Second Language Speech Intelligibility Revisited: Differential Roles of Phonological Accuracy, Visual Speech, and Iconic Gesture

PAGE WHEELER[1] AND KAZUYA SAITO[2]
[1] *University College London, Institute of Education, 20 Bedford Way, United Kingdom, WC1H 0AL*
E-mail: page.wheeler.18@ucl.ac.uk
[2] *University College London, Institute of Education, 20 Bedford Way, United Kingdom, WC1H 0AL*
E-mail: k.saito@ucl.ac.uk

Although intelligibility is a core concept in second language (L2) speech assessment and teaching research, the vast majority of previous work relies on audio-only stimuli. The current study set out to examine how linguistic and visual information jointly interact to determine the degree of speech intelligibility. Both first language (L1) and L2 English listeners were presented with stimuli that varied along 3 factors (vowel error, visual speech, and iconic gesture) and completed an orthographic transcription task. Results revealed that iconic gesture significantly increased all the listeners' intelligibility scores when speech contained vowel errors. When speech did not contain errors, gesture increased intelligibility for L2 listeners but not L1 listeners. Visual speech had no significant effect on intelligibility in either listener group. Vowel error reduced intelligibility by approximately 20–30% for both L1 and L2 listeners. Findings suggest that visual modalities, especially gestures, have the potential to significantly affect the intelligibility of speech containing phonological errors.

Keywords: intelligibility; pronunciation; visual speech; gesture

THERE IS NOW LITTLE DEBATE IN THE field of second language (L2) speaking and pronunciation research that the primary goal of instruction should be intelligibility rather than native-likeness (Levis, 2018). Instead of trying to erase all deviations from standard speech, priority should be given to language errors that are most likely to impede effective communication. Finding out what these language errors are is one of the main aims of intelligibility research. In the last several decades, a number of areas have been examined, including segmentals (Bent et al., 2007), word stress (Field, 2005), rhythm (Tajima et al., 1997), and intonation (Winters & O'Brien, 2013), as well as features beyond pronunciation, such as grammar and lexis (Munro & Derwing, 1995a).

The vast majority of L2 intelligibility research relies on audio-only stimuli. We know from first language (L1) speech perception studies that visual speech (the movements of the lips, mouth, tongue, and teeth) as well as gesture (expressive movements of the arms and hands) have the power to significantly affect comprehension, especially in adverse listening conditions (Hostetter, 2011; Peelle & Sommers, 2015). The current study attempts to expand L2 intelligibility research by incorporating these visual modalities into its design. The first aim of the study is to examine how L1 English listeners' intelligibility scores are

The Modern Language Journal, 0, 0, (2022)
DOI: 10.1111/modl.12779
0026-7902/22/1–21 $1.50/0

influenced by the presence or absence of three factors: vowel error, visible speech, and iconic gesture. The second aim is to see if results differ for L2 English (L1 Mandarin) listeners. A third subsidiary aim is to explore the beliefs, preferences, and behaviors of L1 and L2 English users regarding the effect of visual cues on comprehension.

BACKGROUND

Second Language Intelligibility

Intelligibility is defined as the degree to which a speaker is actually understood. At its narrowest, it refers to listeners' ability to decode individual phonemes or words, and at its broadest, their ability to comprehend a long stretch of discourse; Munro and Derwing (2015) referred to these, respectively, as "local" and "global" intelligibility (p. 381). Intelligibility is most often operationalized as the percentage of words (or key words) correctly transcribed by listeners. A variety of other measurements also appear in the literature, including cloze tests (Riseborough, 1981), sentence verification (Munro & Derwing, 1995b), forced-choice identification (Tajima et al., 1997), comprehension questions (Hahn, 2004), focused interviews (Zielinski, 2008), written summaries (Hahn, 2004), speech reception thresholds (Quené & van Delft, 2010), and transcriptions of nonsense sentences (Kang et al., 2020). Ostensibly, all of these tasks are measuring the same construct, although recent evidence suggests that this may not be the case (Kang et al., 2018). If different measurements are in fact tapping into different constructs, it may not be possible to directly compare results. In addition, regardless of measurement type, it is difficult to ensure that listeners are paying full attention during listening and transcription tasks (Derwing & Munro, 2015). For this reason, it is recommended that researchers adopt multiple measures of L2 intelligibility.

L2 speech intelligibility research has explored a wide range of different linguistic features, but the current study is particularly concerned with the influence of segmentals—one of the areas that has received a fair amount of attention from researchers. Bent et al. (2007) examined recorded sentences from 15 Mandarin speakers and coded all segmental errors. Intelligibility scores for each speaker were calculated through a speech-in-noise transcription task. Analysis revealed that the overall accuracy of vowels and the accuracy of word-initial consonants—but not other consonants—were correlated to intelligibility. In Zielinski's (2008) study, three L1 listeners transcribed the extemporaneous speech of three L2 speakers from different language backgrounds (Korean, Mandarin, and Vietnamese). During the transcription session, listeners shared their thinking process and commented on any difficulties they experienced. Results revealed that intelligibility was reduced by speakers' production of nonstandard segments and nonstandard syllable stress (or both of these in combination). The most misleading nonstandard segments were those in strong syllables, especially nonstandard vowels and nonstandard initial consonants within these syllables. Available research thus suggests that certain segmental substitution errors significantly affect the intelligibility of L2 speech.

One noteworthy methodological issue with these L2 intelligibility studies is a reliance on audio-only stimuli. In some respects, this choice is justified. Real-life listening is sometimes confined to an auditory signal (phone conversations, radio broadcasts, etc.), and focusing on audio eliminates a number of potentially confounding variables—such as the ethnicity of the speaker—that a visual modality introduces. However, given that the majority of daily spoken interaction takes place face to face, investigation into the effects of visual information is also warranted. Drawing on L1 speech perception studies, we aim to expand the scope of the L2 speech intelligibility paradigm by introducing an audiovisual framework for the investigation of speech containing linguistic errors. In the current study, we feature two well-researched modalities in particular: visual speech and iconic gesture.

Visual Speech

Visual speech is defined as "information available from seeing a speaker's mouth, including the lips, tongue, and teeth" (Peelle & Sommers, 2015, p. 170). These visual cues provide information about place of articulation that can help distinguish between certain segmentals (e.g., /b/ vs. /d/), thereby constraining lexical competition when the auditory signal is ambiguous (Tye–Murray et al., 2007). For example, if a listener is unsure whether they have heard "bad" or "dad," seeing a speaker's mouth (with either closed or opened lips) would rule out one or the other.

There is robust evidence that seeing visual speech increases intelligibility in L1-speaker–L1-listener interactions (see Peelle & Sommers, 2015, for an overview). Visual speech is especially beneficial in conditions of noise (Ross et al., 2006;

Sumby & Pollack, 1954). It is less beneficial in quiet conditions (at least for L1 listeners), as intelligibility scores tend to reach ceiling levels based on auditory information alone. For L2 listeners, however, visual speech has been shown to increase intelligibility even in quiet conditions (Hazan et al., 2006; Navarra & Soto–Faraco, 2007).

Though few in number, some scholars have begun to examine the role of visual speech in L1 listeners' understanding of foreign-accented speech. In Yi et al.'s (2013) study, participants transcribed L1 English speech as well as Korean-accented English speech presented in two modalities: audio-only and audiovisual with the speaker's face visible. Results revealed that the presence of visual speech increased the intelligibility of all speakers and the degree of this benefit was greater for L1 speech than Korean-accented speech. Kawase et al. (2014) focused on a set of consonants in Japanese-accented English speech. Results showed that overall intelligibility was higher in the audiovisual condition than in the audio condition, but the degree of benefit was greater for certain consonants that are the same or similar in Japanese and English. Notably, the intelligibility of /ɹ/ was actually lower in the audiovisual condition. The authors attributed this to inaccurate articulatory configurations; unlike L1 speakers, the Japanese speakers tended not to round their lips when pronouncing this sound, which may have misled L1 listeners. This finding suggests that the interaction between visual speech and the intelligibility of foreign-accented speakers is complex. In comparison to L1 speakers, the mouth and lip movements of foreign-accented speakers are more likely to contain errors or nonstandard articulations that might reduce rather than enhance the intelligibility of certain sounds and words. Despite these occasional articulation errors, however, it appears that access to visual speech likely increases the global intelligibility of foreign-accented speech.

An extra consideration in this area of research is the possibility of cultural differences in the relative weighting of auditory and visual information. To investigate this issue, researchers measure a McGurk effect (McGurk & MacDonald, 1976), which occurs when an incongruent audiovisual stimulus is presented to a listener (e.g., audio of someone saying "ba" and a video of someone mouthing "ga") and the listener reports hearing a third sound (in this case, "da"). This effect illustrates how the auditory signal and visual signal are integrated in speech processing. However, not all individuals experience the effect and instead will simply perceive the auditory signal ("ba") or, less often, the visual signal ("ga"). Cultural differences in the likelihood of observing the McGurk effect are interpreted as reflecting differences in auditory–visual weighting. Magnotti et al. (2015) found similar frequencies of the effect in English speakers and Mandarin speakers (two groups of particular importance to the current study), indicating little to no difference in visual weighting. Notably, the study found a huge amount of variation across participants (0–100%), indicating a high degree of individual variability in audiovisual speech integration.

Iconic Gesture

Kendon (2004) defined gestures as "actions that have the features of manifest deliberate expressiveness" (p. 15)—that is, bodily movements (usually of the hands and arms) that are at least somewhat voluntary and serve a communicative purpose. For the purposes of this study, we are primarily concerned with hand movements that co-occur with speech. McNeill's (1992) classification system outlines four kinds of co-speech gesture: iconic, metaphoric, deictic, and beat. Iconic gestures concretely represent the attributes, movements, or spatial relationships of objects or people. An example of an iconic gesture would be someone raising their hand and then jerking it down while saying, "The box fell to the ground." The gesture iconically resembles the movement of the box. This gesture provides similar semantic information to what is conveyed by the word "fell," but might also give additional information, such as the force of the fall or the manner in which it happened. Importantly, the gesture on its own (without the accompanying speech) is not entirely transparent, unlike an emblematic gesture. Metaphoric gestures also convey semantic information, but of abstract concepts (e.g., clenching one's fist to represent the feeling of anger). Deictic gestures are pointing gestures that refer to locations, objects, or ideas, which may or may not be physically present. Finally, beat gestures are movements of the hand that "beat" time with the rhythm of speech, placing emphasis on certain words or phrases without conveying any semantic information. Unlike emblems, such as the thumbs-up or OK sign, which are highly conventionalized and often stand on their own, these four types of co-speech gesture are not culturally specific in form, although there are some subtle crosslinguistic and crosscultural differences (Kita, 2009). Of the four types, iconic gestures are by far the most researched and are also the main focus of the current study.

In the context of L1 users, research suggests that iconic gestures significantly affect the intelligibility of speech. Studies that compare conditions with and without iconic gestures repeatedly find that participants perform better in the gesture condition, especially in noise (Beattie & Shovelton, 1999; Riseborough, 1981; Rogers, 1978). In a more recent study, Drijvers and Özyürek (2017) investigated both visual speech and iconic gesture in a single design. Participants transcribed words presented in several different modalities (audio-only vs. audio with visual speech vs. audio with visual speech and iconic gesture) and across several different levels of noise-vocoding, that is, a method of degrading the clarity of speech in a way that simulates speech perception with a cochlear implant. Results showed that the visual-speech condition was more intelligible than audio-only and that the visual-speech-with-gesture condition was the most intelligible. Benefits were larger under a moderate level of noise-vocoding than under severe levels.

Recently, scholars have also begun to examine the role of iconic gestures in communication involving L2 users. In Sueyoshi and Hardison's (2005) study, 42 ESL learners at different levels of English proficiency were split into three groups and listened to a recorded lecture in three different conditions: audio only, audiovisual with face, and audiovisual with face and gestures. As a percentage of total gestures, 38% were beat, 31% were iconic, 23% were metaphoric, and 8% were deictic. Results of a multiple-choice comprehension test showed that for the low-proficiency listeners, the gesture condition was most intelligible, while for the high-proficiency listeners, the face condition was best. For all proficiency levels, the audio-only condition resulted in the lowest comprehension scores. These results suggest that the benefit of gesture for L2 comprehension is moderated by proficiency level. In a more recent study, Dahl and Ludvigsen (2014) found a facilitative effect of gesture on the ability of seventh- and eighth-grade L2 listeners to comprehend a verbal description of a cartoon image. Drijvers and Özyürek (2020) investigated the effect of both visual speech and iconic gesture on L2 listeners and compared results to those previously found for L1 listeners (2017). While L2 listeners benefited from the presence of iconic gestures, they did so to a lesser degree than L1 listeners. When speech was severely degraded, L2 listeners, unlike L1 listeners, did not benefit from visual speech.

In addition to studies that focus on gesture in particular, there are studies that have investigated the use of visuals more generally in the context of L2 listening assessment. Some researchers (Parry & Meredith, 1984; Wagner, 2010b, 2013) have found that participants who see a video perform better on listening exams than those who have access to the audio only. Others (Batty, 2015) have found little to no difference in test scores between these conditions. Based on the verbal reports of L2 listeners after taking a video-listening exam, Wagner (2008) concluded that listeners varied in their ability to use visual information as an aid to comprehension. Overall, findings in this area suggest that the presence of visual information (including gesture, visual speech, and other nonverbal behavior) can often facilitate comprehension but may not be beneficial in all listening contexts or for all individuals. More research that isolates the effect of gesture is needed in order to better understand this factor, especially as it relates to foreign-accented speech.

MOTIVATION FOR CURRENT STUDY

Despite the fact that most spoken interaction involving L2 users occurs face to face, there is very little research that examines the effect of linguistic error alongside the effect of visual cues. The few studies that do include a visual modality have focused solely on facial cues. To address this gap in the literature, the current study aims to investigate the question of if and to what degree visual speech and iconic gesture affect listeners' comprehension of speech containing phonological deviations. Such an investigation could shed light on speaker behaviors (both verbal and nonverbal) that facilitate or impede effective communication, with subsequent implications for teaching and assessment. We have adopted a design similar to that of Drijvers & Özyürek (2017, 2020) in order to investigate both visual speech and iconic gesture within a single experiment.

For this preliminary study, phonological deviations were operationalized on a segmental level. Similar to the methodology in Banks et al. (2015), we created segmental deviations by shifting vowel sounds within these words, creating a "novel accent." Following Drijvers & Özyürek (2017, 2020), we focused on the intelligibility of common action verbs, as these are highly conducive to iconic gesture pairings. These words were presented in isolation rather than within extended speech in order to isolate and control the variables under investigation. As in other intelligibility research that utilizes single words or meaningless syllables (Field, 2005; Hazan et al., 2006; Kawase et al., 2014; Ross et al., 2006), this allowed us to eliminate a number of potentially confounding variables, including suprasegmental pronunciation features, lexicogrammatical accuracy and complexity, and contextual verbal information.

Two main research questions were formulated:

RQ1. How do visual speech, iconic gestures, and vowel errors affect the intelligibility of speech for L1 listeners?
RQ2. Are L2 listeners affected by these factors in the same way and to the same degree as L1 listeners?

A subsidiary aim was to explore listeners' attitudes regarding the potential effect of visual cues on comprehension. These attitudes are yet another factor that may influence intelligibility results and thus bear consideration. A third RQ addresses this aim:

RQ3. What are L1 and L2 listeners' beliefs, preferences, and behaviors regarding the use of visual cues as an aid to comprehension?

In the experiment, L1 and L2 English listeners watched the same set of 60 stimuli in six conditions (see Figure 1) and completed an intelligibility task. Stimuli consisted of common verbs pronounced with or without vowel errors in three different modalities:

1. Audio only (video with the speaker's mouth obscured)
2. Visual speech (audio + visual speech)
3. Gesture (audio + visual speech + iconic gesture)

Listeners additionally completed the Visual Cue Preference Questionnaire (VCPQ), adapted from Sueyoshi & Hardison (2005), which provided information about their beliefs and preferences regarding facial cues and gesture. Listener background information was collected and analyzed in order to help explain similarities and differences in L1 and L2 listener intelligibility scores.

METHOD

Listeners

L1 Listeners. L1 listeners were 10 L1 users of English (5 females, 5 males, $M_{\text{age}} = 36$, $SD = 11.16$, with 2 listeners choosing not to reveal their age). Listeners were postgraduate students at a university in London, most of whom were British, with one listener from the United States. All L1 listeners had English language teaching experience or some knowledge of English phonology (most had both). Almost all these listeners had studied an L2, reaching various levels of proficiency.

L2 Listeners. L2 listeners were 22 L2 users of English from the same UK university (21 women and 1 man, $M_{age} = 26.64$, $SD = 2.63$). Compared to the L1 listeners, there was a greater percentage of women and a lower mean age. These listeners were L1 users of Mandarin from China ($n = 20$) and Taiwan ($n = 2$). All L2 listeners were highly proficient in English, scoring a minimum of 7 out of 9 on the International English Language Testing System (IELTS) or 100 out of 120 on the Test of English as a Foreign Language (TOEFL). They had lived in an English-speaking country for less than 18 months and reported speaking English at least 50% of the time at school or at home.

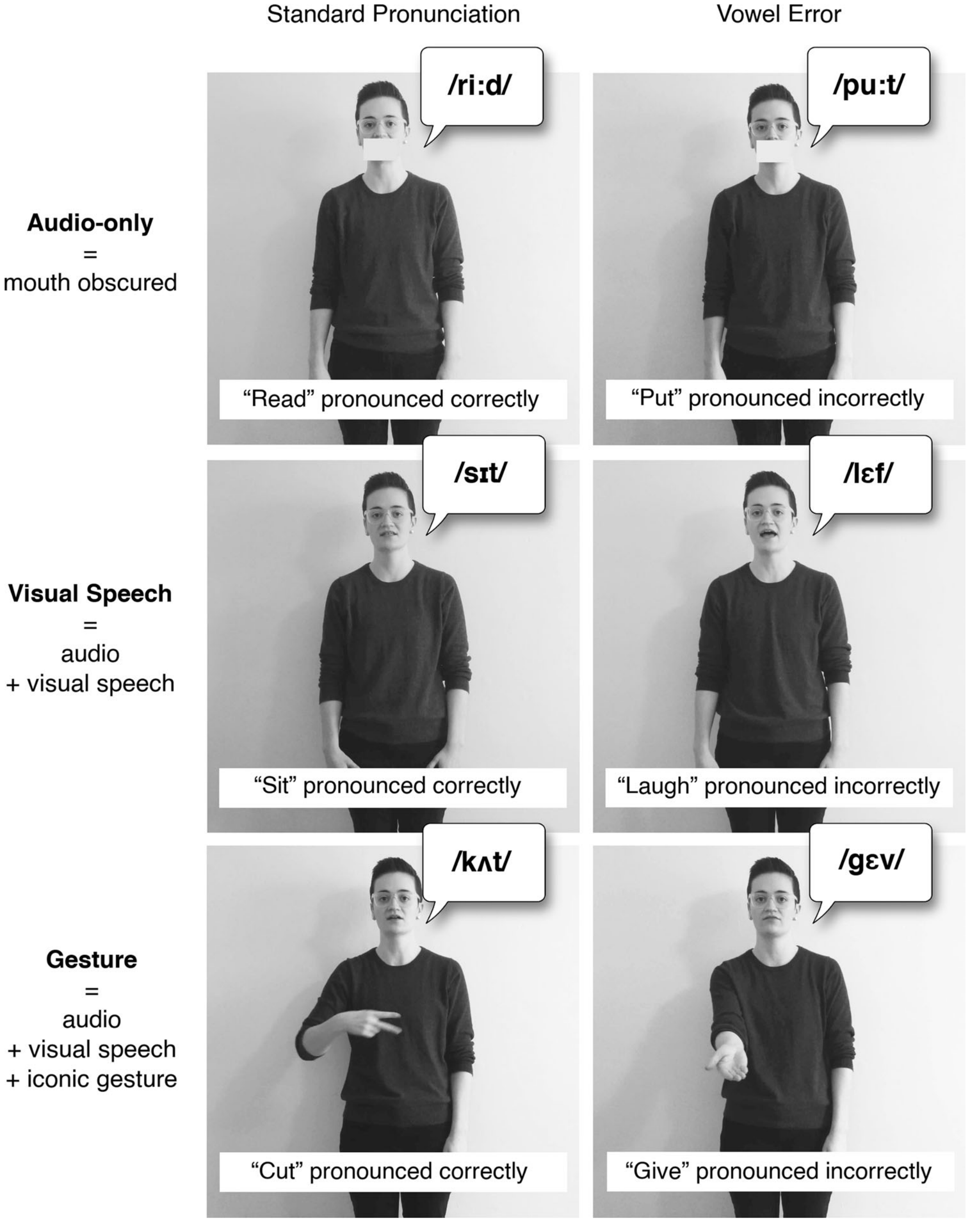
Stimulus Materials

Stimuli consisted of 10 two-second video clips for each of the six conditions, resulting in a total of 60 clips. The speaker was recorded in front of a neutral background from her knees to head. In each clip, the speaker was video-recorded saying a single monosyllabic action verb. In gesture conditions, the speaker additionally performed an iconic gesture. Audio-only conditions were created by obscuring the speaker's mouth (see Figure 1).

Audio and Visual Speech. The first author (an L1 English speaker with a General American [GA] accent) recorded a set of isolated verbs, some of which contained predetermined vowel errors (see Table 1; for the full list, see Appendix A1-A6). The vast majority of participants had never resided in the United States but likely had some familiarity with a GA accent through interaction with American classmates, travelling, or mass media. All verbs fell within the most frequent 2,000 word families in the BNC/COCA-25 (British National Corpus and Corpus of Contemporary American English) wordlist (Nation, 2017), confirmed through the website *The Compleat Lexical Tutor* (Cobb, 2019). In vowel error conditions, vowels were shifted in height, backness, and/or roundness. Large shifts (e.g., from a low front vowel to a high back vowel) were avoided. Audio files were denoised in Audacity (AudacityTeam, 2016) and intensity was scaled to 70 dB in Praat (Boersma & Weenink, 2018) to ensure that audio clarity and decibel level were consistent across the stimuli.

Words containing vowel errors were intended to be ambiguous; as such, shifts that would have resulted in the pronunciation of a different word in GA or received pronunciation (RP; i.e., standard British) English (e.g., /sɪt/ → /sɛt/) were

FIGURE 1
Overview of the Designs and Conditions Used in the Experiment


Standard Pronunciation
Vowel Error
/ri:d/
/pu:t/
Audio-only = mouth obscured
"Read" pronounced correctly
"Put" pronounced incorrectly
/sɪt/
/lɛf/
Visual Speech = audio + visual speech
"Sit" pronounced correctly
"Laugh" pronounced incorrectly
/kʌt/
/gɛv/
Gesture = audio + visual speech + iconic gesture
"Cut" pronounced correctly
"Give" pronounced incorrectly

not included. Previous studies have made similar efforts to avoid the complicating factor of word status (Munro & Derwing, 2006). It was impractical, however, to ensure that "incorrect" pronunciations would be considered inaccurate in every English dialect. Therefore, only GA and RP pronunciations, likely to be familiar to all listeners, were considered. After vowel error stimuli were recorded, the audio was reviewed and phonetically transcribed using the international phonetic alphabet by a researcher unfamiliar with the experiment. Transcriptions were checked to ensure that vowel sounds matched those intended (see Appendix A) and no large vowel shifts were created. One stimulus item was removed after this process.

Gesture. In gesture conditions, the speaker performed a scripted iconic gesture alongside the spoken verb. In each stimulus, the stroke of the gesture (i.e., the part of the movement that bears

TABLE 1
Examples of Stimuli in Vowel Error Conditions

Vowel shift	Examples
ɪ ↔ i:	mix (/mɪks/) pronounced as /mi:ks/
	feed (/fi:d/) pronounced as /fɪd/
ɛ ↔ ɪ	smell (/smɛl/) pronounced as /smɪl/
	give (/gɪv/) pronounced as /gɛv/
i: ↔ ɛ	teach (/ti:tʃ) pronounced as /tɛtʃ/
	press (/prɛs/) pronounced as /pri:s/
æ ↔ ɛ	laugh (/læf/) pronounced as /lɛf/
	let (/lɛt/) pronounced as /læt/
u: ↔ ʊ	move (/mu:v/) pronounced as /mʊv/
	push (/pʊʃ/) pronounced as /pu:ʃ/

meaning) co-occurred with the spoken word. The preparation phase of the gesture (i.e., when the arm begins to move from a resting position) began approximately 150–200 milliseconds after the start of the clip.

Intelligibility Task

Intelligibility was measured through an orthographic transcription task. The 60 stimuli were put in a pseudo-random order in which no condition occurred more than two times in a row. Stimuli were edited into a single video (10 minutes in length) using Kdenlive (2016) software, with 6 seconds of silence in between stimuli to leave ample time for transcription. Audio and video of item numbers (1–60) preceded each stimulus. Listeners recorded their transcriptions on a paper handout.

Questionnaire

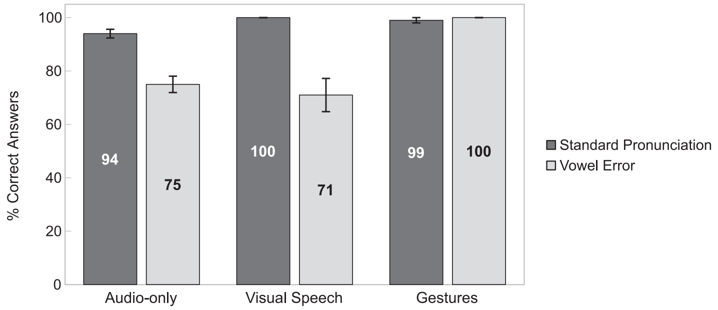
The questionnaire (see Appendix B) was adapted from one section of the VCPQ created by Sueyoshi & Hardison (2005). All items were 5-point Likert scales (1 = *strongly disagree*, 5 = *strongly agree*). Items 1–9 concerned listeners' beliefs, preferences, and behaviors regarding visual cues in daily life (e.g., "In face-to-face communication, I pay attention to the speaker's lip movements"). Items 10–13 focused specifically on the videos in the experiment (e.g., "In the videos that I just watched, I paid close attention to the speaker's lip movements"). Small adaptations were made to the questionnaire to suit the purposes of the current study. For example, the statement "It is easier to understand English when I can see the speaker's face" was split into two statements—one focusing on the comprehension of L1 English speech and one on L2 English speech—in order to investigate whether the perceived benefit of visual cues differs for foreign-accented speech in comparison to L1 speech. L1 Mandarin listeners and L1 English listeners received slightly different versions of the questionnaire, although the items were made as parallel as possible.

Piloting the Gesture Stimuli

All gestures were piloted with a group of 5 L1 English users from the United States (2 men and 3 women, aged 32–67), none of whom took part in the main experiment. Listeners completed the task via online video chat with the first author. The primary purpose of the pilot test was to ensure that gestures were appropriate and not misleading. The procedure was based on Drijvers & Özyürek (2017). First, listeners watched all of the gesture stimuli (with the audio removed and with mouth obscured) and were asked to write down the verbs they thought each movement communicated. Stimuli were presented following the same timing procedure as in the main experiment. Listeners were then shown the "correct" answers (i.e., the intended verbs alongside the gesture stimuli) and were asked to rate how well each verb matched the gesture on a 7-point Likert scale (1 = *doesn't fit the movement at all*, 7 = *fits the movement very well*).

Listeners' appropriacy ratings (from 1–7) were averaged for each stimulus. Five gesture stimuli that were not considered appropriate by listeners (i.e., those that received a mean score of less than 5) were discarded. For the remaining set of stimuli, we calculated a gesture recognition rate, the percentage of accurate answers. When exact matches as well as synonyms (e.g., "close" instead

FIGURE 2
Percentage of Correctly Identified Verbs per Condition (L1 Listeners)



Note. Error bars represent standard error.

of "shut") were coded as correct, the recognition rate was 80%. This is fairly high, but notably not 100%. This reflects that the meaning of an iconic gesture without its co-occurring speech is usually not entirely transparent. Overall, the results of the pilot study suggest that the gestures used in the main experiment will likely help disambiguate the spoken message, although they will not be sufficient in and of themselves for speech to be intelligible.

Procedure

The intelligibility task was administered with each listener individually in a quiet room. These instructions were given orally and provided on the task handout:

> You will hear the base form of a verb (e.g., 'walk,' 'find,' 'steal'). Some verbs are mispronounced, so you may not recognize them. For each video, which verb do you think the speaker is trying to communicate? If you are not sure, please try to guess.

Listeners watched the video on an 11.6-inch computer screen, wearing headphones. They started with a practice phase of six items (one per condition) and were allowed to ask questions before proceeding. Practice items were not included in the analysis. Listeners then watched the 10-minute video consisting of 60 stimuli and completed the transcription task. The video was played from start to finish without a break. After finishing this task, listeners filled out the VCPQ. Finally, listeners were asked to provide some biographical information (see Online Supporting Information A). In total, the completion of all tasks took 30–50 minutes per listener.

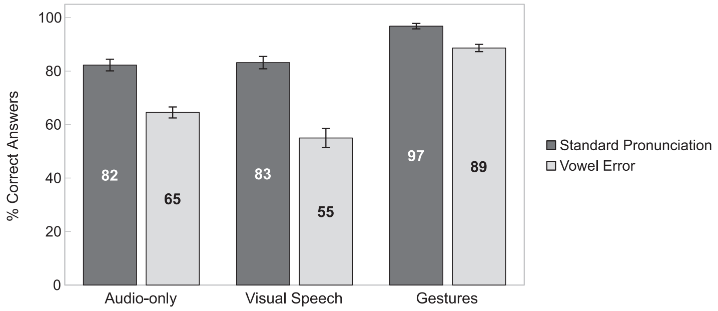
RESULTS FOR RESEARCH QUESTIONS 1 AND 2

In order to answer our first and second RQs, we examined the effects of pronunciation (+ error, – error) and modality (audio only, visual speech, and gesture) on L1 and L2 listeners' intelligibility scores. Mean scores across conditions are summarized in Figures 2 and 3. Standard deviations per condition are reported in Online Supporting Information B. In order to explore multiple within-participant and between-participant independent variables in a single analysis, a mixed-effects binomial logistic regression analysis was performed using the lmer functions from the lme4 package (Version 1.1-21; Bates et al., 2015) in the R statistical environment. In comparison to an analysis of variance (ANOVA), this mixed-effects model makes fewer assumptions of the dataset (it does not assume homoscedasticity or sphericity) and is able to take into account random variation across participants and stimulus items. The model included participants' binary intelligibility scores for each stimulus (0 for incorrect, 1 for correct) as the dependent variable with one between-participants factor (group: L1 vs. L2 listeners) and two within-participants factors (pronunciation and modality) as fixed effects. Both participants and stimuli were included as random effects. All the predictor variables (group, pronunciation, modality) were treatment coded. The following model of main and interaction effects was constructed:

$$\text{Intelligibility scores} = \text{Group} + \text{Pronunciation} + \text{Modality} + \text{Group} \times \text{Pronunciation} + \text{Group} \times \text{Modality} + \text{Group} \times \text{Pronunciation} \times \text{Modality}$$

In order to check the adequacy of the sample size for a statistical analysis involving

FIGURE 3
Percentage of Correctly Identified Verbs per Condition (L2 Listeners)



Note. Error bars represent standard error.

TABLE 2
Summary of Mixed-Effects Binomial Logistic Regression Analysis of Intelligibility Scores Relative to Listener Type, Pronunciation, Error, and Modality

Effects	Variables	*F*	*p*	Variance	*SD*	Conditional R^2	Marginal R^2
Fixed effects	Intercepts	9217.930	< .001*				
	Group	41.903	< .001*				
	Pronunciation	93.194	< .001*				
	Modality	68.070	< .001*				
	Group pronunciation	0.515	.473				
	Group × modality	3.033	.048*				
	Group × pronunciation × modality	2.604	< .001*				
Random effects	Participants			5.435	2.485	.796	.443
	Stimuli			0.272	0.521		

*$p < .05$.

cross-random factors of participants (N = 32) and stimuli (N = 60), resulting in a total of 1,920 observations, we followed Westfall et al.'s (2014) recommended procedure for calculating statistical power for a crossed-design model. The formula determines the optimal number of participants based not only on the number of predictors but also on the number of stimuli. The results of the prior analysis suggested an optimal sample size of 17.9 for the fully crossed model (i.e., where the same participants judge the same stimuli multiple times) to reach a medium effect size of $d = 0.5$ and sufficiently strong power of .80, which is in line with Larson–Hall's (2015) field-specific benchmark. The actual sample size of the current study (N = 32) was substantially larger than the suggested sample size (N= 17.9).

According to the results of the regression analyses (summarized in Table 2), the model significantly explained 44.3% of the variances in the outcomes of listeners' intelligibility judgments. According to Plonsky & Ghanbar's (2018) guidelines, the size of the effects reported here ($R^2 = .433$) can be considered medium to large. Significant main effects were found for group (L1 vs. L2), F(1,1400.498) = 41.903, $p < .001$; modality (audio-only vs. visual speech vs. gesture), F(2,1325.070) = 68.070, $p < .001$; and pronunciation (– error vs. + error), F(1,1400.498) = 93.194, $p < .001$. Interestingly, group status (L1 vs. L2) was found to be unrelated to the role of pronunciation (± errors) in intelligibility scores, F(1,1400.498) = .515, $p = .473$, indicating that vowel error affected L1 and L2 listeners to a comparable degree.

Further post-hoc tests were performed to investigate the effect of listener group and pronunciation on intelligibility. Results of an independent samples t test revealed that L1 listeners' intelligibility scores (M = 89.83) were significantly higher than those of L2 listeners (M = 78.41),

$t(30) = 5.63$, $p < .001$, with a large effect size ($d = 2.15$). A paired samples t test showed that L1 and L2 listeners' intelligibility scores were significantly higher in standard pronunciation conditions ($M = 90.63$) than in vowel error conditions ($M = 73.33$), $t(31) = 12.88$, $p < .001$, with a large effect size ($d = 2.28$).

Interestingly, the regression analysis revealed significant two-way interaction effects of group and modality, $F(2,1325.070) = 3.033$, $p = .048$, and three-way interaction effects of group, modality, and pronunciation. A set of multiple comparison analyses were conducted using the Wilcoxon signed ranks test to further examine the impact of modality on intelligibility scores for L1 and L2 listeners. Parametric tests could not be used as not all conditions were normally distributed. The effect size r was calculated from the Z statistic (Rosenthal, 1991).

L1 Listeners

The effect of visual cues in standard pronunciation conditions was not investigated as all scores in these conditions were near ceiling level for L1 listeners. To further explore the effect of modality in vowel error conditions (audio only + error vs. visual speech + error vs. gesture + error), three pairwise multiple comparisons were made using the Wilcoxon signed ranks test. This conservative nonparametric test was chosen given the relatively small number of participants ($n = 10$). A Bonferroni correction was applied, resulting in a corrected α level of .017. Results showed that under vowel error conditions, listeners' intelligibility scores were significantly greater when they had access to gestural information, both in comparison to the audio-only condition, $Z = -2.844$, $p = .004$, $r = .64$, and the visual speech condition, $Z = -2.530$, $p = .011$, $r = .57$. Interestingly, there was no significant difference between the audio-only and visual speech conditions, $Z = -0.526$, $p = .599$, $r = .12$.

L2 Listeners

A total of six Wilcoxon signed ranks tests were run to investigate the effect of modality with and without vowel errors. The α level was set to .008 (Bonferroni corrected). A significant difference was found between the visual speech and gesture modalities, both in standard pronunciation conditions, $Z = -3.677$, $p < .001$, $r = .55$, and vowel error conditions, $Z = -4.076$, $p < .001$, $r = .61$; both effect sizes were large. However, contrary to predictions, there was no significant difference between the audio-only modality and the visual speech modality in either standard pronunciation conditions, $Z = -0.532$, $p = .595$, $r = .08$, or vowel error conditions, $Z = -2.057$, $p = .04$, $r = .31$. The effects of visual cues are summarized in Table 3.

RESULTS FOR RESEARCH QUESTION 3: THE QUESTIONNAIRE

Mean scores and standard deviations for the 13 Likert items of the questionnaire are reported in Online Supporting Information B. Based on the hypothesis that certain statements were correlated to each other as aspects of a larger construct, internal consistency reliability tests were run for two groups of Likert items—a group relating to visual speech and another relating to gesture. For statements regarding visual speech (items 1, 3, 8, 10, and 13), Cronbach's alpha was .82, suggesting that these items could be combined into a multi-item scale. The gesture group (items 2, 4, 5, 6, 7, 9, 11, and 12) was also internally consistent with a Cronbach's alpha of .79. Removing items did not substantially improve internal consistency, so all items were retained. The new multi-item scales of "attitude toward visual speech" and "attitude toward gesture" were created by calculating the mean of all the items within the scale. After performing a power transformation ($\textit{transformed data} = \textit{raw data}^{2.1}$), the scales were found to be normally distributed.

Results of an independent samples t test showed that L2 listeners had a significantly more positive attitude toward visual speech ($M = 3.47$) than L1 listeners ($M = 2.76$), $t(30) = 2.39$, $p = .02$, with a large effect size ($d = 0.91$). There was no difference between groups in attitude toward gesture. Listeners generally agreed that gesture was beneficial for comprehension ($M_{\text{L2}} = 4.19$, $M_{\text{L1}} = 3.95$). A paired samples t test revealed that all listeners had a significantly more positive attitude toward gesture ($M = 4.11$) than visual speech ($M = 3.25$), $t(31) = 6.28$, $p < .001$, with a large effect size ($d = 1.11$).

DISCUSSION

There are two primary findings of this study. First, iconic gestures significantly increased intelligibility when (a) speech contained vowel errors and/or (b) the listener was an L2 user of the language. The only situation in which iconic gesture did not facilitate intelligibility was when L1 listeners perceived speech in a standard accent; in this case, intelligibility was near ceiling in every

TABLE 3
Effect of Visual Cues on Intelligibility for L1 and L2 Listeners

Effect	L1 English listeners ($n = 10$)	L2 English listeners ($n = 22$)
Positive gesture effect[a]		
Standard pronunciation	n.s.	$r = .55^{**}$
Vowel error	$r = .57^{*}$	$r = .61^{**}$
Visual speech effect[b]		
Standard pronunciation	n.s.	n.s.
Vowel error	n.s.	n.s.

Note. n.s. = no significant effect was found. $r > .5$ is considered a large effect.
[a] Gesture effect = intelligibility of visual speech vs. gesture.
[b] Visual speech effect = intelligibility of audio-only vs. visual speech.
$^{*}p < .01$, significant at a corrected α level of .017.
$^{**}p < .001$, significant at an α level of .008.

modality. The second main finding was that vowel error significantly reduced intelligibility for both L1 and L2 listeners.

Research Question 1: Visual Cues, Vowel Errors, and L1 Listeners

Results showed that vowel error reduced intelligibility and that iconic gesture increased intelligibility for L1 listeners. Visual speech had no significant effect.

The Effect of Gesture. For L1 listeners, iconic gestures facilitated intelligibility when speech contained vowel errors but not when speech was pronounced in a standard, familiar accent. Intelligibility scores in the gesture + error condition were significantly higher than scores in the visual speech + error condition, with a large effect size ($r = .57$). These results suggest that when speech contains linguistic errors, iconic gestures can help disambiguate the message by providing additional semantic information. Previous studies have shown that gesture facilitates comprehension to a greater degree in conditions of noise or noise-vocoding where the auditory signal is difficult to decode (Drijvers & Özyürek, 2017; Riseborough, 1981; Rogers, 1978). The current study has similarly found that iconic gesture is more likely to affect understanding when auditory information is ambiguous, although the ambiguity in this case is due to phonological error.

It is unclear whether other types of gesture would have the same effect. Hostetter's (2011) meta-analysis of 63 gesture studies found that gestures were particularly beneficial when they conveyed information about space or movement (as iconic gestures often do) rather than abstract concepts. It is therefore likely that the size of the benefit found in the current study is partially a result of the type of gesture under investigation. However, Hostetter's meta-analysis found no significant difference in effect size between studies examining spontaneous gesture and those using scripted gesture, suggesting that the use of scripted gestures in the current study may not be a significant moderating factor.

The Effect of Visual Speech. A large standard deviation (19.7%) was observed in the visual speech + error condition. This aligns with previous research that has found large individual variation in listeners' abilities to benefit from visual speech (Grant et al., 1998) and visual information in general (Wagner, 2008). This may be due to differences in lipreading ability and ability to integrate visual and auditory information. Somewhat unexpectedly, no significant differences were found between audio-only conditions and visual speech conditions. This result seems to contradict previous research that has found an enhancement effect for visual speech (Kawase et al., 2014; Sumby & Pollack, 1954). Three possible reasons might account for the finding of the present study: lack of saliency, lack of informativeness, and measurement insensitivity.

Unlike much previous research, the speaker's lips were not particularly salient in the present study. In most audiovisual speech perception research, stimuli consist of video close-ups of the speaker's face on a large monitor. In the current study, however, the speaker was seen at a medium distance on a relatively small screen (1366 × 768). While some previous studies have recorded from a medium distance, these are usually presented on a larger screen. For example, Drijvers and Özyürek

(2017) used a 1650 × 1080 monitor. Given the effect of distance on the usefulness of visual information (Zheng & Samuel, 2019), it may be that the speaker's lips were simply too small to capture the attention of listeners. Questionnaire responses confirm that L1 listeners paid little attention to the speaker's lips in the experiment ($M = 2.20$).

In vowel error conditions, lack of informativeness may also have influenced visual speech intelligibility scores. As with the /ɹ/ sound in Kawase et al.'s (2014) study, the articulations of the mouth and lips for the vowel sounds in these conditions were inaccurate. Unlike some other research, the current study did not include incongruent stimuli (e.g., hearing "gev" but seeing "gave"), nor did it include degraded speech matched with clear visuals (e.g., hearing "gave" or "gaze" but clearly seeing "gave"). Rather, it used congruent stimuli which contained errors in both channels (e.g., hearing "gev" and seeing "gev"). It is therefore not surprising that visual speech was unhelpful in vowel error conditions.

Finally, the measurement itself may be unreliable. It could be, for example, that there is some unknown factor that made the words chosen for the visual speech conditions particularly difficult to understand. Especially given the large standard deviation in the visual speech + error condition, it may be necessary to include a larger number of stimuli per condition and/or a larger number of listeners to achieve a sufficiently sensitive measurement.

The Effect of Vowel Errors

In alignment with previous research (Bent et al., 2007; Zielinski, 2008), vowel errors were found to reduce the intelligibility of speech. The effect size of this decrease was large ($d = 2.28$). In the current study, intelligibility was reduced by 20–29% in audio-only and visual speech conditions for L1 listeners. This represents a huge loss, especially considering that even a small percentage of misunderstood words (anything more than 3–5%) can significantly affect the global intelligibility of spoken discourse (Nation, 2001). It should be noted that stimuli consisted of monosyllabic words in which a single segmental error is likely to have a larger effect than in multisyllabic words. Additionally, to control for the influence of surrounding words on listeners' intelligibility scores, words were presented out of context. This methodological decision may have made these words particularly difficult to understand. However, as Field (2005) pointed out, added context (e.g., a key word embedded in a sentence or longer stretch of discourse) is only helpful to the extent that this context can be accurately decoded by the listener. If the surrounding context is also somewhat difficult to decode, then it may not increase the intelligibility of the key word.

Research Question 2: The Effect of Language Background

Gesture: L2 vs. L1 Listeners. Similar to the L1 listener group, L2 listeners experienced a large beneficial effect of gesture in vowel error conditions ($r = .61$). It is not possible to say whether the degree of benefit in these conditions differed meaningfully between groups because of ceiling effects in the L1 group. In standard pronunciation conditions, the effect of gesture clearly differed between groups. Unlike L1 listeners, L2 listeners benefited from gesture when there were no vowel errors (with a large effect size of $r = .55$). This aligns with previous research that has found a positive effect of gesture on L2 listening comprehension in normal listening conditions (Dahl & Ludvigsen, 2014; Sueyoshi & Hardison, 2005). The current study expands these findings to highly proficient L2 listeners. Unlike L1 listeners, L2 listeners are able to benefit from gesture in "normal" conditions because their intelligibility scores do not reach ceiling based on auditory information alone. Lower baseline scores leave room for improvement that does not exist for L1 listeners. This suggests that when speech is relatively easy to decode (i.e., when it contains no phonological errors and is presented in quiet conditions), gesture may have a greater potential to affect understanding for L2 listeners than L1 listeners. However, when listening conditions become more difficult (e.g., due to noise or, in this case, vowel error), gesture significantly affects understanding for both L1 and L2 listeners.

Visual Speech: L2 vs. L1 Listeners. As with L1 listeners, the largest standard deviation occurred in the visual speech + error condition, suggesting a high degree of individual variability. Similar to L1 listeners, L2 listeners did not have significantly different scores in visual speech and audio-only conditions. This runs counter to research that suggests visual speech increases intelligibility for L2 listeners (Hazan et al., 2006; Navarra & Soto–Faraco, 2007). As stated earlier, this could be because mouth movements lacked salience, because they lacked informativeness (in the vowel error conditions), or because the measurement lacked sensitivity.

Even if these issues were addressed, however, it is unlikely that visual speech would have a large effect. Gesture, at least iconic gesture, may have a relatively greater effect on intelligibility than information from a speaker's mouth and lips. This finding could shed some light on L2 listening comprehension research. While some studies have found a facilitative effect of video on L2 listening comprehension (Sueyoshi & Hardison, 2005; Wagner, 2010b), others have found no effect (Batty, 2015). This may be partially due to a difference in the types of visual cues being presented. We can speculate that video texts that include a greater number of iconic gestures may be more likely to find a benefit in audiovisual conditions.

Vowel Error: L2 vs. L1 Listeners. The language background of listeners was found to be unrelated to the effect of vowel error. In vowel error conditions, the percentage of intelligibility loss for L2 listeners was comparable to that observed for L1 listeners, reaching 21.5–33.9% in audio-only and visual speech conditions.

Research Question 3: Attitudes About Visual Speech and Gestures

Questionnaire results revealed that Mandarin L1 users tended to have a more positive attitude toward visual speech and facial cues (as a comprehension aid) than English L1 users. It is possible that this reflects a cultural difference, but it more likely reflects an L1–L2 difference. Most of the questions in the "attitude toward visual speech" scale focused on listening to English. As speech in an L2 is usually not completely intelligible from the auditory signal alone, it is not surprising that L2 listeners rated visual speech as more beneficial than L1 listeners for comprehension. L2 listeners' belief in the facilitative effect of visual information is well documented in listening assessment research (Cubilo & Winke, 2013; Wagner, 2010a). This belief did not appear to have any effect on listeners' intelligibility scores.

Questionnaire results also showed that L1 and L2 listeners believed that gestures were more helpful for comprehension than visual speech—both in the experiment and in day-to-day life. This belief aligns with the results from the experiment, in which gesture increased intelligibility, but visual speech did not. Listeners reported that visual cues were especially helpful when listening to L2 speakers (when compared to L1 speakers). At least in the case of L1 listeners, this belief was borne out in the experiment; gesture increased intelligibility when there were vowel errors (a type of error that is sometimes made by L2 speakers), but not in standard pronunciation conditions. This suggests that both the actual and perceived benefit of gesture is dependent on the intelligibility of the auditory signal: when the auditory signal is more difficult to decode, the benefit of gesture is greater.

LIMITATIONS

With an eye toward future replications, there are a few limitations of this study that need to be discussed. First, results concerning L2 listeners cannot be generalized to all populations but must be limited to Mandarin L1 users. Results are further limited to a population that is highly proficient in their L2, has a background in linguistics and/or language teaching, belongs to a particular age group, and has lived in an English-speaking country for a relatively short amount of time.

Additionally, the current study necessarily lacked a certain amount of ecological validity. Words, gestures, and pronunciation errors were carefully scripted and performed; recordings were presented in a laboratory setting; and only local intelligibility at the word level was investigated. Moreover, phonological deviations consisted entirely of vowel errors and are thus clearly unrepresentative of natural foreign-accented speech. Similarly, gestures were limited to a particular type and do not reflect the diversity and complexity of gesture as it is used in normal spoken interaction. This type of controlled design that focuses on single words or syllables in conjunction with one particular phonological feature (Field, 2005; Kawase et al., 2014) or one particular gesture type (Drijvers & Özyürek, 2017) is often used in intelligibility research in order to eliminate confounding variables, thereby increasing the reliability and replicability of results and allowing for the investigation of cause-and-effect relationships. While controlled experiments of this type have a strong place and purpose in intelligibility research, they must be complemented by more naturalistic research which uses extemporaneous speech, unscripted gestures, and measures of global intelligibility. Additional controlled experiments that focus on other phonological features and other gesture types are also needed.

Finally, small sample sizes and ceiling effects limited the analyses we could run. As a consequence, nonparametric tests were used to make post-hoc comparisons between conditions. Research with larger sample sizes and designs that avoid ceiling effects (by including noise, for example) could help confirm the findings of the current study and provide more nuanced results.

IMPLICATIONS

If iconic gestures improve the intelligibility of speech containing vowel errors, it is possible that L2 learners who make such errors could benefit from attending to their gesture use. We say "attention to gesture use" rather than "acquisition of gesture forms," as the latter is much more controversial and not a direct implication of the current study. This study has focused on iconic gestures that, by definition, are not culturally or linguistically specific—that is, they would not need to be explicitly taught in order to be used or understood in an L2. Although it is not necessary to teach these forms, it might be useful to raise awareness of their potential power in communication. In a recent book, Gregersen and MacIntyre (2017) provided a number of exercises for the L2 classroom that serve this purpose.

The findings of this study might also have implications for face-to-face speaking assessments. If iconic gesture has the power to affect the intelligibility of L2 speakers, as the current study suggests it might, it is reasonable to consider whether gesture or other nonverbal behavior should be assessed as an aspect of speaking proficiency, as some scholars have suggested (Plough et al., 2018). It is possible that gesture is in fact already a hidden factor in many oral assessments. Through rater reports and stimulated verbal recalls, several recent studies have found that nonverbal behavior (including gesture), while not listed in the criteria, is heavily relied upon when rating the "interactional competence" of participants in peer-to-peer candidate exams (Ducasse & Brown, 2009; May, 2011). Jenkins and Parra (2003) came to a similar conclusion in their investigation of interview-style exams, suggesting that simply excluding gesture from speaking criteria does not eliminate its influence. Rather, at least in some contexts, gesture may operate as a hidden variable, affecting scores and the reliability of tests in an uncontrolled manner.

A similar debate surrounds L2 listening assessment. Just as with speaking, it can be argued that visual speech and gesture are part of the construct of listening. Sueyoshi and Hardison (2005) made this case, as have a number of other researchers, including Wagner (2008) and Ockey (2007). In alignment with other research (Grant et al., 1998; Wagner, 2008), the current study found a wide range of variability in listeners' ability to benefit from visual cues, especially visual speech. Wagner (2008) argued that this ability should be measured as part of the listening construct. Whether it would be practical to add a visual component to either listening or speaking assessment is unclear, but the current study joins others in questioning the authenticity and validity of audio-based assessment.

CONCLUSION

The main aim of this study was to incorporate a visual modality into L2 intelligibility research. Previous studies that have focused on the effect of visual cues, especially those that include gesture, have been limited to the context of standard L1 speech. The current study is the first to explore the effect of iconic gesture on the intelligibility of speech containing phonological errors. Findings revealed a complex interaction among vowel accuracy, visual cues, and listener background in determining the intelligibility of speech. When the auditory signal was difficult to decode (when speech contained vowel errors and/or when the listener was an L2 user of the language), iconic gesture significantly increased intelligibility. These findings add important information to ongoing debates surrounding the place of gesture in L2 teaching and assessment.

Open Research Badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://www.iris-database.org.

REFERENCES

AudacityTeam. (2016). *Audacity(R): Free audio editor and recorder* (Version 2.1.2) [Computer software]. http://www.audacityteam.org

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015). Cognitive predictors of perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, *137*, 2015–2024. https://doi.org/10.1121/1.4916265

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Batty, A. O. (2015). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning.

Language Testing, *32*, 3–20. https://doi.org/10.1177/0265532214531254

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica*, *123*, 1–30. https://doi.org/10.1515/semi.1999.123.1-2.1

Bent, T., Bradlow, A. R., & Smith, B. L. (2007). Segmental errors in different word positions and their effects on intelligibility of non-native speech. In O. S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 331–347). John Benjamins Publishing. https://doi.org/10.1075/lllt.17.28ben

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer* (Version 6.0.4) [Computer software]. https://www.praat.org/

Cobb, T. (2019). *Compleat Web VP* (Version 2.1) [Computer software]. https://www.lextutor.ca//vp/comp/

Cubilo, J., & Winke, P. (2013). Redefining the L2 listening construct within an integrated writing task: Considering the impacts of visual-cue interpretation and note-taking. *Language Assessment Quarterly*, *10*, 371–397. https://doi.org/10.1080/15434303.2013.824972

Dahl, T. I., & Ludvigsen, S. (2014). How I see what you're saying: The role of gestures in native and foreign language listening comprehension. *Modern Language Journal*, *98*, 813–833. https://doi.org/10.1111/modl.12124

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research* (Vol. *42*). John Benjamins. https://doi.org/10.1075/lllt.42

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, *60*, 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Drijvers, L., & Özyürek, A. (2020). Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language and Speech*, *63*, 209–220. https://doi.org/10.1177/0023830919831311

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Rater's orientation to interaction. *Language Testing*, *26*, 423–443. https://doi.org/10.1177/0265532209104669

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*, 399–423. https://doi.org/10.2307/3588487

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*, *103*, 2677–2690. https://doi.org/10.1121/1.422788

Gregersen, T., & MacIntyre, P. D. (2017). *Optimizing language learners' nonverbal behavior: From tenet to technique*. Channel View Publications. https://doi.org/10.21832/9781783097371

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*, 201–223. https://doi.org/10.2307/3588378

Hazan, V., Sennema, A., Faulkner, A., Ortega–Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America*, *119*, 1740–1751. https://doi.org/10.1121/1.2166611

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, *137*, 297–315. https://doi.org/10.1037/a0022128

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *Modern Language Journal*, *87*, 90–107. https://doi.org/10.1111/1540-4781.00180

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*, 115–146. https://doi.org/10.1111/lang.12270

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

Kawase, S., Hannah, B., & Wang, Y. (2014). The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers. *The Journal of the Acoustical Society of America*, *136*, 1352–1362. https://doi.org/10.1121/1.4892770

Kdenlive. (2016). Kdenlive (Version 15.12.3) [Computer software]. https://kdenlive.org

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press. https://doi.org/10.1017/CBO9780511807572

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*, 145–167. https://doi.org/10.1080/01690960802586188

Larson–Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R*. Routledge. https://doi.org/10.4324/9780203875964

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press. https://doi.org/10.1017/9781108241564

Magnotti, J. F., Mallick, D. B., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental Brain Research*, *233*, 2581–2586. https://doi.org/10.1007/s00221-015-4324-7

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language*

Assessment Quarterly, *8*, 127–145. https://doi.org/10.1080/15434303.2011.565845

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748. https://doi.org/10.1038/264746a0

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*, 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

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

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*, 520–531. https://doi.org/10.1016/j.system.2006.09.004

Munro, M. J., & Derwing, T. M. (2015). Intelligibility in research and practice: Teaching priorities. In M. Reed & J. M. Levis (Eds.), *The handbook of English pronunciation* (pp. 375–396). Wiley. https://doi.org/10.1002/9781118346952.ch21

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, I. S. P. (2017). *The BNC/COCA Level 6 word family lists* (Version 1.0.0) [Data file]. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources

Navarra, J., & Soto–Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*, 4–12. https://doi.org/10.1007/s00426-005-0031-5

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*, 517–537. https://doi.org/10.1177/0265532207080771

Parry, T. S., & Meredith, R. A. (1984). Videotape vs. audiotape for listening comprehension tests: An experiment. *OMLTA Journal*, *1984*, 47–53.

Plough, I., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing*, *35*, 427–445. https://doi.org/10.1177/0265532218772325

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior*, *68*, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *Modern Language Journal*, *102*, 713–731. https://doi.org/10.1111/modl.12509

Quené, H., & van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication*, *52*, 911–918. https://doi.org/10.1016/j.specom.2010.03.005

Riseborough, M. G. (1981). Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication. *Journal of Nonverbal Behavior*, *5*, 172–183. https://doi.org/10.1007/BF00986134

Rogers, W. T. (1978). The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances. *Human Communication Research*, *5*, 54–62. https://doi.org/10.1111/j.1468-2958.1978.tb00622.x

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). SAGE. https://doi.org/10.4135/9781412984997

Ross, L. A., Saint–Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*, 1147–1153. https://doi.org/10.1093/cercor/bhl024

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*, 661–699. https://doi.org/10.1111/j.0023-8333.2005.00320.x

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*, 212–215. https://doi.org/10.1121/1.1907309

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

Tye–Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification*, *11*, 233–241. https://doi.org/10.1177/1084713807307409

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, *5*, 218–243. https://doi.org/10.1080/15434300802213015

Wagner, E. (2010a). Test-takers' interaction with an L2 video listening test. *System*, *38*, 280–291. https://doi.org/10.1016/j.system.2010.01.003

Wagner, E. (2010b). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*, 493–513. https://doi.org/10.1177/0265532209355668

Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly*, *10*, 178–195. https://doi.org/10.1080/15434303.2013.769552

Wells, J. C. (2008). *Longman pronunciation dictionary* (3rd ed.). Longman.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: Gen-*

eral, *143*, 2020–2045. https://doi.org/10.2139/ssrn.2449567

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

Yi, H. G., Phelps, J. E., Smiljanic, R., & Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America*, *134*, EL387–EL393. https://doi.org/10.1121/1.4822320

Zheng, Y., & Samuel, A. G. (2019). How much do visual cues help listeners in perceiving accented speech? *Applied Psycholinguistics*, *40*, 93–109. https://doi.org/10.1017/S0142716418000462

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*, 69–84. https://doi.org/10.1016/j.system.2007.11.004

APPENDIX A

Stimuli List

Frequency from lextutor BNC/COCA-25: 1K = within top 1,000 word families; 2K = within top 2,000 word families.

TABLE A1
Audio-Only – Vowel Error Condition

Word	Frequency	GA pronunciation
tell	1K	/tɛl/
sing	1K	/sɪŋ/
read	1K	/ri:d/
cook	1K	/kʊk/
serve	1K	/sɝ:v/
lead	1K	/li:d/
lock	1K	/lɑ:k/
search	2K	/sɝ:tʃ/
bend	2K	/bɛnd/
bump	2K	/bʌmp/

GA = General American; phonetic transcriptions are from *The Longman Pronunciation Dictionary* (Wells, 2008).

TABLE A2
Audio-Only + Vowel Error Condition

Word	Frequency	Vowel change
put	1K	/pʊt/ → /pu:t/
rest	1K	/rɛst/ → /ri:st/
teach	1K	/ti:tʃ/ → /tɛtʃ/
move	1K	/mu:v/ → /mʊv/
pass	1K	/pæs/ → /pɛs/
guess	1K	/ɡɛs/ → /ɡɪs/
let	1K	/lɛt/ → /læt/
tap	2K	/tæp/ → /tɛp/
match	2K	/mætʃ/ → /mɛtʃ/
cheat	2K	/tʃi:t/ → /tʃɛt/

TABLE A3
Visual Speech – Vowel Error Condition

Word	Frequency	GA pronunciation
fill	1K	/fɪl/
meet	1K	/mi:t/
fit	1K	/fɪt/
come	1K	/kʌm/
win	1K	/wɪn/
leave	1K	/li:v/
pack	1K	/pæk/
drag	2K	/dræɡ/
sink	2K	/sɪŋk/
risk	2K	/rɪsk/

TABLE A4
Visual Speech + Vowel Error Condition

Word	Frequency	Vowel change
feed	1K	/fi:d/ → /fɪd/
laugh	1K	/læf/ → /lɛf/
fish	1K	/fɪʃ/ → /fɛʃ/
step	1K	/stɛp/ → /stɪp/
send	1K	/sɛnd/ → /si:nd/
lose	1K	/lu:z/ → /lʊz/
check	1K	/tʃɛk/ → /tʃæk/
lend	2K	/lɛnd/ → /lɪnd/
spread	2K	/sprɛd/ → /spri:d/
spell	2K	/spɛl/ → /spæl/

TABLE A5
Iconic Gesture – Vowel Error Condition

Word	Frequency	GA pronunciation	Iconic gesture
call	1K	/kɑ:l/	Hand (Y-shape) is brought to ear
shoot	1K	/ʃu:t/	Hand (L-shape) makes “bang bang” recoil motion
shut	1K	/ʃʌt/	Hands touch at 90° angle (open palm), then close together
drink	1K	/drɪŋk/	Hand (C-shape) is brought to mouth, tilts back
cut	1K	/kʌt/	Hand (V-shape) makes scissor motion
lift	1K	/lɪft/	Both hands (open palm, facing upward) move upward
drop	1K	/drɑ:p/	Hand starts in clenched position (palm facing down), then releases
pat	2K	/pæt/	Fingers lightly touch shoulder two times
fan	2K	/fæn/	Hand (open palm, fingers spread) tilts up and down close to face
chop	2K	/ʧɑ:p/	Right hand (open palm facing left) strikes left hand (open palm facing up)

TABLE A6
Iconic Gesture + Vowel Error Condition

Word	Frequency	Vowel change	Iconic gesture
give	1K	/gɪv/ → /gɛv/	Hand (open palm, facing up) extends forward
catch	1K	/kæʧ/ →/kɛʧ/	Hands (curved) clasp together
push	1K	/pʊʃ/ → /pu:ʃ/	Both hands (open palms, facing out) extend away from body
sleep	1K	/sli:p/ → /slɛp/	Hands (palms together) press to side of face
smell	1K	/smɛl/ → /smɪl/	Open hand makes “whiff” motion near nose
press	1K	/prɛs/ → /pri:s/	Right hand (open palm, facing down) meets left hand (open palm, facing up) and both move downward
mix	2K	/mɪks/ → /mi:ks/	Left hand in C-shape as if holding a bowl; right hand (clenched fist) makes fast circular motion as if stirring
spin	2K	/spɪn/ → /spi:n/	Index finger (pointing up) makes rotating motion
flip	2K	/flɪp/ → /flɛp/	Hand (open palm, facing up) turns over (open palm, facing down)
twist	2K	/twɪst/ → /twɛst/	Both hands (clenched position) rotate in opposite directions

APPENDIX B

Visual Cue Preference Questionnaire for L1 Participants

The purpose of this questionnaire is to find out more about your attitudes and preferences regarding visual information and gestures (movements of the arms and hands). There are no right or wrong answers. Your personal information will be kept confidential. If there are any questions that you do not want to answer, you may leave them blank. Thank you!

Please circle the number that expresses your opinion.

1 = strongly disagree, 2 = disagree, 3 = I'm not sure, 4 = agree, 5 = strongly agree

		Strongly disagree	**Disagree**	**I'm not sure**	**Agree**	**Strongly agree**
1.	It is easier to understand L1 English users when I can see the speaker's face.	1	2	3	4	5
2.	It is easier to understand L1 English users when I can see the speaker's gestures (hand and arm movements).	1	2	3	4	5
3.	It is easier to understand L2 English users when I can see the speaker's face.	1	2	3	4	5
4.	It is easier to understand L2 English users when I can see the speaker's gestures (hand and arm movements).	1	2	3	4	5
5.	When I talk in a second language, I use gestures more frequently than when I talk in English. *(leave blank if you rarely/never talk in a second language)*	1	2	3	4	5
6.	When I talk in a second language, I think people understand my speech better when I use gestures. *(leave blank if you rarely/never talk in a second language)*	1	2	3	4	5
7.	I think my friends understand my speech in English better when I use gestures.	1	2	3	4	5
8.	In face-to-face communication, I pay attention to the speaker's lip movements.	1	2	3	4	5
9.	In face-to-face communication, I pay attention to the speaker's gestures.	1	2	3	4	5
10.	In the videos that I just watched, I paid close attention to the speaker's lip movements.	1	2	3	4	5
11.	In the videos that I just watched, I paid close attention to the speaker's gestures.	1	2	3	4	5

		Strongly disagree	Disagree	I'm not sure	Agree	Strongly agree
12.	In the videos that I just watched, I believe that watching the speaker's gestures helped my understanding.	1	2	3	4	5
13.	In the videos that I just watched, I believe that seeing the speaker's lips helped my understanding.	1	2	3	4	5

14. Please write any comments you wish about this research. Which videos were the most difficult for you? Did you think visual cues were helpful? (Optional)

Visual Cue Preference Questionnaire for L2 Participants

The wording of items 5, 6, and 7 differed slightly for L2 participants. All other items remained the same.

		Strongly disagree	Disagree	I'm not sure	Agree	Strongly agree
5.	I use gestures more frequently when I talk in English than when I talk in Chinese.	1	2	3	4	5
6.	I think people understand my speech in English better when I use gestures.	1	2	3	4	5
7.	I think my friends understand my speech in Chinese better when I use gestures.	1	2	3	4	5

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.