneme word reaction times when listening to more strongly accented speech was found for native English listeners, but not for native Mandarin listeners. Native English listeners were faster for common-phoneme words compared to unique-phoneme words spoken by strongly accented native Mandarin speakers, indicating an advantage for common-phoneme over unique-phoneme words spoken by nonnative speakers, with the opposite effect for native Mandarin listeners, who were faster for unique-phoneme words compared to common-phoneme words. So overall, nonnative productions of common-phoneme words are more accurate than unique-phoneme words, but for the most accented productions, nonnative listeners are faster to respond to these unique, often mispronounced, productions. Given the reaction time differences between native and non-native listeners, it is also possible that this patterning may result simply from a speed-accuracy tradeoff. In comparison, listeners showed no difference in intelligibility for common-phoneme versus unique-phoneme words in native English (less strongly accented) speech. This result is to be expected because these two groupings carry little significance to native English speakers, and thus, no difference in ability to produce the sounds would be expected. When listening to native Mandarin speech, however, listeners were more accurate for common-phoneme compared to unique-phoneme words. Speakers with stronger foreign accents likely experienced greater difficulty in pronouncing the phonemes unique to English not found in Mandarin, and these words were likely pronounced in the least nativelike way, making it more difficult for native listeners to recognize them as words. Interestingly, there was, however, a three-way interaction between Speaker Accentedness, Uniqueness, and Listener L1 in reaction times for Mandarin-accented speech tokens, indicating that native English listeners experienced greater divergence in reaction times for common-phoneme compared to unique-phoneme words as speaker accentedness increased compared to native Mandarin listeners, who experienced much less divergence. This may indicate that native Mandarin listeners have some advantage over native English listeners in being able to understand the non-nativelike pronunciations of unique-phoneme words spoken by strongly accented native Mandarin speakers because overall their response speed did not differ as greatly when listening to common-phoneme and unique-phoneme words and because they were actually faster responding to unique-phoneme words compared to common-phoneme words. These results are in line with Sereno et al. (2002), which found that although Dutch listeners were equally fast when responding to Dutch-accented unique-phoneme and common-phoneme stimuli, they were slower for native English unique-phoneme stimuli compared to common-phoneme stimuli. This would lend further support for an ISIB-L effect, with native Mandarin listeners experiencing an extra speed advantage compared to native English listeners for unique-phoneme words produced by native Mandarin speakers.

Recent research on ISIB effects has also investigated the nature of language input. Xie and Fowler (2013) examined native Mandarin listeners living in Beijing (hearing mainly Mandarin-accented English input) or the US (hearing primarily native English input), as well as native English listeners living in the US, by comparing their transcription accuracy of English-like nonwords spoken by a native Mandarin speaker and a native English speaker. The focus was on word-final stop voicing, and nonwords were used to ensure results were not influenced by lexical effects. An ISIB-L was found for Mandarin listeners in Beijing and in the US, meaning native Mandarin listeners were better than native English listeners at understanding the Mandarin-accented speaker, regardless of country of residence. Interestingly, an ISIB-T effect was found only for native Mandarin listeners in Beijing, meaning native Mandarin listeners in Beijing were better at comprehending Mandarin-accented English than native English speech, suggesting that the nature of learner input does impact L2 speech comprehension. Li and Mok (2015) also found that learners of Mandarin with the greatest exposure to Mandarin-accented English (those in an immersive environment in Mandarin-speaking Beijing) tended to have higher accuracy scores than those with less Mandarin-accented English exposure (those in the classroom setting in Cantonese-speaking Hong Kong), suggesting that ISIB effects may not solely be a result of a shared phonological interlanguage, but instead may also result from type of L2 input.

The present results reveal both ISIB-L and ISIB-T effects for native Mandarin listeners. The presence of an ISIB-L effect confirms Xie and Fowler's (2015) results. Interestingly, we also found an ISIB-T effect even though all our native Mandarin participants were living in the U.S. at the time of testing. The present ISIB effects do not seem to be strongly affected by the nature and extent of the language input, with no significant correlations between exposure and accuracy on the lexical decision task. Our findings suggest that language proficiency of both the speaker and listener may be the more important factor.

6. Conclusion

The current results show a native English interlocutor advantage, clear evidence of ISIB effects dependent on interlocutor proficiency, and an advantage in comprehension of common-phoneme words over unique-phoneme words for Mandarin-accented English speech. Results from the present experiment demonstrated evidence of gradience in ISIB effects modulated by speaker and listener proficiency, with effects typically occurring at the more extreme ends of speaker and listener accentedness. The present results indicate that the presence of an ISIB is highly dependent on many factors, including listener proficiency, speaker proficiency, phoneme type, and the acoustics of specific speech tokens. These data support the notion of an Interlanguage Speech Intelligibility Benefit for talkers and listeners who share the same first and second language. This comprehension benefit holds for both nonnative listeners over native listeners and for nonnative speech over native speech. The current data clearly introduce listener and speaker proficiency as critical variables in understanding this shared language effect.

CRediT authorship contribution statement

Sheyenne Fishero: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Funding acquisition. **Joan A. Sereno:** Resources, Validation, Writing – review & editing, Supervision, Project administration. **Allard Jongman:** Resources, Validation, Writing – review & editing, Supervision, Project administration.

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.

Acknowledgements

Portions of this study were presented at the 179th and 182nd Meeting of the Acoustical Society of America. We thank the members of the KU Ling 850 seminar for their feedback on this project. This project was partially supported by a research scholarship from the KU Department of Linguistics.

Appendix A

See Table 5.

Table 5
Word and Nonword Stimuli for Lexical Decision Task.

Unique Phoneme Words	Common Phoneme Words	Unique Phoneme Nonwords	Common Phoneme Nonwords
bad [bæd]	come [kʌm]	bav [bæv]	eef [if]
be [bi]	cup [kʌp]	biv [bɪv]	eem [im]
beat [bit]	cut [kʌt]	doov [duv]	een [in]
big [bɪg]	eat [it]	eath [iθ]	foo [fu]
bit [bɪt]	feet [fit]	gee [gi]	fup [fʌp]

Table 5 (*continued*)

Unique Phoneme Words	Common Phoneme Words	Unique Phoneme Nonwords	Common Phoneme Nonwords
but [bʌt]	fun [fʌn]	geen [gin]	fusk [fʌsk]
dad [dæd]	keep [kip]	gith [gɪθ]	keek [kik]
deep [dip]	key [ki]	goov [guv]	keem [kim]
done [dʌn]	me [mi]	ib [ɪb]	keet [kit]
due [du]	mean [min]	ig [ɪg]	leet [lit]
food [fud]	meet [mit]	kag [kæg]	lun [lʌn]
give [gɪv]	new [nu]	loov [luv]	lup [lʌp]
gun [gʌn]	none [nʌn]	meeb [mib]	mees [mis]
had [hæd]	one [wʌn]	mip [mɪp]	moop [mup]
hand [hænd]	piece [pis]	nuth [nʌθ]	mup [mʌp]
him [hɪm]	sea [si]	pud [pʌd]	nuck [nʌk]
league [lig]	seam [sim]	seeb [sib]	nup [nʌp]
list [lɪst]	seat [sit]	thab [θæb]	nuss [nʌs]
live [lɪv]	seen [sin]	theeb [θib]	ook [uk]
love [lʌv]	some [sʌm]	thid [θɪd]	oon [un]
man [mæn]	soon [sun]	thiv [θɪv]	oot [ut]
miss [mɪs]	speak [spik]	thook [θuk]	pook [puk]
month [mʌnθ]	stuff [stʌf]	thuv [θʌv]	poot [put]
need [nid]	sun [sʌn]	vad [væd]	soom [sum]
pass [pæs]	team [tim]	veed [vid]	spus [spʌs]
sit [sɪt]	tough [tʌf]	veen [vin]	teep [tip]
think [θɪŋk]	two [tu]	veep [vip]	toos [tus]
thus [ðʌs]	week [wik]	voo [vu]	tuke [tuk]
who [hu]	what [wʌt]	voog [vug]	tup [tʌp]
with [wɪθ]	you [ju]	voost [vust]	yufe [juf]

Appendix B

See Table 6.

Table 6
Accentedness Judgment Task Stimuli.

Words
king [kɪŋ]
least [list]
leave [liv]
lose [luz]
moon [mun]
must [mʌst]
news [nuz]
sleep [slip]
tea [ti]
young [jʌŋ]

Appendix C. Acoustic analysis of subset of experimental tokens

In a comparison between word-initial voiceless and voiced stops, the [t] in *to* and the [d] in *do,* as well as the the [p] in *peace and the* [b] in *beat* were compared across all six speakers' productions (see Table 7). No clear differences were found between native English and nonnative English speakers in VOT and voicing duration patterns.

In a comparison of the [i]-[ɪ] contrast for the vowel [i], all six participants produced F1 between 300 and 393 Hz and F2 between 2385 and 2947 Hz, resulting in F2-F1 values between 1997 and 2600 Hz (see Table 8). For the vowel [ɪ], all five speakers other than one native Mandarin speaker produced the vowel with an F2-F1 difference of 1533 Hz or less. In contrast, one native Mandarin female speaker (NMF2) classified as the most strongly accented speaker out of the six lexical decision task speakers (rated 3.76 out of 5 by native English judges) produced the [ɪ] in *sit* in a very [i]-like way, with a

Table 7
Acoustic Measures for *to*/*do* and *peace*/*beat* (relative VOT defined as VOT in relation to word duration, relative voicing duration defined as voicing in relation to word duration, and relative burst duration defined as burst in relation to word duration).

	Native English (NEF)	Native English (NEM)	Native Mandarin (NMF1 – weakly accented)	Native Mandarin (NMM1 – weakly accented)	Native Mandarin (NMF2 – strongly accented)	Native Mandarin (NMM2 – strongly accented)
Relative VOT of [t] in *to*	34.7%	25.3%	31.7%	28.9%	29.2%	19.9%
Relative VOT of [d] in *do*	29.7%	8.4%	9.0%	22.2%	4.3%	18.9%
Relative voicing duration of [t] in *to*	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Relative voicing duration of [d] in *do*	29.7%	0.0%	0.0%	22.2%	0.0%	18.9%
Relative burst duration of [t] in *to*	0.6%	0.6%	0.5%	0.4%	0.5%	0.8%
Relative burst duration of [d] in *do*	0.5%	0.4%	0.5%	0.3%	1.2%	0.5%
Relative VOT of [p] in *peace*	16.8%	10.8%	15.2%	23.3%	15.3%	12.9%
Relative VOT of [b] in *beat*	2.3%	2.5%	2.9%	2.1%	1.5%	3.4%
Relative voicing duration of [p] in *peace*	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Relative voicing duration of [b] in *beat*	0.6%	2.5%	0.0%	0.0%	1.5%	0.0%
Relative burst duration of [p] in *peace*	0.2%	0.2%	0.2%	0.1%	0.4%	0.4%
Relative burst duration of [b] in *beat*	0.6%	0.5%	0.7%	0.5%	0.8%	0.6%

Table 8
Acoustic Measures for *seat*/*sit* (relative vowel duration defined as vowel duration in relation to word duration).

	Native English (NEF)	Native English (NEM)	Native Mandarin (NMF1 – weakly accented)	Native Mandarin (NMM1 – weakly accented)	Native Mandarin (NMF2 – strongly accented)	Native Mandarin (NMM2 – strongly accented)
F1 in *seat* (Hz)	393	333	393	300	347	393
F2 in *seat* (Hz)	2715	2385	2901	2529	2947	2390
F1 in *sit* (Hz)	448	486	532	440	393	486
F2 in *sit* (Hz)	1974	1925	2065	1786	2854	1786
Relative vowel duration of [i] in *seat*	25.8%	28.1%	23.8%	30.6%	24.9%	41.0%
Relative vowel duration of [ɪ] in *sit*	41.9%	24.6%	13.6%	20.8%	38.1%	25.6%

lower-than-expected F1 of 393 Hz and a higher-than-expected F2 of 2854 Hz, creating an F2-F1 difference of 2461 Hz. This particular speaker's [i] formant frequency values for F1 and F2 in the word *seat* were 347 and 2947 Hz, respectively, meaning her [ɪ] production closely resembled her [i] production. In comparison, all other speakers, regardless of L1, clearly distinguished their [i] and [ɪ] productions as seen in Table 8 below. Accuracy in the lexical decision task for the non-nativelike *sit* token drastically differed from accuracy of the other *sit* tokens in the experiment. Lexical decision task accuracy for this particular token was 8.33%, whereas accuracy for all other *sit* tokens was 88.33%, demonstrating that the non-nativelikeness of this native Mandarin speaker's production led to difficulty in perception of the ambiguous token during the lexical decision task.

In a comparison between word-final voiced and voiceless stops, the [k] in *weak* and the [g] in *league* were compared (see Table 9). All six speakers produced the vowel in *weak* with a much shorter relative and absolute duration compared to the vowel in *league.* One measure that indicated divergence between native English and most native Mandarin speakers was the voicing of the final obstruent [g] in *league*. While all six speakers produced the final [k] in *weak* with very little closure voicing (<19% of the closure), differences arose among speaker groups for the final [g] in *league*. While both native English participants produced the final [g] in *league* with 100% of the closure voiced, only one native Mandarin speaker (NMM2) shared this pattern. The three other native Mandarin speakers (NMF1, NMM1, and NMF2) produced this [g] with 27.0%, 23.7%, and 5.9% voicing, respectively, indicating a pattern of devoicing of this final voiced [g] not found in the native English productions. This could result from the fact that Mandarin does not allow final stops and does not contain a voiced [g], and thus, producing this English word that breaks Mandarin phonotactic rules may be difficult for native Mandarin speakers. Accuracy during the lexical decision task for the fully voiced native and nativelike tokens was 77.78%, whereas accuracy for the largely devoiced non-nativelike tokens was

Table 9
Acoustic Measures for *weak*/*league* (relative pre-stop vowel duration defined as vowel duration in relation to word duration, relative closure voicing duration defined as stop voicing duration in relation to stop closure duration, relative closure duration defined as closure duration in relation to word duration, and relative burst duration defined as burst duration in relation to word duration).

	Native English (NEF)	Native English (NEM)	Native Mandarin (NMF1 – weakly accented)	Native Mandarin (NMM1 – weakly accented)	Native Mandarin (NMF2 – strongly accented)	Native Mandarin (NMM2 – strongly accented)
Relative pre-stop vowel duration of *weak*	29.4%	21.1%	28.8%	30.4%	31.0%	28.8%
Relative pre-stop vowel duration of *league*	37.3%	50.9%	52.5%	48.4%	44.5%	37.4%
Relative closure voicing duration of [k] in *weak*	18.9%	0%	15.5%	0%	9.4%	18.7%
Relative closure voicing duration of [g] in *league*	100%	100%	27.0%	23.7%	5.9%	100%
Relative closure duration of [k] in *weak*	25.6%	34.0%	33.8%	22.3%	33.6%	17.4%
Relative closure duration of [g] in *league*	19.4%	26.0%	25.8%	20.3%	38.6%	23.9%
Relative burst duration of [k] in *weak*	1.1%	0.4%	0.6%	0.9%	2.0%	0.9%
Relative burst duration of [g] in *league*	1.5%	0.5%	1.1%	0.7%	1.1%	0.5%

Table 10
Acoustic Measures for *tough*/*love* (relative pre-fricative vowel duration defined as vowel duration in relation to word duration, and relative voicing duration defined as fricative voicing duration in relation to fricative duration).

	Native English (NEF)	Native English (NEM)	Native Mandarin (NMF1 – weakly accented)	Native Mandarin (NMM1 – weakly accented)	Native Mandarin (NMF2 – strongly accented)	Native Mandarin (NMM2 – strongly accented)
Relative pre-fricative vowel duration of *tough*	24.8%	31.5%	16.5%	38.8%	28.9%	37.0%
Relative pre-fricative vowel duration of *love*	43.3%	42.0%	41.2%	56.4%	46.2%	48.6%
Relative voicing duration of [f] in *tough*	6.1%	0%	0%	0%	0%	0%
Relative voicing duration of [v] in *love*	18.0%	0%	53.9%	62.1%	0%	100%
F1 in *love* (Hz)	703	606	752	630	923	728
F2 in *love* (Hz)	1557	1289	1337	1264	1386	1313

52.78%, indicating that the tokens not pronounced in a native-like way were likely more difficult to process in the lexical decision task.

In a comparison between word-final voiceless and voiced fricatives, the [f] in *tough* and the [v] in *love* were compared (see Table 10). The vowel duration was shorter for the vowel in *tough* compared to *love* for all speakers. All six speakers similarly produced [f] with little to no voicing. In contrast, speakers differed in their voicing patterns of the final fricative in *love*. The two native English speakers (NEF and NEM) in the study produced the final [v] in *love* with 18.0% and 0% of the final fricative voiced. Only one native Mandarin speaker (NMF2) similarly produced this fricative as devoiced. In contrast, the other three native Mandarin speakers (NMF1, NMM1, and NMM2) in the present study produced the final [v] in *love* as voiced for the majority of the fricative duration (53.9%, 62.1%, and 100%). While voiced fricatives are typically expected to be realized with some degree of voicing in word-final position in English, word-final fricative devoicing like what was found in the present study for both native English speakers and one native Mandarin speaker has previously been documented in English (Docherty, 1992). Lexical decision task accuracy for the native English speakers' devoiced [v] tokens was 100%, but accuracy when perceiving the native Mandarin speaker (NMF2) who devoiced her final [v] was at 0%, despite the fact that the final [v] devoicing found for this token was the most nativelike in terms of being devoiced. In comparison, accuracy in perception for the native Mandarin speakers who maintained voicing in their final [v] to a much greater degree was 87.88%. A fully devoiced final [v] in the word *love* creates a nonword *luff*, which could explain the 0% accuracy in the lexical decision task for the native Mandarin speaker's devoiced [v] token, but because [v] devoicing did not lower accuracy scores for native English speech, [v] devoicing does not seem to be driving this low accuracy rate. Upon inspection of the first two formants of the vowel in all tokens of *love*, it seems that the speaker (NMF2) whose token received 0% accuracy also had an extremely high F1 (923 Hz) compared to all other speakers (ranging from 606 to 752 Hz), indicating that her vowel quality was also not nativelike. Therefore, while there were acoustic differences across tokens of *love* in terms of relative voicing duration of the final fricative, this particular cue is not what seems to be driving differences in accuracy for different speakers' tokens.

References

Algethami, G., Ingram, J., & Nguyen, T. (2011). The interlanguage speech intelligibility benefit: The case of Arabic-accented English. In J. Levis & K. LeVelle (Eds.). In *Proceedings of the 2nd Pronunciation in Second Language Learning and Teaching Conference,* Sept. 2010. (pp. 30–42), Ames, IA: Iowa State University.

Anwyl-Irvine, A. L., Massonié, J., Flitton, A., Kirkham, N. Z., & Evershed, J. K. (2019). Gorilla in our midst: An online behavioral experiment builder. *Behavioral Research Methods, 52*, 388–407.

Baese-Berk, M. (2009). Perceptual adaptation to foreign accented speech. *Journal of the Acoustical Society of America, 125*, 2765.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America, 114*, 1600–1610.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707–729.

Boersma, P. & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.0.49. Retrieved from http://www.praat.org/.

Canty, A., & Ripley, B. (2021). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-27. https://CRAN.R-project.org/package=boot.

Chen, H. C. (2015). Acoustic analyses and intelligibility assessments of timing patterns among Chinese English learners with different dialect backgrounds. *Journal of Psycholinguistic Research, 44*(6), 749–773.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116*(6), 3647–3658.

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Retrieved from https://corpus.byu.edu/coca/.

Docherty, G. (1992). *The timing of voicing in British English obstruents (Netherlands phonetic archives* (9) Berlin; New York: Foris Publications.

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics, 15*(1), 47–65.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). York Press.

Flege, J. E., & Bohn, O.-S. (2021). The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second language speech learning* (pp. 3–83). Cambridge University Press.

Han, J.-I., Choi, T.-H., Lim, I.-J., & Lee, J.-K. (2011a). The interlanguage speech intelligibility benefit for Korean learners of English: Perception of English front vowels. *Korea Journal of English Language and Linguistics, 11*(2), 385–413.

Han, J.-I., Choi, T.-H., Lim, I.-J., & Lee, J.-K. (2011b). The interlanguage speech intelligibility benefit for Korean learners of English: Production of English front vowels. *Journal of the Korean Society of Speech Sciences, 3*(2), 53–61.

Hayes-Harb, R., Smith, B. L., Bent, T., & Bradlow, A. R. (2008). The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *Journal of Phonetics, 36*(4), 664–679.

Imai, S., Walley, A. C., & Flege, J. E. (2005). Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. *Journal of the Acoustical Society of America, 117*(2), 896–907.

Koo, S.-H. (2018). The combined effect of listeners' language background and L2 English teaching background on mutual intelligibility: A mixed-methods approach. *The SNU Journal of Education Research, 27*(2), 1–28.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1–26.

Lee, W.-S., & Zee, E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association: Illustrations of the IPA, 33*(1), 109–112.

Li, G., & Mok, P.K. (2015). Interlanguage speech intelligibility benefit for Mandarin: Is it from shared phonological knowledge or exposure to accented speech. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow.

McDonald, J. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language, 55*(3), 381–401.

McLaughlin, D. J., & Van Engen, K. J. (2020). Task evoked pupil response for accurately recognized accented speech. *The Journal of the Acoustical Society of America, 147*, EL151.

Millard, S. P. (2013). *EnvStats: An R Package for Environmental Statistics*. Springer.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*(1), 111–131.

Palan, S., & Schitter, C. (2018). Prolific.ac – A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27.

Perception Research Systems (2007). *Paradigm Stimulus Presentation*. Retrieved from http://www.paradigmexperiments.com.

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

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*, 209–231.

Sereno, J., McCall, J., Jongman, A., Dijkstra, T., & van Heuven, W. (2002). Perception of native and accented speech by native and non-native listeners. *Abstract of the 43rd Annual Meeting of the Psychonomic Society*, *7*, 59.

So, C. K., & Attina, V. (2014). Cross-language perception of Cantonese vowels spoken by native and non-native speakers. *Journal of Psycholinguistic Research, 43*(5), 611–630.

Stibbard, R. M., & Lee, J. (2006). Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis. *Journal of the Acoustical Society of America, 120*(1), 433–442.

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. [Data file]. Available from http://www.iphod.com.

Van Wijngaarden, S. J. (2001). Intelligibility of native and non-native Dutch speech. *Speech Communication, 35*(1–2), 103–113.

Xie, X., & Fowler, C. A. (2013). Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *Journal of Phonetics, 41*(5), 369–378.

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *Journal of the Acoustical Society of America, 143*(4), 2013–2031.

123Apps LLC (2021). *Online Voice Recorder*. https://online-voice-recorder.com/.