BRITISH COUNCIL

Institute of Education UCL

The effect of vowel accuracy, visual speech, and iconic gesture on intelligibility

by Page Wheeler

British Council's Master's Dissertation Awards 2020

Winner

www.teachingenglish.org.uk/publications/elt-masters-dissertations

The Effect of Vowel Accuracy, Visual Speech, and Iconic Gesture on Intelligibility

MA TESOL

Page Wheeler

August 2019

Acknowledgements

I would like to thank Kazuya Saito for his insight, guidance, and kindness. Rarely have I felt so encouraged and motivated to do something well. I am also grateful to Ana Pellicer-Sanchez, Andrea Révész, and Amos Paran for all of their excellent classes over the course of the programme. A special acknowledgement goes to my sister Cory who read and helped edit every word of this dissertation (a few times over). Finally, thanks to all the students who participated in this study.

Abstract

Intelligibility, defined as the extent to which speech is understood by an interlocutor, is a core concept in L2 pronunciation research. There is a growing number of empirical studies that investigate which aspects of language are most likely to affect intelligibility and therefore deserve priority in the classroom and in oral assessment. The vast majority of these studies, however, rely on audio-only stimuli. There is relatively little known about how visual information may interact with these linguistic properties in face-to-face communication. The current study addresses this gap by incorporating a visual modality into its design. In Experiment 1, 10 L1 English users were presented with stimuli that varied along three factors (vowel error, visual speech, and iconic gesture) and completed an orthographic transcription task. In Experiment 2, 22 L2 English (L1 Mandarin) users completed the same task. Results revealed that iconic gesture significantly increased intelligibility when speech contained vowel errors ($r = .57$ for L1 listeners; $r = .83$ for L2 listeners). 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 experiment. Vowel error reduced intelligibility by approximately 20% for both L1 and L2 participants. Findings suggest that visual cues, especially gestures, have the potential to significantly affect L2 speech intelligibility. These effects may be moderated by the language background of the listener.

Contents

Figures

Tables

1 Introduction

There is now little debate in the field of L2 speaking and pronunciation research that the primary goal of instruction should be *intelligibility* rather than *native-likeness* (Derwing & Munro, 2015; 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 the primary aim of intelligibility research. In the last several decades, a number of areas have been examined, including segmentals (e.g. Bent, Bradlow, & Smith, 2007), word stress (e.g. Field, 2005), rhythm (e.g. Tajima, Port, & Dalby, 1997), and intonation (e.g. Winters & O'Brien, 2013), as well as features beyond pronunciation such as grammar and lexis (e.g. Munro & Derwing, 1995a). Although certain errors do seem to have a greater impact than others, it has become clear that the gravity of a mistake depends not only on the type of error, but also on the background of the listener (Stringer & Iverson, 2019).

The vast majority of L2 intelligibility research relies on audio-only stimuli, although a few researchers have begun to question this methodological stance (Kawase, Hannah, & Wang, 2014). We know from 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 (Peelle & Sommers, 2015; Hostetter, 2011). On a theoretical level, researchers argue that it is a mistake to dismiss these visual cues as somehow subsidiary to the auditory signal; audio and visual modalities are tightly integrated and *jointly* determine the content of a message (McNeill, 1992; Goldin-Meadow, 2003). Given what we know about the power of visual cues, it is notable that they are rarely studied in intelligibility research.

The current study attempts to expand L2 intelligibility research by incorporating visual modalities. The first aim of the study is to examine how intelligibility (as perceived by L1 English listeners) is 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. I will begin by providing a review of research in L2 intelligibility, audiovisual speech perception, and co-speech gesture (Chapter 2). All permutations of L1 and L2 user interactions will be explored:

- L1 speaker – L1 listener (L1S – L1L)
- L1 speaker – L2 listener (L1S – L2L)
- L2 speaker – L1 listener (L2S – L1L)
- L2 speaker – L2 listener (L2S – L2L)

Chapter 3 explains the rationale and research questions of the current study in light of this research. Chapter 4 outlines the methodology of two experiments designed to meet the two respective aims of the study. Results and discussion follow in Chapters 5 and 6. Finally, Chapter 7 summarizes findings and concludes with areas for further research.

2 Literature Review

2.1 Second Language Intelligibility

2.1.1 Defining the Construct

What has thus far been broadly referred to as "intelligibility" requires a more precise definition. Although synonymous in general parlance, the terms "intelligibility" and "comprehensibility" are regarded as distinct constructs by L2 researchers (Munro & Derwing, 1995a; Derwing & Munro, 1997, 2015). *Intelligibility* is 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 refer to these respectively as "local" and "global" intelligibility (2015, p. 381).

Intelligibility is most often operationalised 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 (e.g. Smith & Nelson, 1985), sentence verification (e.g. Munro & Derwing, 1995b), forced-choice identification (e.g. Tajima et al., 1997), comprehension questions (e.g. Hahn, 2004), focused interviews (e.g. Zielinski, 2008), written summaries (e.g. Hahn, 2004), speech reception thresholds (e.g. Quené & van Delft, 2010), and transcriptions of nonsense sentences (Kang, Thomson, & Moran, 2018b). Ostensibly, all of these tasks are measuring the same construct, although recent evidence suggests that this may not be the case (Kang, Thomson, & Moran, 2018a). If different measurements are in fact tapping into different constructs, it may not be possible to directly compare results.

In contrast to intelligibility, c*omprehensibility* is defined as the *ease* with which a speaker is understood, regardless of the extent to which they are *actually* understood. It is usually measured by listener ratings on a 7 or 9 point scale (from "easy to understand" to "very difficult to understand"). Studies that employ both intelligibility and comprehensibility tasks have shown that while these constructs are correlated, they are distinct (Derwing & Munro, 1997; Munro & Derwing, 1995a; Winters & O'Brien, 2013). For example, a listener might be able to decode 100% of a speaker's words and yet not rate that speaker as very easy to understand if listening requires a great deal of effort; this speaker would be highly intelligible, but not easily comprehensible.

A third construct discussed by Munro and Derwing is *accentedness*, generally defined as the extent to which speech deviates from a reference accent, usually measured by listener ratings. Importantly, speech that is highly accented can also be intelligible (Munro & Derwing, 1995a). Researchers agree that intelligibility and comprehensibility should be the goal of L2 instruction rather than reduction in accentedness. In this paper, I will focus primarily on intelligibility, although comprehensibility research is also relevant.

Before turning our attention to recent empirical findings, it is important to note that these constructs are not inherent to the speaker, but are a product of a bidirectional interaction. A listener who is familiar with a speaker's accent or whose own accent is similar may perceive that speaker as more intelligible (Stringer & Iverson, 2019) and more comprehensible (Saito, Tran, Suzukido, Sun, Magne, & Ilkhan, 2019). Other moderating factors include the listener's proficiency, willingness to expend effort, and age (Munro & Derwing, 2015).

2.1.2 Intelligibility and Segmentals

Intelligibility research has explored a wide range of different linguistic features (see Levis, 2018 for an overview), but this paper will focus on just one of these—segmentals. Consonant and vowel sounds have traditionally constituted a large part of pronunciation teaching and research. A recent review of 75 pronunciation instruction studies found that 77% included segmentals (Thomson & Derwing, 2015). Given this attention, it is crucial to understand if and when segmentals actually affect intelligibility. Vowel sounds are especially important to study because they appear to be harder to acquire naturally (Neri, Cucchiarini, & Strik, 2006; Munro & Derwing, 2008). In other words, if vowel contrasts are not directly instructed, they may not be learned even over long periods of time. If these lasting errors do not affect intelligibility, they are merely unproblematic aspects of accentedness; if they do affect intelligibility, however, it is important to include them in pronunciation syllabi.

Some evidence that segmentals are implicated in intelligibility comes from Jenkins' analysis of L2S – L2L ("lingua franca") communication (2000, 2002). Based on close analysis of communication breakdowns in recorded interactions between L2 speakers, Jenkins concluded that the majority of miscommunications (68%) were caused by phonological error, with a great deal of these caused by segmental error. This research provides some interesting insights, but it does not tell us much about intelligibility *per se*. Intelligibility is inherently a perception of the interlocutor and cannot be measured by a third party listening in on a conversation. In another corpus-based lingua franca study, Deterding (2013) addressed this limitation by including participants in the transcription process, allowing him some access into their perceptions. Results of this study similarly found that segmental mistakes were responsible for a large percentage of misunderstandings. Such corpus-based studies are appealing because of their ecological validity, but their results are difficult to interpret. As Sewell (2017) points out, an *a posteriori* approach makes it difficult to pinpoint the exact cause of a communication breakdown as several possible causes can co-occur; for example, multiple linguistic errors on the part of the speaker in conjunction with a lack of vocabulary knowledge on the part of the listener might all simultaneously lead to a breakdown in understanding. Furthermore, Sewell argues, simply because a certain linguistic error is not present in a conversation does not mean it does not influence intelligibility; it might in fact mean that this error is of such great importance that speakers have already learned to avoid it.

Other evidence concerning segmentals and intelligibility comes from research using more controlled transcription tasks. 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 accuracy of vowels, but not consonants, was correlated to intelligibility scores and that errors in word-initial position were more harmful than errors in other positions. Because this research is correlational, however, it is again difficult to say anything firm about causation. It could be, for example, that speakers who made vowel errors tended to make other significant pronunciation errors as well and these other mistakes were the true cause of reduced intelligibility.

In Zielinski's (2008) study, three L1 listeners transcribed the extemporaneous speech of three L2 speakers from three different language backgrounds (Korean, Mandarin, and Vietnamese). During the transcription session, listeners were encouraged to share their thinking process and comment on any difficulties they experienced. Results revealed that nonstandard segments in strong syllables played an important part in reducing the intelligibility of all three speakers. Although these results are intriguing, the study was designed as "a series of detailed case studies" (p. 71) and thus is quite limited in its generalisability. Partial confirmation comes from Levis and Im (2015) who used a similar procedure with a similar number of participants.

Perhaps the most promising area in segmental intelligibility research is the concept of functional load (FL). Catford (1987) defines the functional load of a phonemic contrast as "the number of pairs of words in the lexicon that it serves to keep distinct" (p. 88). In other words, segmental substitutions that result in a greater number of minimal pairs have a higher functional load than those that result in relatively few. For example, i/ɪ is a contrast that distinguishes between many minimal pairs (e.g. feet/fit and lead/lid), whereas u/ʊ distinguishes relatively few (e.g. pool/pull); therefore, i/ɪ has a high FL and u/ʊ has a low FL. Catford argues that high FL contrasts should be prioritised in L2 pronunciation teaching and provides a list which categorizes contrasts on a scale from 1 (low FL) to 10 (high FL). Brown (1988) gives a somewhat more complex description of functional load, outlining 12 different variables that must be considered in the calculation, including not only the number of minimal pairs, but also the acoustic similarity of the contrast, the number of pairs that belong to the same part of speech, and other factors. The most important of these, Brown argues, are the abundance of minimal pairs and the "cumulative frequency" of the pair, which is calculated by adding the individual frequencies of each phoneme. Brown created a ranking of contrasts similar to Catford's, ranging from 1% (low FL) to 100% (high FL).

Munro and Derwing (2006) were the first to apply the concept of functional load to a comprehensibility study. Listeners provided comprehensibility ratings for extemporaneous L2 speech that contained high FL consonant errors, low FL consonant errors, or both. Results showed that while low FL errors had a relatively small effect on comprehensibility, high FL errors had a large effect. A more recent study with Japanese speakers of English similarly showed that high FL substitutions (especially consonants) were correlated to lower comprehensibility ratings (Suzukido & Saito, 2019). Although these studies investigated comprehensibility rather than intelligibility (*ease* of understanding rather than *extent* of understanding), they suggest that functional load might be a usable framework for both constructs.

For those who wish to adopt a functional load framework, there are a few issues that should be considered. First, we must remember that Catford's (1987) and Brown's (1988) FL rankings do not include all possible phonemic contrasts. Instead, they only include the FLs of "conflations commonly made by English language learners and typically practiced in pronunciation drill books" (Brown, p. 602). It is unclear if these "typical" conflations are based on empirical evidence. In any case, it is important to note that the lists are not exhaustive and may need to be updated at some point. Second, the exact connection between minimal pairs, functional load, and Munro and Derwing's (2006) findings is not immediately apparent. While minimal pairs are a defining factor of functional load computations, Munro and Derwing actively avoided minimal pair substitutions when choosing the stimuli of their study "so that the effects of word status would not complicate the interpretation of results" (p. 524). Out of the 32 tokens used in their study, only 3 substitutions resulted in a minimal pair, whereas 29 did not (e.g. laywer → naywer). It would seem that the exact mechanism by which functional load affects intelligibility is complex, going beyond the likelihood of pronouncing a different word and tapping into some other factor.

In summary, available research suggests that segmentals significantly affect the intelligibility of L2 speech, justifying its place in pronunciation instruction. There is still some uncertainty concerning which segmentals (or segmental contrasts) are of particular importance, although functional load may provide a useful framework. Methodologically, the majority of recent research uses extemporaneous speech of sentence length or longer, which makes it difficult to control for confounding factors. Further research that controls for these factors could strengthen existing evidence. The most significant methodological issue, however, and one which is almost never discussed, is the reliance on audio-only stimuli. The implication is that visual information is not part of the construct of intelligibility or, if it is, it makes little difference. The validity of such a position will be examined in the next two chapters.

2.2 Visual Speech

2.2.1 Audiovisual Speech Perception

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, Sommers, & Spehar, 2007). For example, if a listener is unsure whether they have heard "bad" or "dad," seeing a speaker's mouth would rule out one or the other. Other segmental contrasts cannot be distinguished in the same way because they share similar visual expressions (e.g. /b/ vs. /m/). These phonemes are said to share the same "viseme."

Studies investigating the effect of visual information on listeners' understanding of speech are more likely to use the term "speech perception" than "speech intelligibility," although methodologically they are the same. Similar to the studies in Chapter 1, speech perception tasks include transcription (e.g. Ma, Zhou, Ross, Foxe, & Parra, 2009), forced-choice identification (e.g. Jongman, Wang, & Kim, 2003), the measurement of speech reception thresholds (e.g. Macleod & Sommerfield, 1987), and comprehension questions (e.g. Arnold & Hill, 2001).

2.2.2 Visual Speech in L1S – L1L Interactions

There is robust evidence from this field of research that seeing visual speech increases intelligibility in L1S – L1L interactions (see Peelle & Sommers, 2015 for an overview). Visual speech is especially beneficial in conditions of noise (Sumby & Pollack, 1954; Erber, 1969; Macleod & Summerfield, 1987; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Ma et al., 2009). On the segmental level, the benefit depends on the saliency of the viseme (Neilsen, 2004). This facilitatory effect varies significantly across listeners, with some benefiting far more than others, possibly because of differences in lip-reading ability and ability to integrate visual and auditory information (Grant, Walden, & Seitz, 1998). In quiet conditions, visual speech makes less of a difference as intelligibility scores reach ceiling levels based on auditory information alone. Even in these cases, however, seeing a speaker's mouth can ease the cognitive demands of listening (Peelle & Sommers, 2015); in other words, visual information can also increase the *comprehensibility* of speech.

2.2.3 Visual Speech in L1S – L2L Interactions

Similarly, visual speech has been shown to increase intelligibility in interactions with L2 listeners. Unlike L1S – L1L communication, this benefit is found in clear conditions as well as conditions of noise (Wang, Behne, & Jiang, 2009; Navarra & Soto-Faraco, 2007; Hazan, Sennema, Faulkner, Ortega-Llebaria, Iba, & Chung, 2006; Hardison, 1999, 2005). As some researchers have pointed out, listening in a second language is its own adverse listening condition, somewhat analogous to the noise of L1S – L1L studies.

An extra consideration in this area of research is the possibility of cultural differences in the relative weighting of auditory and visual information. Some researchers have argued, for example, that Asian listeners may rely more heavily on auditory information than non-Asian listeners (Sekiyama & Tohkura, 1993; Burnham & Lau, 1998). Two main reasons are given for this hypothesis: first, tone languages like Mandarin convey more information in the acoustic signal and listeners may therefore have less need to rely on visual information; second, in countries where there is a culture of face avoidance, people might rely more heavily on auditory information. To test this hypothesis, researchers measure a “McGurk effect” (McGurk & MacDonald, 1976). The McGurk effect 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. Speakers from a variety of language backgrounds have been tested, but Mandarin speakers are of particular importance to the current study. Some research has found a weaker McGurk effect in this population when compared to American listeners (e.g. Sekiyama, 1997). However, a recent study with a much larger sample size found little difference. Magnotti, Mallick, Feng, Zhou, Zhou, and Beauchamp (2015) compared 162 Mandarin speakers from China to 145 English speakers from the US and found similar frequencies of the McGurk effect across groups (48% for the Mandarin speakers and 44% for the English speakers) with culture and language accounting for only 0.3% of the variance. It is safe to conclude that there is little to no difference in visual weighting between these two groups, at least amongst younger demographics (see the full article for a methodological critique of previous research in this area). Notably, the study found a huge amount of variation across participants (0 – 100%), indicating a high degree of individual variability in audiovisual speech integration.

2.2.4 Visual Speech in L2S – L1L Interactions

All of the studies discussed so far have focused on the visible speech of L1 speakers. Such research has found that seeing visual speech enhances intelligibility in most cases, with a high degree of individual variation across listeners. Would the same benefit be found for L2 speakers? Research in this area is relatively scarce, but there have been a handful of relevant studies in recent years. Yi, Phelps, Smiljanic, and Chandrasekaran (2013) found that visual speech enhanced intelligibility for Korean accented speech, but the benefit was less than the benefit found for L1 speech. Kawase et al. (2014) focused on a set of consonant sounds in Japanese accented speech. The consonants included three that are found in Japanese as well as English (/b, s, l/) and three that are found in English, but not Japanese (/ʋ, h, ɹ/). Results showed that overall intelligibility was higher in the audiovisual condition than in the audio condition, but the degree of benefit was much higher in the former set of consonants than in the latter. Notably, the intelligibility of /ɹ/ was actually lower in the audiovisual condition. The authors attribute this to inaccurate articulatory configurations; unlike L1 speakers, the Japanese speakers tended not to round their lips when pronouncing this sound. Because this articulation is not what an L1 listener would expect, it may have reduced intelligibility. This finding suggests that the interaction between visual speech and intelligibility is more complex for L2 speakers than L1 speakers, with a beneficial effect not necessarily guaranteed. For most segmentals, however, there was some degree of benefit. Banks, Gowen, Munro, and Adank (2015) also used Japanese accented stimuli and similarly found that listeners understood more words in the audiovisual condition than in the audio-only condition. Interestingly, in an accompanying experiment, the researchers created an artificial “novel accent” and found the same effect. Most recently, Zheng and Samuel (2019) employed a slightly different methodology, comparing the intelligibility of two Mandarin accented speakers across different “distances” rather than modalities. In one condition, video recordings were close to the speaker’s face and in another condition, the recordings were relatively distant (four meters away). Results showed that intelligibility scores were generally higher in the close condition, suggesting that the benefit of visual speech depends on distance from the speaker. Overall, research indicates that access to visual speech makes L2 speakers more intelligible, but this benefit is relatively small and it is possible that in some instances, it may actually be harmful.

2.2.5 Summary

Studies in audiovisual speech perception clearly show that seeing an L1 speaker’s mouth helps listeners understand. The benefit of visual speech for L2 speaker intelligibility is less certain. The extent of any benefit is likely dependent upon the degree to which individual segments are correctly articulated (in terms of the movements of the mouth, lips, tongue, and teeth). As inaccurate phonemes no doubt often co-occur with inaccurate visemes, one might speculate that the greater the segmental accuracy of a speaker, the greater the potential for visual enhancement (as long as ceiling effects are avoided). Seeming to contradict this hypothesis, Zheng and Samuel (2019) found that the benefit of closeness was greater for a speaker with a relatively strong accent than for a speaker with a relatively weak accent. The exact relationship between phonological accuracy, visemic accuracy, and intelligibility remains unclear. Another unanswered question is whether misleading articulations such as a lack of lip-rounding for /ɹ/ would have the same effect on L2 listeners, who have so far been absent from this area of research.

2.3 Co-Speech Gesture

2.3.1 Gesture and Speech

Kendon (2004) defines gestures as “actions that have the features of manifest deliberate expressiveness”—that is, bodily movements (usually of the hands and arms) that are at least somewhat voluntary and serve a communicative purpose. Actions that are inadvertent (e.g. crying or nervous fidgeting) or that serve a practical aim with no communicative purpose (e.g. eating) are not considered gestures. For the purposes of this study, I am primarily concerned with hand movements that co-occur with speech, sometimes referred to as “gesticulations” (McNeill, 1992; Kendon, 1980). This excludes culturally specific “emblems,” such as the thumbs up or OK sign, as these are highly conventionalised and often stand on their own, apart from speech (McNeill, 1992).

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. The 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, these four types of co-speech gesture are not culturally specific in form, although there are some subtle cross-linguistic and cross-cultural differences (Kita, 2009). Of the four types, iconic gestures are by far the most researched and are also the focus of the current study.

Most researchers now believe that gesture and speech form part of a tightly integrated system during both production and comprehension (McNeill, 1992, 2005; Goldin-Meadow, 2003; Kendon, 2004; Kita & Özyürek, 2003; Graziano & Gullberg, 2018). McNeill explains that the way in which gesture and speech are able to express a similar underlying idea in a different way *at the same time* (synchronously) is a strong indication that “at the moment of speaking, the mind is doing the same thing in two ways, not two separate things” (2005, pp. 22–23). Recent neurobiological research supports this claim by showing that the processing of gestures activates neural responses similar to those activated in speech processing (see Özyürek, 2014 for a review). Overall, there is strong evidence that gesture is an integral, inseparable part of human conversation.

2.3.2 Iconic Gesture in L1S – L1L Interactions

As with visual speech research, gesture has been proven to significantly affect the intelligibility of speech between L1 users. Studies that compare conditions with and without iconic gestures repeatedly find that participants perform better in the gesture condition, especially in noise (Rogers, 1978; Riseborough, 1981; Beattie & Shovelton, 1999; Holler, Shovelton, & Beattie, 2009; Drijvers & Özyürek, 2017). Hostetter’s (2011) meta-analysis of 63 gesture studies found a medium effect size. Further analysis revealed that gestures were particularly beneficial when they conveyed information about space or movement rather than abstract concepts (i.e. iconic rather than metaphoric gestures). No significant difference was found between studies examining spontaneous gesture and those using scripted gesture—a finding which validates the methodologies employing the latter.

2.3.3 Iconic Gesture in L1S – L2L Interactions

It wasn’t until the 1990s that SLA researchers started conducting empirical gesture studies (see Stam & Buescher, 2018; Gullberg, 2008 for overviews). Among these studies, there are relatively few that look specifically at gesture’s effect on intelligibility. One notable exception is Sueyoshi and Hardison (2005). In this 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. The gestures, which had been spontaneously performed by the speaker, were coded using McNeill’s taxonomy and counted for relative frequency; as a percentage of total gestures, 38% were beat, 31% were iconic, 23% were metaphoric, and 8% were deictic. During pauses in the recording, participants answered multiple choice comprehension questions (what Munro and Derwing would call a test of “global intelligibility”). Results 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 order to confirm these results, however, it is necessary to examine a greater variety of speakers, listeners, and intelligibility task types, while also taking into account the linguistic variables in the speech itself.

Dahl and Ludvigsen (2014) expanded upon Sueyoshi and Hardison’s findings by using a picture dictation task rather than comprehension questions. Participants watched recordings of a speaker describing four different cartoons and then drew what had been described. For half of the participants, gestures were visible in the frame, while for the other half, gestures were cropped out in editing (resulting in a close-up of the speaker’s face). Pictures were then coded for accuracy. Results revealed that gestures had a significant beneficial effect for L2 listeners. This supports Sueyoshi and Hardison’s (2005) findings. Both studies suggest that gesture increases the intelligibility of L1 speakers for L2 listeners in clear conditions (without noise). Drijvers and Özyürek (2017, 2019) found similar results in conditions of noise. Based on word level transcriptions, highly proficient L2 listeners were seen to benefit from iconic gesture, although the enhancement effect was greater for L1 listeners.

On a methodological note, it is important that these L2 comprehension/perception studies include an audiovisual condition without gesture. Studies that compare a gesture condition to an audio-only condition (e.g. Beattie & Shovelton, 1999) are conflating the effect of visual speech and gesture. On the other hand, the way in which Sueyoshi and Hardison (2005) and Dahl and Ludvigsen (2014) create their visual no-gesture condition is slightly problematic. In both studies, a close-up was used for the no-gesture condition (for practical reasons) and a medium shot was used for the gesture condition. Given the effect of distance on speech perception (Zheng & Samuel, 2019), it may be important to control for perceived closeness to the speaker’s face. Drijvers and Özyürek (2017, 2019) controlled for this variable by using medium shots (from the knees to the head) in all audiovisual conditions.

2.3.4 Iconic Gesture in L2S – L1L/L2L Interactions

As with visual speech research, the majority of gesture research has focused on L1 speakers. Very little is known about the effect of gesture in L2 speech intelligibility. What we do know is that speakers tend to use more co-speech gestures when speaking in a second language (e.g. Gullberg, 1998; Sherman & Nicoladis, 2004). Research suggests that these gestures are sometimes used as a communication strategy to help in situations of expressive difficulty (Gullberg, 2008). As of yet, however, there are no intelligibility studies that empirically investigate the actual effect of L2 gestures on understanding. There are a few case studies which look at L2 gesture in speaking assessments (see Stam & McCafferty, 2008 for a brief overview), but these are based on ratings of oral proficiency rather than measurements of actual understanding.

2.3.5 Summary

Gesture and speech are part of an integrated system in which both modalities contribute to meaningful communication. Although it is clear from L1 research that the removal of gestural information can reduce the intelligibility of a speaker, there is relatively little known about this phenomenon in L2 research. The little research that does exist focuses on L2 *listeners* rather than L2 *speakers*.

3 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 in the presence of visual cues. The few studies that do include a visual modality have explored a limited set of pronunciation features and have focused solely on facial cues, leaving gesture out of the conversation. A recent L1 speech perception study that could be used as a model for such research comes from Drijvers and Özyürek (2017, 2019), who investigated the intelligibility of speech across several different modalities (audio-only vs. audio + visual speech vs. audio + visual speech + iconic gesture vs. iconic gesture only) and across several different levels of speech clarity (levels of noise-vocoding). They also compared the effect of listener background (L1 vs. L2) over two experiments. As the authors note, it is rare to investigate the effect of visual speech *and* iconic gesture in the same experiment. Their study provides a controlled design that allows for such comparisons.

Drijvers and Özyürek included many of the elements that are central to the current study, but whereas they defined speech clarity in terms of acoustic degradation, The current study investigated clarity in terms of the presence or absence of *phonological error*, specifically segmental error. Segmental errors were artificially created by systematically shifting vowel sounds in monosyllabic words, creating a “novel accent” similar to that of Banks et al. (2015). Two main research questions were explored:

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

In this study, L2 listeners were Mandarin L1 users from China and Taiwan. Based on previous research, it was possible to make a few hypotheses. Regarding the first question, it was predicted that vowel errors would reduce intelligibility for L1 listeners. It was also predicted that the addition of visual cues would increase intelligibility for L1 listeners, especially when there were vowel errors. Based on previous research, visuals are less likely to have an effect in standard speech conditions (without vowel errors), as intelligibility may reach ceiling levels based on audio information alone.

Regarding the second question, it was predicted that L2 listeners would similarly suffer from vowel errors and benefit from visual information. In contrast to predictions made for L1 listeners, it was hypothesized that L2 listeners would benefit from visual cues in *both* pronunciation conditions (standard speech as well as vowel error). Finally, it was hypothesized that both L1 and L2 listeners would benefit more from visual speech in the standard pronunciation condition than in the vowel error condition, based on the findings of Yi et al. (2013) and Kawase et al. (2014).

4 Methodology

4.1 Introduction

Two experiments were conducted, partially modelled on Drijvers and Özyürek's (2017, 2019) design. In the first experiment, L1 English users from the UK and the US watched 60 stimuli in six conditions (see Figure 1) and completed an intelligibility task. Stimuli consisted of common monosyllabic verbs pronounced with or without vowel errors in three different modalities:

- audio-only (video with the speaker's mouth obscured)
- visual speech (audio + visual speech)
- gesture (audio + visual speech + iconic gesture)

Mean intelligibility scores in each condition were computed. In the second experiment, Mandarin speakers from China and Taiwan who were highly proficient in English as a second language completed the same intelligibility task. All participants additionally completed the Visual Cue Preference Questionnaire (VCPQ), adapted from Sueyoshi and Hardison (2005), which provided information about their beliefs and preferences regarding facial cues and gesture.

4.2 Participants

In Experiment 1, participants were 10 L1 users of English (5 women and 5 men, M_{age} = 36, *SD* = 11.16, with 2 participants choosing not to reveal their age). Participants were postgraduate students at University College London, most of whom were British, with one participant from the US. All participants had English language teaching experience or some knowledge of English phonology (most had both). Almost all participants had studied a second language, reaching various levels of proficiency.

In Experiment 2, participants were 22 L2 users of English from the same university (21 women and 1 man, M_{age} = 26.64, SD = 2.63). Compared to the participants in the first experiment, there was a greater percentage of women and a lower mean age. Participants were L1 users of Mandarin from China (n = 20) and Taiwan (n = 2). All participants were highly proficient in English, scoring a minimum of 7 on the IELTS or 100 on the TOEFL. They reported speaking English at least 50% of the time at school or at home. Participants had lived in the UK or another English-speaking country for 9–16 months. Some participants were also fluent in a third language. In order to control for daily English use and length of residence in the UK, two participants (from an original group of 24) were not included in the data analysis. All participants in both experiments reported having normal hearing and normal or corrected vision that would not impede perception of the video in the intelligibility task.

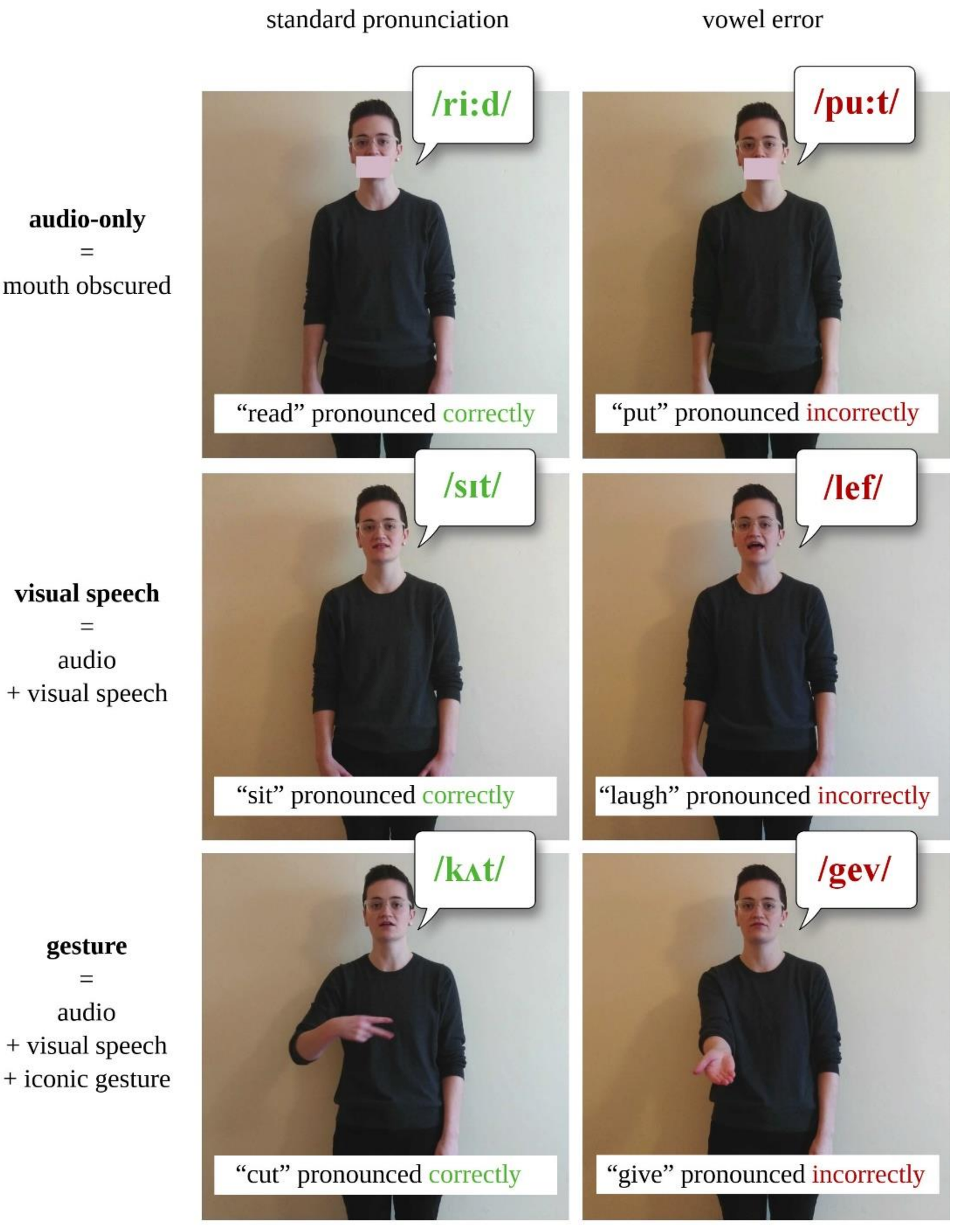


Figure 1 Overview of the Design and Conditions Used in the Experiments

4.3 Materials

4.3.1 Stimulus Materials

In order to control for a single pronunciation variable, extemporaneous speech was not used. Instead, the author (an American L1 English speaker) recorded a chosen set of words, some of which contained particular pre-determined vowel errors, in a fashion similar to Field (2005) and Hahn (2004). Materials consisted of 60 2-second video clips covering 6 different conditions (10 per condition). Extra stimuli were recorded as a precaution in case any needed to be removed after gesture piloting (see section 4.4) or for any other reason.

In each clip, the speaker was video-recorded saying a single monosyllabic action verb (see Appendix A for the full list). All verbs fell within the most frequent 2,000 word families in the BNC-COCA-25 wordlist, confirmed through Cobb's website *The Compleat Lexical Tutor* (2019). Only relatively frequent verbs were chosen so that the results of L2 participants would not be influenced by vocabulary knowledge. The speaker for the stimulus materials had a General American (GA) accent as confirmed by comparing her speech to pronunciations provided in the *Longman Pronunciation Dictionary* (Wells, 2008). The speaker was recorded in front of a neutral background from her knees to her head.

In vowel error conditions, vowels were systematically shifted one step toward the front or back or one step more open or closed (see Table 1). The systematic shifting of vowel sounds is similar to the methodology used by Banks et al. (2015) to create their "novel accent." Out of 30 total substitutions made, 20 were high FL contrasts (e.g. æ/e), 5 were low FL contrasts (e.g. u:/ʊ), and 5 were of unknown FL (i:/e) according to the ranks of Catford (1987) and Brown (1988), where "high" FL is defined as a FL of greater than 5 (Catford) or 50% (Brown). Presumably, the contrast of unknown FL (i:/e) does not appear in Catford or Brown's lists because it was considered to be an error that is not commonly made by L2 learners (see section 2.1.2). High FL contrasts were approximately balanced across vowel error conditions, comprising at least 50% of stimuli.

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/
e ↔ ɪ	smell (/smel/) pronounced as /smɪl/ give (/gɪv/) pronounced as /gev/
i: ↔ e	teach (/ti:tʃ) pronounced as /tetʃ/ press (/pres/) pronounced as /pri:s/
æ ↔ e	laugh (/læf/) pronounced as /lef/ let (/let/) pronounced as /læt/
u: ↔ ʊ	move (/mu:v/) pronounced as /mʊv/ push (/pʊʃ/) pronounced as /pu:ʃ/

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 RP English (e.g. /sɪt/ → /set/) were not included. 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 participants, were considered. Although the avoidance of minimal pairs might seem to artificially exclude the most egregious of errors, it was shown in Munro and Derwing’s (2006) study that segmental substitutions (especially high FL substitutions) are likely to cause misunderstandings regardless of minimal pair status. In addition, substitutions that result in a different word are actually relatively rare in L2 speech (Sewell, 2017). In order to check that the intended vowel shifts were performed and shifts did not exceed the “one step” rule, audio recordings of all vowel error stimuli were phonetically transcribed by a researcher familiar with the IPA who was unfamiliar with the chosen words. One stimuli was removed after this process.

In gesture conditions, the speaker performed a scripted iconic gesture alongside the spoken verb. In order to confirm the appropriacy of the gesture, a small piloting study was conducted before the main experiment (see section 4.4). In each stimulus, the “stroke” of the gesture (the part of the movement that bears meaning) co-occured with the spoken word. The “preparation” phase of the gesture (when the arm begins to move from a resting position) began approximately 150–200 milliseconds after the start of the clip.

Audio and video were separated during editing. Audio files were de-noised in Audacity (Team, 2016). The audio was then trimmed into clips of approximately 2 seconds in length and intensity was scaled to 70 dB in Praat (Boersma & Weenink, 2018). Audio and video were then reintegrated in Kdenlive (2016). Following Drijvers and Özyürek (2017, 2019), audio-only conditions were created by obscuring the speaker’s mouth (see Figure 1).

4.3.2 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, with six seconds of silence in between stimuli to leave ample time for transcription. Audio and video of item numbers (1–60) preceded each stimulus. A paper handout with instructions and numbered lines was prepared for participants to record their transcriptions (see Appendix B). Intelligibility per condition was operationalised as the percentage of words correctly transcribed in that condition.

4.3.3 Questionnaire

The questionnaire (see Appendix C) was adapted from one section of the "Visual Cue Preference Questionnaire" (VCPQ) created by Sueyoshi and Hardison (2005) and also used by Dahl and Ludvigsen (2014). All items were 5-point Likert scales (where 1 = *strongly disagree* and 5 = *strongly agree*). Items 1–9 concerned participants' beliefs, preferences, and behaviours 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 Sueyoshi and Hardison's version of the questionnaire to suit the purposes of the current study. For example, whereas Sueyoshi and Hardison used the statement, "It is easier to understand English when I can see the speaker's face," I split this into two questions—one for *L1* English speech and one for *L2* English speech. L1 Mandarin participants and L1 English participants received slightly different versions of the questionnaire, although the items were made as parallel as possible.

4.4 Piloting the Gesture Stimuli

Before the final set of stimuli was chosen, all gestures were piloted with a group of 5 L1 English users from the US (2 men and 3 women, aged 32–67), none of whom took part in the main experiments. Participants completed the task via online video chat with the author. The primary purpose of the pilot test was to ensure that gestures were appropriate and not misleading. The procedure for the test was based on Drijvers and Özyürek (2017). First, participants watched the gesture stimuli (without audio and with mouth obscured) and were asked to write down the verb they thought the movement communicated. Answers that matched the intended verb or gave a synonym (e.g. "close" instead of "shut") were coded as correct. Then, the intended gesture/verb pairing was revealed and participants were asked to rate how well this verb matched the movement on a 7-point Likert scale (1 = *doesn't fit the movement at all*, 7 = *fits the movement very well*, 4 = *I'm not sure*). Any gestures that were not considered appropriately matched (those that received a mean score of less than 5), were discarded.

For the final stimuli chosen, it was then possible to calculate a "gesture recognition rate," the percentage of correct answers based on gesture alone. When exact matches as well as synonyms 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. Additionally, it must be kept in mind that synonyms will not be coded as correct in the intelligibility task of the main experiment. Overall, the results of the pilot study suggested that the gestures used in the main experiments would likely help disambiguate the spoken message, although they would not be *sufficient* in and of themselves for speech to be intelligible.

4.5 Procedure

Prior to the start of the experiment, participants were provided with an information sheet (see Appendix E) and signed a consent form (see Appendix F), which informed them that their identity would be kept confidential and that they could withdraw from the study at any time. Experiment 1 and Experiment 2 followed the same procedure. First, instructions for the intelligibility task were given orally as well as provided on the 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 actress is trying to communicate? If you are not sure, please try to guess.*

During the task, participants were individually placed in a quiet room and watched the video on an 11.6” computer screen, wearing headphones. They started with a trial of six items (one per condition) and were allowed to ask questions before proceeding. A few participants asked a version of this question: “If I feel like the audio and gesture are giving different information, which one should I pay attention to?” Being postgraduate students familiar with experimental designs like that used in the “McGurk effect,” these participants were likely wondering how they should behave in the case of incongruent audio-visual stimuli, which are actually not present in the current study. The researcher responded to this question indirectly by saying, “Imagine that the actress is a real person who you have encountered somewhere. Perhaps she has an accent that you are unfamiliar with. How would you interpret her speech in this situation?” During the transcription task, participants were able to record answers for the vast majority of items, with some participants leaving a few items blank. Based on the researcher’s observations and discussions that followed, it seemed that all participants felt they had guessed for at least some of the words. L1 participants tended to say that they had guessed for a few items, whereas L2 participants ranged in their reaction to the task, with some saying they had guessed for a few and others feeling like that had guessed for a lot. Some participants circled or marked items when they felt very uncertain; on average, these items comprised about 5–15% of their total responses.

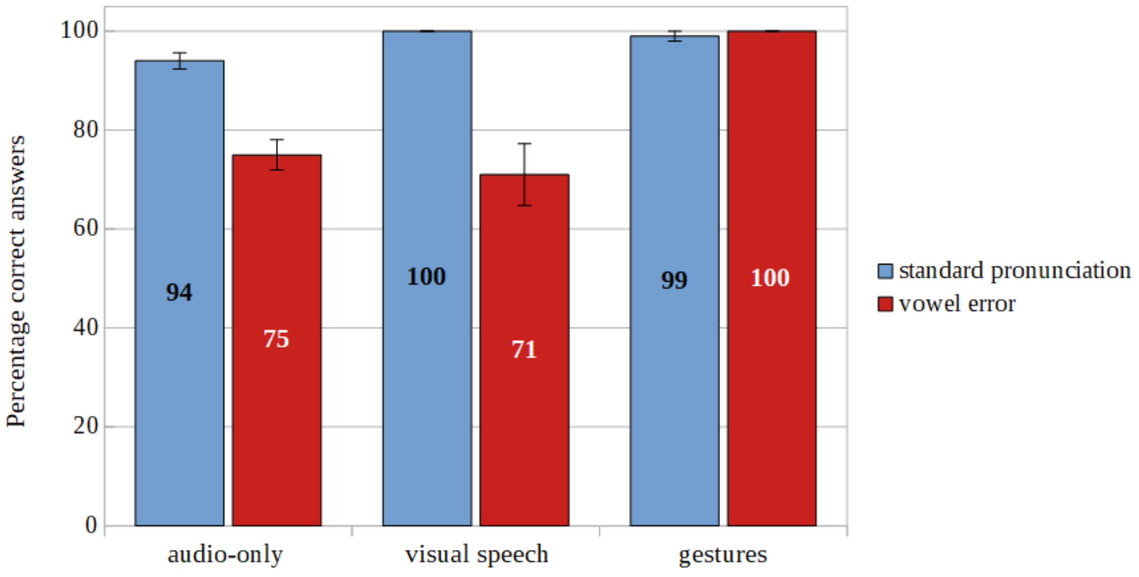
After finishing the transcription task, participants completed the Visual Cues Preference Questionnaire. Finally, participants were asked to provide some biographical information (see Appendix D). In a short debriefing session, participants were thanked for their participation.

Some participants offered their thoughts on the experiment and their own hypotheses about possible results and implications. In total, the experiment lasted 30 to 50 minutes per participant.

5 Results

5.1 Results: L1 Listeners

Experiment 1 explored the effect of vowel error and visual cues on the intelligibility of speech for L1 listeners. Transcription task scores across conditions reveal a few interesting patterns (see Figure 2). First, it is clear that vowel error conditions are generally less intelligible than correct pronunciation conditions. Second, it appears that visual cues do not make too much of a difference for L1 listeners under standard pronunciation conditions; speech is largely intelligible across modalities. However, visual cues (at least gesture) do seem to make a difference in vowel error conditions. To confirm that these differences between conditions were statistically significant, several nonparametric tests were run. Parametric tests could not be used as several of the conditions were not normally distributed (due to ceiling effects).



Error bars represent *SE*.

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

5.1.1 The Effect of Vowel Error

The effect of vowel error was tested using the Wilcoxon Signed Ranks test, the nonparametric equivalent of a paired samples *t*-test. Two tests were run—one in the audio-only modality and one in the visual speech modality. No test was performed in the gesture modality as intelligibility scores were nearly identical in this condition (*Mn* = 99% vs. 100%). A Bonferroni correction was applied to protect against inflation of the familywise Type I error rate, resulting in a corrected α level of .025. The effect size *r* was calculated from the *Z* statistic (Rosenthal, 1991):

$$r = \frac{Z}{\sqrt{N}} \tag{1}$$

Results showed that in the audio-only modality, intelligibility scores were significantly higher in the standard pronunciation condition (*Mn* = 94; *Mdn* = 90) than the vowel error condition (*Mn* = 75; *Mdn* = 80), *T* = 0, *p* = .004 (two-tailed), with a large effect size (*r* = .60). Similarly, in the visual speech modality, intelligibility scores were significantly higher in the standard pronunciation condition (*Mn* = 100; *Mdn* = 100) than in the vowel error condition (*Mn* = 71; *Mdn* = 70), *T* = 0, *p* = .008 (two-tailed), with a large effect size (*r* = .57). We can conclude that vowel errors reduced intelligibility in these conditions, as hypothesized. Interestingly, there was no observable detrimental effect of vowel error in the gesture modality, suggesting that such an effect was fully mitigated by the enhancement effect of gesture.

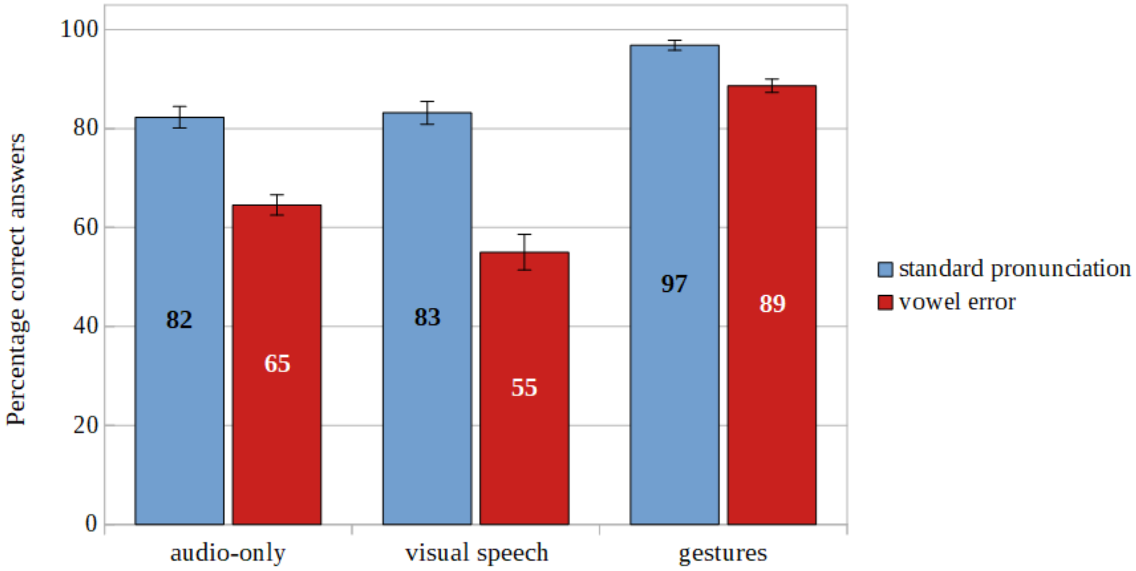
5.1.2 The Effect of Visual Cues

Another set of Wilcoxon Signed Ranks tests was used to investigate the effect of visual cues. Possible differences were examined in the vowel error conditions only, as intelligibility scores were near ceiling when standard pronunciation was used. The α level was corrected to .0167. Results showed that intelligibility scores were significantly higher in the gesture condition (*Mn* = 100; *Mdn* = 100) than the visual speech condition (*Mn* = 71; *Mdn* = 70), *T* = 0, *p* = .008 (two-tailed), with a large effect size (*r* = .57). In fact, intelligibility scores reached 100% in the gesture + vowel error condition, suggesting that the benefit of gesture was able to completely compensate for the negative effect of vowel error. Contrary to predictions, however, scores did not significantly differ in the visual speech condition and the audio-only condition (*Mn* = 75; *Mdn* = 80), *T* = 11, *p* = .609 (two-tailed). Surprisingly, the mean score in the visual speech + vowel error condition was actually *lower* than the mean score in the audio-only + vowel error condition by 4 percentage points, although this difference was not significant. Notably, there was a large standard deviation in the visual speech + vowel error condition (*SD* = 19.69%), suggesting a significant amount of individual variation.

Overall, results show that gesture, but *not* visual speech, has a significant beneficial effect when L1 listeners are perceiving speech with vowel errors. When perceiving standard speech, visual cues make less of a difference as the auditory signal alone provides enough information to reach high degrees of intelligibility. These results partially confirm predictions for the first research question.

5.2 Results: L2 Listeners

In general, intelligibility scores for L2 listeners were lower than for L1 listeners across conditions (see Figure 3). Mean scores suggest that similar to L1 listeners, L2 listeners were negatively affected by the presence of vowel errors. Unlike L1 listeners, L2 listeners appeared to benefit from gesture in both the vowel error conditions *as well as* the standard pronunciation conditions. However, the benefit of gesture did not fully mitigate the harmful effect of vowel error, as it had with L1 listeners; in other words, there is an observable negative effect of vowel error in each modality, including gesture. A repeated measures ANOVA was run to investigate differences in intelligibility scores across two factors (pronunciation and modality). Before this test could be performed, it was necessary to meet the assumptions of a parametric test.



Error bars represent *SE.*

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

5.2.1 Meeting the Assumptions of a Parametric Test

In order to better meet the assumptions of normal distribution, the raw data was transformed using SPSS. The best transformation to correct for skewness was:

$$new\ variable = (110 - original\ variable)^{1/3} \tag{2}$$

Another transformation was considered in order to reduce kurtosis in the gesture + vowel error condition:

$$new\ variable = sign(Y)\left[\frac{(Y+1)^{1/10}-1}{1/10}\right] \qquad (3)$$

$$\text{where} \quad Y = (100 - original\ variable)^{1/3} - 2.9$$

Because SPSS does not have a *sign* function, this was performed in multiple steps with “if” conditions:

$$\text{if } \; Y > 0, \; new\ variable = \left[\frac{(Y+1)^{1/10}-1}{1/10}\right]$$

$$\text{if } \; Y < 0, \; new\ variable = \left[\frac{(Y+1)^{1/10}-1}{1/10}\right] \qquad (4)$$

The two resulting sets were then combined using the “max” function. This transformation effectively normalised the kurtosis value for the gesture + vowel error condition, but introduced a new kurtosis problem in another condition. No variation of transformation (3) was found that effectively normalised kurtosis across *all* conditions, so it was ultimately discarded and transformation (2) was used instead.

While the transformed data was not entirely normal, it was deemed sufficiently normal to run parametric tests. In nearly all conditions, the absolute value of *z*-scores for skewness and kurtosis were less than 1.96, often provided as a guideline for normalcy (Field, 2018; Kerr, Hall, & Kozub, 2002). The only exception was the *z*-score for kurtosis in the gesture + vowel error condition, which was 2.39. Overall, the data can be considered *approximately* normally distributed.

Other assumptions were also examined, including absence of outliers and homogeneity of covariance. Visual inspection of boxplots revealed an outlier in the visual speech + vowel error condition. Analyses were run with and without the outlier and results were essentially the same; the reported results include the outlier case. Finally, Mauchly’s Test of Sphericity did not show significance for modality or for the interaction between modality and pronunciation. Therefore, the data met the assumption of homogeneity of covariance.

5.2.2 Results of a Two-Factor Repeated Measures ANOVA

ANOVA results revealed a significant main effect of modality on intelligibility scores, *F*(2, 42) = 101.38, *p* < .001, partial η2 = .828, as well as a significant main effect of pronunciation (standard pronunciation versus vowel error), *F*(1, 21) = 139.25, *p* < .001, partial η2 = .869. In addition, there was a significant interaction between modality and pronunciation, *F*(2, 42) = 5.196, *p* = .01, partial η2 = .198. To unpack these results, further post-hoc tests were performed.

5.2.3 The Effect of Vowel Error

Three paired samples *t*-tests were run to investigate the effect of vowel error. The α level was corrected to .0167. Effect size *r* was not calculated from the *t*-value as this can lead to overestimations (Dunlap et al., 1996). Instead, *r* was calculated from Cohen’s *d* (Cohen, 1988; Rosenthal, 1994):

$$r = \frac{d}{\sqrt{d^2 = 4}} \qquad d = \frac{M_2 - M_1}{SD_{pooled}} \tag{5}$$

Results revealed that intelligibility scores were significantly lower in vowel error conditions in all modalities: in the audio modality, *t*(21) = 5.48, *p* < .001 (two-tailed), *r* = .66; in the face modality, *t*(21) = 8.65, *p* < .001 (two-tailed), *r* = .72; and in the gesture modality, *t*(21) = 4.91, *p* < .001 (two-tailed), *r* = .61; all effect sizes were large. Like L1 listeners, L2 listeners found speech less intelligible when there were vowel errors. Unlike the L1 listener group, the effect was apparent in every modality, including gesture.

It is possible to compare the effect of vowel error across groups more precisely. Field (2005) used the following formula to quantify loss of intelligibility:

$$decrement = \frac{standard\ condition\ score - nonstandard\ condition\ score}{standard\ condition\ score} \tag{6}$$

When this formula is applied to participants in Experiment 1 and 2, we find very similar decrement percentages across groups. In the audio-only condition, both groups experienced a decrement of 20%. This suggests that vowel error harms L1 listeners and L2 listeners to the same extent. Loss of intelligibility was less in the gesture conditions (0% for L1 listeners and 8.2% for L2 listeners), suggesting that the presence of gesture is able to partially or fully reverse the negative effect of vowel error.

5.2.4 The Effect of Visual Cues

Four paired samples *t*-tests were run to investigate the effect of visual cues. The α level was corrected to .0125. A significant difference was found between the visual speech and gesture modalities, both in standard pronunciation conditions, *t*(21) = 5.59, *p* < .001 (two-tailed), *r* = .62, and vowel error conditions, *t*(21) = 10.78, *p* < .001 (two-tailed), *r* = .83; both effect sizes are large.

However, contrary to predictions, there was no significant difference between the audio-only modality and the visual speech modality (at an α level of .0125), either in correct pronunciation conditions (*p* = .482) or vowel error conditions (*p* = .04). Although not statistically significant, it is notable that mean scores were lower in the visual speech + vowel error condition than in the audio + vowel error condition. This mirrors the pattern found in Experiment 1. Possible explanations will be explored in the following chapter. Similar to L1 listeners, L2 listeners showed a great deal of individual variation in the visual speech + vowel error condition (*SD* = 16.83%).

While it would be interesting to compare the benefit of gesture in Experiment 1 and Experiment 2, this proves difficult. If we simply compare the effect sizes, it would seem that the gesture benefit is larger for L2 listeners (*r* = .83) than L1 listeners (*r* = .57). However, a direct comparison of effect sizes may be misleading. First, L1 listeners' intelligibility scores reached ceiling in the gesture condition, imposing a limit on any achievable effect. It is entirely possible that the difference between the visual speech + vowel error condition and the gesture + vowel error condition would have been larger had ceiling levels been avoided (such as in conditions of noise). Second, it must be taken into account that increases in intelligibility become exponentially more difficult to achieve as one reaches 100% (i.e. an increase from 80% to 90% is not equivalent to an increase from 50% to 60%) (Sommers, Tye-Murray, & Spehar, 2005). A comparison of effect sizes would therefore be biased against L1 listeners, who had higher baseline scores. In order to more fairly compare participants from different populations, many speech perception researchers calculate a "visual enhancement" effect or "gesture enhancement" effect using the following equations (e.g. Grant et al., 1998; Yi et al., 2013; Drijvers & Özyürek, 2017):

$$visual\ enhancement = \frac{visual\ speech\ score - audio\ score}{100 - audio\ score}$$

$$gesture\ enhancement = \frac{gesture\ score - visual\ speech\ score}{100 - visual\ speech\ score} \tag{7}$$

Using this measurement, someone who correctly transcribed 80% of the words in the audio-only condition and 90% in the visual speech condition would receive the same visual enhancement score as someone who correctly transcribed 20% of the words in the audio-only condition and 60% in the visual speech condition (Sommers et al., 2005). This computation corrects for the bias against participants with high baseline scores, but it is less useful when intelligibility scores reach 100% in the enhanced condition; in these cases, the enhancement effect will likewise be 100% regardless of the score in the baseline or reference condition. In the current study, a gesture enhancement effect can be legitimately calculated for L2 participants (the effect is 74.1%), but it cannot be calculated in a comparable manner for L1 participants (it would simply be 100%). In summary, it is difficult to conclude whether L1 or L2 listeners benefited more from the presence of gestures in vowel error conditions. What can be said is that both groups experienced a large benefit.

5.3 Results of the Questionnaire

Mean scores and standard deviations for the 13 Likert items of the questionnaire are reported in Tables 2 and 3. In previous analyses of this questionnaire, Sueyoshi and Hardison (2005) and Dahl and Ludvigsen (2014) ran a series of independent samples *t*-tests to look for differences between groups. While this practice is common, it is not entirely statistically sound as such data is unlikely to meet the assumptions of a parametric test. Moreover, it is debatable whether single Likert items can even be treated as interval data (Larson-Hall, 2010). Therefore, instead of running *t*-tests on individual items, the current study explored the possibility of using more rigorous methods.

Table 2. *Preferences for Visual Cues in Daily Life*

Item	Focus	Mandarin L1 Mean	*SD*	English L1 Mean	*SD*
1	Preference for seeing a speaker’s face to understand L1 English	4.00	1.02	3.40	1.08
2	Preference for seeing a speaker’s face to understand L2 English	4.64	.73	4.10	1.20
3	Preference for seeing a speaker’s gestures to understand L1 English	3.95	1.00	3.90	.99
4	Preference for seeing a speaker’s gestures to understand L2 English	4.64	.58	4.50	.53
5	More gestures used in L2 than L1	3.59	.80	3.20	1.23
6	Perceived contribution of gestures to comprehension of participant’s L2 speech	3.82	.73	3.80	.92
7	Perceived contribution of gestures to comprehension of participant’s L1 speech	3.45	1.01	3.20	.92
8	Attention paid to a speaker’s lip movements (in daily life)	3.27	.99	2.20	.63
9	Attention paid to a speaker’s gestures (in daily life)	4.09	.87	3.80	.63

Table 3. *Preferences for Visual Cues in the Experiment*

Item	Focus	Mandarin L1 Mean	*SD*	English L1 Mean	*SD*
10	Attention paid to the speaker’s lip movements (in the experiment)	2.86	1.04	2.20	1.32
11	Attention paid to the speaker’s gestures (in the experiment)	4.45	.74	4.40	.97
12	Seeing the speaker’s gestures helped comprehension (in the experiment)	4.82	.40	4.60	.52
13	Seeing the speaker’s lips helped comprehension (in the experiment)	3.27	.94	2.10	1.10

5.3.1 Creating Multi-Item Scales

While the questionnaire was not intentionally constructed by Sueyoshi and Hardison (2005) as a set of multi-item scales, it is possible that certain statements are in fact correlated to each other as aspects of a larger construct. For example, item 1 (“It is easier to understand L1 English users when I can see the speaker’s face”) and item 8 (“In face-to-face communication, I pay attention to the speaker’s lip movements”) could both be tapping into general beliefs about the potential informativeness of facial cues. Based on this hypothesis, internal consistency reliability tests were run for two groups of Likert items—a group of items relating to visual speech and another group of items relating to gesture. For statements regarding visual speech (items 1, 2, 8, 10, and 13), Cronbach’s Alpha was .81, suggesting that these items could indeed be combined into a multi-item scale. Similarly, gesture statements (items 3, 4, 5, 6, 7, 9, 11, 12) hung together, albeit with a slightly lower Cronbach’s Alpha of .74. 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 (see Table 4). After squaring the raw data, the scales were found to be normally distributed. It was thus possible to perform *t*-tests.

Table 4. *Attitude Toward Visual Speech and Gesture (Scales Range from 1–5)*

Multi-Item Scale	Mandarin L1 Mean	*SD*	English L1 Mean	*SD*	*t*-value
Attitude Toward Visual Speech	3.61	.15	2.80	.21	**2.76***
Attitude Toward Gesture	4.45	.74	4.40	.97	--

Notes. t-values are based on transformed data * *p* = .01

5.3.2 Differences Between and Within Groups

Results of an independent samples *t*-test showed that L2 listeners had a significantly more positive attitude toward visual speech than L1 listeners, $t(30) = 2.76$, $p = .01$, with a large effect size ($d = 1.1$). This is not to say that all L1 listeners had a negative attitude and all L2 listeners had a positive attitude. A closer analysis of individual item responses revealed that there was division across agree/disagree lines for several of the items. For example, Mandarin L1 participants were somewhat mixed in their response to item 8 (with 45% agreeing or strongly agreeing, 27% disagreeing or strongly disagreeing, and 27% unsure). Taking this variation into account, it is still possible to conclude that Mandarin L1 participants tended to have a more positive attitude toward visual speech than English L1 participants. There was no such difference between groups in attitude toward gesture. Participants generally agreed that gesture was beneficial for comprehension ($Mn_{\text{Mandarin}} = 4.45$, $Mn_{\text{English}} = 4.4$). A paired samples *t*-test revealed that when the two groups were combined, all participants had a significantly more positive attitude toward gesture than visual speech, $t(31) = 5.56$, $p < .001$, with a large effect size ($d = 1.01$). This finding aligns with the results of Experiments 1 and 2, in which gesture but not visual speech facilitated intelligibility.

Although statistical tests were not run on individual Likert items, there are a few differences across these items that are easily observable. Admittedly, it has not been proven that these differences are significant. First, it seems that participants had a stronger preference for seeing a speaker’s face to understand L2 English ($Mn_{\text{Mandarin}} = 4.64$, $Mn_{\text{English}} = 4.10$) than L1 English ($Mn_{\text{Mandarin}} = 4.00$, $Mn_{\text{English}} = 3.40$). Similarly, participants had a stronger preference for seeing a speaker’s gestures to understand L2 English ($Mn_{\text{Mandarin}} = 4.64$, $Mn_{\text{English}} = 4.50$) than to understand L1 English ($Mn_{\text{Mandarin}} = 3.95$, $Mn_{\text{English}} = 3.90$). This aligns with the results of the experiment: gesture had a larger beneficial effect in conditions containing vowel errors (comparable to an unfamiliar accent) than in standard pronunciation conditions.

One final note regarding terminology deserves mention. Within the multi-item scale of “attitude toward visual speech,” there are references to lip-reading as well as seeing a speaker’s entire face. We must be careful, however, not to conflate “facial cues” and “lip-reading” as the former is a broader construct which includes movements of the eyes and eyebrows. While most participants agreed that seeing a speaker’s face aided comprehension, answers were more mixed when it came to attention given to a speaker’s lips. This discrepancy suggests that participants believe there is other useful information from the face apart from lip movements.

6 Discussion

There are two primary findings of this study. First, iconic gestures significantly increase the intelligibility of speech in two situations: (a) when speech contains vowel errors and/or (b) when the listener is an L2 user of the language. The only situation in which gesture does not facilitate intelligibility is when L1 listeners perceive speech in a standard accent; in this case, intelligibility is near ceiling in every modality. The second main finding of the study is that vowel error significantly reduces the intelligibility of speech for both L1 and L2 listeners and that the size of this reduction is similar for both groups. Results are summarized in Table 5. A more detailed discussion of the results follow.

Table 5. *Effect of Visual Cues and Vowel Error in L1 and L2 Listeners*

Effect	Condition	L1 English Listeners	L2 English Listeners
Positive Gesture Effect *intelligibility of visual speech vs. gesture*	standard pronunciation	*n.s.*	*r* = .62** (large)
	vowel error	*r* = .57* (large)	*r* = .83** (large)
Visual Speech Effect *intelligibility of audio-only vs. visual speech*	standard pronunciation	*n.s.*	*n.s.*
	vowel error	*n.s.*	*n.s.*
Negative Vowel Error Effect *intelligibility of standard pronunciation vs. vowel error*	audio-only	*r* = .60* (large)	*r* = .66** (large)
	visual speech	*r* = .57* (large)	*r* = .72** (large)
	gesture	*n.s.*	*r* = .61** (large)

n.s. = no significant effect was found

* *p* < .01, significant at a corrected α level of .0167–.025

** *p* < .001, significant at a corrected α level of .0125–.0167

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

Experiment 1 investigated the effect of visual speech, iconic gesture, and vowel error on the intelligibility of speech for L1 listeners. Results partially confirm hypotheses. As predicted, vowel error reduced intelligibility and iconic gesture increased intelligibility. However, contrary to predictions, visual speech had no significant effect.

6.1.1 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 + vowel error condition were significantly higher than scores in the visual speech + vowel error condition, with a large effect size (r = .57). In fact, gestures were shown to fully mitigate the negative effect of vowel error, bringing scores in the gesture + vowel error condition to 100%. These results suggest that when speech contains linguistic errors, iconic gestures can help to disambiguate the message by providing additional semantic information. It can be argued that this interaction between gesture and nonstandard speech is analogous to the interaction between gesture and speech in noise or noise-vocoded speech; in all of these adverse listening conditions, gesture can significantly affect understanding.

6.1.2 The Effect of Visual Speech

Somewhat unexpectedly, visual speech did not facilitate intelligibility for L1 listeners. No significant differences were found between audio-only conditions and visual speech conditions. This result seems to contradict previous research which has found an enhancement effect for visual speech (e.g. Sumby & Pollack, 1954; Kawase et al., 2014). There are three possible reasons 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 11.6” screen (1366 x 768). While some studies have used medium shots, these are usually presented on a larger screen. For example, Drijvers and Özyürek (2017) used a 1650 x 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 participants. Questionnaire responses confirm that L1 participants paid little attention to the speaker’s lips in the experiment (*Mn* = 2.20).

In vowel error conditions, lack of informativeness may have been another factor influencing visual speech intelligibility scores. In these stimuli, visual speech simply *reinforced* the incorrect phonological information present in the auditory signal. 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 which made the words chosen for the visual speech conditions particularly difficult to understand. Especially given the large standard deviation in the visual speech + vowel error condition (*SD* = 19.7%), it may be necessary to include a larger number of stimuli per condition and/or a larger number of participants to achieve a sufficiently sensitive measurement of intelligibility.

6.1.3 The Effect of Vowel Errors

As predicted, vowel errors reduced the intelligibility of speech for L1 listeners. The effect size of this decrease was large (*r* = .60 in the audio-only modality; *r* = .57 in the visual speech modality). In the gesture modality, however, no such decrease was observed; this suggests that the detrimental effect of vowel error was fully mitigated by the addition of gesture. The finding that vowel accuracy affects intelligibility aligns with previous research (Bent et al., 2007; Zielinski, 2008). The fact that most of the vowel substitutions had a high functional load may have contributed to the size of the effect. Unfortunately, there were not enough low functional load contrasts in the current study to investigate the moderating influence of this factor. It should also 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 (Sewell, 2017).

6.2 Research Question 2: The Effect of Language Background

In Experiment 2, L2 listeners completed the same intelligibility task given to L1 listeners in Experiment 1. Analyses revealed areas of similarity between the two groups, as well as some areas of difference. Similar to L1 listeners, L2 listeners found speech accompanied by gesture to be relatively more intelligible and speech containing vowel errors to be relatively less intelligible. No significant effect of visual speech was found. Unlike L1 listeners, L2 listeners benefited from gesture even in standard pronunciation conditions and were harmed by vowel error in every modality. Baseline intelligibility scores were lower in the L2 group compared to the L1 group. In general, results partially confirmed predictions (see Table 5 for a summary of results).

6.2.1 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* = .83). Due to lower baseline scores, however, the gesture + vowel error condition did not reach 100% intelligibility as it had for L1 listeners. It is not possible to say whether the degree of gesture benefit differed meaningfully between L1 and L2 groups because of ceiling effects in Experiment 1. The two groups did clearly differ, however, in standard pronunciation conditions. Unlike L1 listeners, L2 listeners benefited from gesture even when there were no vowel errors (with a large effect size of *r* = .62). This aligns with previous research that has found a positive effect of gesture on L2 listening comprehension in normal listening conditions (Sueyoshi & Hardison, 2005; Dahl & Ludvigsen, 2014). The current study expands these findings to L2 listeners who are highly proficient in the language. 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 in day-to-day conversation in quiet environments, gesture has a greater potential to affect understanding for L2 listeners than L1 listeners.

6.2.2 Visual Speech: L2 vs. L1 Listeners

Similar to the L1 participants, L2 listeners did not have significantly different scores in the visual speech conditions compared to the audio-only conditions. Again, this could be because mouth movements lacked salience, because they lacked informativeness (in the vowel error conditions), or because the measurement lacked sensitivity. As with L1 listeners, the largest amount of variation occurred in the visual speech + vowel error condition. This aligns with previous research that has found large individual variation in participants’ abilities to benefit from visual speech (Grant et al., 1998).

6.2.3 Vowel Error: L2 vs. L1 Listeners

Vowel error reduced intelligibility for L2 listeners, with a large effect size (*r* = .61–.72). The percentage of intelligibility loss was very similar between L1 and L2 listeners, reaching 20% for both groups in audio-only conditions. A decrement of 20% 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). In gesture conditions, participants were less reliant on accurate pronunciation; decrement due to vowel error was 8% for L2 listeners and 0% for L1 listeners.

6.3 Attitudes about Visual Speech and Gestures

Questionnaire results revealed that Mandarin L1 users tended to have a more positive attitude toward visual speech/facial cues (as a comprehension aid) than English L1 users. It is possible that this reflects a cultural difference, but it more likely reflects a L1/L2 difference. Most of the questions in the “attitude toward visual speech” multi-item scale focused on listening to English. This means that participants in Experiment 1 were primarily thinking about the effect of visual speech in their first language, whereas participants in Experiment 2 were primarily thinking about their second language. When listening in an L1, speech is largely intelligible without visual cues (unless there is noise). However, speech in a second language is usually not completely intelligible from the auditory signal alone. It is therefore not surprising that second language listeners rated visual speech as more beneficial for comprehension. It is not clear if this belief had any effect on participants’ intelligibility scores.

Questionnaire results also showed that L1 and L2 participants tended to believe that gestures were more helpful for comprehension than visual speech—both in the experiment as well as in day-to-day life. This belief aligns with the results from the experiments, in which gesture increased intelligibility but visual speech did not. Participants 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 bore out in the experiment; gesture increased intelligibility when there were vowel errors (a type of error which 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 language background of the speaker. The distinction between L1 and L2 speakers made in this study’s version of the questionnaire is a novel addition. It may be useful to make this distinction in future uses of the questionnaire, depending on the aim of the research.

6.4 Limitations

Before turning to the implications of these findings, there are a few limitations that need to be discussed. First, results concerning L2 listeners cannot be generalized to all populations, but must be limited to Mandarin L1 users. The same effects on intelligibility may not have been observed in a different speaker-listener pairing (Stringer & Iverson, 2019). 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. All of these factors influence intelligibility, limiting the generalizability of results.

A second limitation regards a problem in design which came to light in the debriefing sessions at the end of the experiment. During these discussions, several participants in Experiment 1 and one participant in Experiment 2 explained that during the transcription task, they had been trying to figure out if there was any pattern to the pronunciation mistakes. A few of these participants (perhaps 3 in total) were able to correctly guess that it was vowel error by the end of the task. This knowledge may have helped these participants deduce correct answers, although their scores did not appear to deviate from the norm. Other participants who looked for patterns came to incorrect conclusions. For example, a couple of participants guessed that inaccurate pronunciations were based on errors commonly made by a particular population of L2 learner. It is possible that this false assumption harmed their performance on the task, although again no significant deviation was found. An easy solution to the problem of pattern-finding would have been to include distractor items that contained other types of pronunciation mistakes but would not be included in the final analysis.

A final limitation of the current study is its artificial nature. In order to control for particular variables and investigate causal relationships, the experiments 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. This type of controlled experiment has its place, but must be complemented by research which uses extemporaneous speech, unscripted gestures, and measures of global intelligibility.

6.5 Implications

The primary novel finding of this study is that iconic gesture can significantly affect the intelligibility of speech in two situations: (a) when speech contains segmental errors and/or (b) when the listener is an L2 user of the language. This finding has a number of implications for language assessment and teaching.

6.5.1 Speaking Assessment

Currently, gesture is not included in most speaking assessment criteria, certainly not in high stakes exams like the TOEFL or IETLS (see Hughes & Reed, 2016 for an overview of these criteria). If gesture has the power to affect the intelligibility of L2 speakers, as the current study suggests, should gesture be assessed in speaking exams? Bachman and Palmer’s (1996) model of “test usefulness” provides a helpful framework for this discussion. According to this model, the usefulness of a test can be evaluated on the basis of six qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality.

First, let us consider construct validity. Gesture should only be included in assessment if it is part of the construct of speaking. The more fundamental question is whether gesture is part of the construct of language. Many gesture theorists would argue that yes, gesture *is* part of language (see Kendon, 2000 for a discussion). McNeill writes:

> Language itself is inseparable from gesture. While gestures enhance the material carriers of meaning, *the core is gesture and speech together.* They are bound more tightly than saying gesture is an “add-on” or “ornament” implies. They are united as a matter of thought itself. (2016, p. 3)

Most theorists agree that gesture is part of day-to-day conversation, co-constructing meaning alongside speech. In general, when people speak, they also gesture and when they stop speaking, they stop gesturing (Graziano & Gullberg, 2018). It is only in exceptional circumstances that speaking happens without gesturing. Arguably then, adding gesture to speaking criteria would increase test validity.

It is likely that gesture is in fact already an important component of many speaking assessments, just one that is not written down in the test criteria. Jenkins and Parra (2003) investigated the role of nonverbal behaviour (including gesture) in an interview-style speaking exam designed to evaluate the proficiency of international teaching assistants at a US university. After closely analysing video recordings of 8 examinations and interviewing the evaluators, they concluded that nonverbal behaviour significantly affected ratings for “borderline” students; students who were highly linguistically proficient passed regardless of their nonverbal competence, but linguistically weaker students could pass or fail depending on the appropriateness of their nonverbal behaviour. As a result of the study, nonverbal and paralinguistic interaction was added to the institution’s scoring rubric under “communicative competence.” Several more recent studies have examined the role of nonverbal behaviour in the assessment of peer-to-peer candidate interactions. Through rater reports and stimulated verbal recalls, researchers have found that nonverbal behaviour, while not written down in the criteria, is heavily relied upon when rating the “interactional competence” of candidates in these exams (Ducasse & Brown, 2009; Ducasse, 2013; May, 2011). In light of these findings, Plough, Banerjee, and Iwashita (2018) argue that “the time has come for empirical investigations of NVB [nonverbal behaviour] in speaking tests with a view to incorporating it into a definition of IC [interactional competence] and thus the speaking construct” (p. 434). Gesture is one of many elements within nonverbal behaviour that they believe should be included within oral assessment.

Now let us turn to another quality in Bachman and Palmer’s test usefulness model—practicality. Adding gesture to assessment criteria might increase validity, but are there enough available resources to implement such a shift? Plough et al. (2018) admit that it would require a significant amount of time and research to develop a set of criteria which clearly outlined appropriate and inappropriate nonverbal behaviour, especially given cultural differences in some aspects of this behaviour. After criteria were established, raters would then need to be trained in how to interpret them reliably. Plough et al. think that such a development is difficult yet achievable, but others are less certain. May (2011) explains that including body language in assessment “would entail a consensus as to exactly what constitutes effective body language in a particular context; perhaps this Pandora’s box has remained closed for very good reasons” (p. 140). Regardless of the validity and authenticity of including gesture as part of assessment, it may simply be impractical.

A final quality to consider is “impact.” What effect would such a development have on teachers and students? If gesture is explicitly listed in test criteria, teachers may feel compelled to instruct students on appropriate gesture use. It is debatable whether this is possible or desirable (see section 6.5.4). If they are not given proper training, teachers may feel uncomfortable incorporating gesture into the syllabus. It is also unclear how students would react to such instruction.

In summary, it is not clear whether adding gesture to speaking assessment would increase test usefulness. While it may increase validity, other factors such as practicality and impact must also be considered. The decision to include or exclude gesture from speaking criteria should be made carefully, after weighing all of these factors. At the very least, it should be taken into account that gesture has the potential to affect the intelligibility (and perceived proficiency) of an L2 speaker. In face-to-face exams, excluding gesture from speaking criteria does not mean that gesture will not have an impact on assessment. If the issue is not clearly addressed in rater training sessions, gesture could become a hidden variable, affecting scores and the reliability of the test in an uncontrolled manner.

6.5.2 Listening Assessment

Results of the current study suggest that gesture has the potential to affect not only the intelligibility of L2 *speakers*, but also the intelligibility of interactions involving L2 *listeners*. In Experiment 2, L2 listeners benefited from gesture in both pronunciation conditions (standard pronunciation as well as vowel error). Given this effect, it is important to consider whether a visual modality should be included in L2 listening exams. In a recent review of 20 academic

English listening exams, none included a video of the speaker (Kang, Arvizu, Chaipuapae, & Lesnov, 2019). It is possible that visuals might enhance the usefulness of these exams.

Just as one can argue that gesture is a part of the construct of speaking, it can also be argued that gesture is a part of the construct of listening. Sueyoshi and Hardison (2005) make this case, as do a number of other researchers, including Wagner (2008, 2010), Ockey (2007), and Shin (1998). 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 (including gestures) to help them answer comprehension questions. He argued that this ability should be measured as part of the listening construct. In a subsequent study, Wagner (2010) found that participants who saw a video performed better on a listening exam than participants who listened to the audio only. As listeners typically have access to visual information in daily life, including this modality in listening assessment could increase test authenticity and validity.

On the other hand, it may not be practical to create audiovisual listening stimuli that pair easily with note-taking tasks or multiple choice questions. In Coniam's (2001) study, groups of listeners took either an audio-only version of a listening exam or an audiovisual version. Participants in the audiovisual group felt that they had received no benefit from the video and actually found it quite distracting to constantly shift their attention from the test paper to the screen. In this experiment, no difference in exam performance was found between groups. Similarly, Batty (2015) found almost no difference in test scores between listeners who took an audio-based test and video-based test. Batty presents a cogent argument against the inclusion of visuals in listening assessment, explaining that the type of semi-spontaneous speech and gesture used in Wagner's experiments is not well-suited to the creation of multiple choice questions with sufficiently difficult distractors. I am inclined to agree with Batty that high-stakes exam developers would find it difficult to use extemporaneous speech as test material. However, developers may be able to create their own audiovisual exam material in which both gesture and speech are highly scripted. Of course, this added factor would likely increase the time and money necessary to develop the test. Moreover, in order for listeners to actually engage with visual cues, traditional listening tasks may not be usable.

In the end, developers of listening exams and speaking exams face a similar dilemma. While the authenticity and validity of these exams is brought into question by theorists as well as empirical research (including the current study), the practicality of incorporating visual cues is a major concern. More research into nonverbal behaviour and language assessment is necessary in order to determine if and how such behaviour can be integrated into reliable exams.

6.5.3 Gesture as Input in the L2 Classroom

Results from Experiment 2 suggest that gestures facilitate understanding for L2 listeners. One clear implication of this finding is that language teachers who use gestures may be able to enhance students' comprehension and, by extension, their learning. Empirical research has shown that learners memorize new vocabulary items more easily and retain them longer when the words are presented alongside gestures (Kelly, McDevitt, & Esch, 2009; Lewis & Kirkhart, 2018). Given the possible benefits, some researchers argue that teachers should be trained in how to effectively use gesture in the classroom (Allen, 2000; Sime, 2006). Further research could help determine whether such training is effective.

6.5.4 Gesture as Output of L2 Learners

This study found that the use of iconic gestures significantly improved the intelligibility of speech containing segmental errors. It is possible then that L2 learners who make such errors could benefit from attending to their gesture use. I 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 are *not* culturally or linguistically specific—that is, they would not need to be explicitly taught in order to be used or understood in a second language. 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) provide a number of exercises for the L2 classroom that serve this purpose. In my view, encouraging students to notice their own gesturing behaviour and the behaviour of others is a reasonable, achievable goal. However, this encouragement may not be necessary in contexts where students are already fairly alert to this behaviour. Based on her review of research on gesture use in classrooms, Goldin-Meadow concludes that "gesture seems to be functioning very effectively in teacher-student exchanges at the moment without any intervention from us" (2003, p. 245). Further research that looks into the possible effects of gesture instruction could help determine if it is a worthwhile use of class time.

Beyond its immediate effect on intelligibility, gesturing in a second language has other benefits. It was mentioned above that *seeing* gestures can aid in vocabulary acquisition. Recent research has shown that *performing* gestures is even more helpful when learning new words (Allen, 1995; Tellier, 2008; Macedonia, Müller, & Friederici, 2011; Macedonia & Knöshe, 2011; Macedonia & Klimesch, 2014). The link between performing gestures and improved learning outcomes is supported by a number of cognitive theories (Macedonia & von Kriegstein, 2012).

Other benefits of gesture mentioned by SLA researchers include its role in creating affective bonds, maintaining conversations and eliciting appropriate input from an interlocutor.

Drawing on Vygotsky, McCafferty (2002) argues that gestures play an important role in creating zones of proximal development (ZPD) that are ideal for language learning. For example, a L2 speaker might gesture when unsure if they are using the correct word and in doing so, elicit the desired word from their interlocutor. In this way, gesturing not only increases intelligibility, but creates an opportunity for learning. Along similar lines, Gullberg (2008) relates gesture use to Krashen’s Comprehensible Input Hypothesis (1994), suggesting that gestures can help elicit optimal input. Beyond facilitating learning in a direct manner, gesturing can also contribute to a positive atmosphere between interlocutors by creating “a sense of shared physical, symbolic, psychological, and social space” (McCafferty, 2002, p. 201). This positive affect may then lead to further opportunities for language use. Observed effects on learning, affect, and intelligibility (including the results of the current study) all point toward a role for gesture in second language acquisition.

7 Conclusion

The main aim of this study was to incorporate a visual modality into L2 intelligibility research. Findings reveal a complex interaction between segmental accuracy, visual cues, and listener background in determining the intelligibility of speech. Providing support for previous research in segmentals, the current study found that vowel substitutions reduced intelligibility for both L1 and L2 listeners by approximately 20%. In these more difficult listening conditions (similar to speech in noise), iconic gesture had a large beneficial effect on intelligibility for all participants. For L2 listeners, gesture facilitated understanding even when speech contained no error. Overall, these findings indicate that gesture has the power to dramatically affect intelligibility. On the other hand, the current study found no significant effect of visual speech. This may be because the design of the study limited the salience of these visual cues or that the measurement of intelligibility was not sufficiently sensitive. Several recent studies using a larger number of stimuli and/or participants did find a visual speech enhancement for L2 speakers (Yi et al., 2013; Kawase et al., 2014), although the size of this effect was generally rather small. The findings of the current study suggest that gesture may have a comparatively larger effect. The difference in enhancement effect between visual speech and gesture may be partly due to the nature of the information they provide. While iconic gestures express semantic information that is not dependent on the segmental accuracy of speech, visual speech provides phonological information which is tightly bound to the acoustic signal itself. Unlike speech in noise, speech containing phonological errors is not necessarily disambiguated by seeing the way in which it is articulated.

L2 intelligibility researchers may be reluctant to investigate the effect of visual cues, regardless of their potential effect on understanding, on the grounds that they are construct-irrelevant. Nonverbal behaviour is not part of language, some might argue, and therefore is not important to study. There are two responses that can be made to such an argument. First, on a theoretical level, it is debatable whether “language” and “gesture” should or can be separated. Second, on a practical level, nonverbal behaviour is not something that can be easily ignored in face-to-face oral examinations. Just because gesture is not listed in the criteria does not mean gesture is not being assessed. On the contrary, the evidence available suggests that gesture does influence proficiency ratings. This effect may be operating in unknown ways, unbeknownst to even the evaluators themselves. Goldin-Meadow explains:

> Speakers are not always aware of the ideas they express in gestures. Listeners pick up on these ideas, but may themselves not be aware of having done so. An entire exchange can take place without either speaker or listener being consciously aware of information passed between them. (2003, p. 245)

This unconscious exchange deserves the attention of L2 intelligibility researchers.

Further research could help clarify and expand the results of the current study. For example, a similar experiment which added noise to some or all of the conditions would avoid ceiling levels and make it possible to compare enhancement effects between groups. It would also be interesting to test other L1 and L2 populations in order to investigate moderating variables, such as language background, age, and proficiency level. In addition, future studies could manipulate different aspects of the stimuli, including the word class, gesture type, type of linguistic error, complexity of speech (i.e. using sentences or mini-lectures instead of words), or the source of speech (i.e. using spontaneous instead of scripted speech). Finally, the intelligibility task itself is a possible moderating factor that needs to be investigated. To conclude, there is so little known about the interaction between L2 speech intelligibility and visual cues that avenues for future research point in almost every direction.

References

Allen, L. Q. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal*, *79*(4), 521–529. https://doi.org/10.1111/j.1540-4781.1995.tb05454.x

Allen, L. Q. (2000). Nonverbal accommodations in foreign language teacher talk. *Applied Language Learning*, *11*(1), 155–76.

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339–355. https://doi.org/10.1348/000712601162220

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

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*(4), 2015–2024. https://doi.org/10.1121/1.4916265

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*(1), 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/

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly*, *22*(4), 593–606. https://doi.org/10.2307/3587258

Burnham, D., & Lau, S. (1998). The effect of tonal information on auditory reliance in the McGurk effect. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*.

Catford, J. C. (1987). Phonetics and the Teaching of Pronunciation. In J. Morley (Ed.), *Current perspectives on pronunciation: Practices anchored in theory* (pp. 83–100). TESOL.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System*, *29*(1), 1–14. https://doi.org/10.1016/S0346-251X(00)00057-9

Dahl, T. I., & Ludvigsen, S. (2014). How I see what you're saying: The role of gestures in native and foreign language listening comprehension. *The Modern Language Journal*, *98*(3), 813–833. https://doi.org/10.1111/modl.12124

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1–16. https://doi.org/10.1017/s0272263197001010

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

Deterding, D. (2013). *Misunderstandings in English as a lingua franca: An analysis of ELF interactions in South-East Asia* (Vol. 1). Walter de Gruyter. https://doi.org/10.1515/9783110288599

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*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Drijvers, L., & Özyürek, A. (2019). Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language and Speech*. Advance online publication. https://doi.org/10.1177/0023830919831311

Ducasse, A. M. (2013). Such a nice gesture: Paired Spanish interaction in oral test discourse. *Journal of Language Teaching and Research*, *4*(6), 1167–1175. https://doi.org/10.4304/jltr.4.6.1167-1175

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423–443. https://doi.org/10.1177/0265532209104669

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. https://doi.org/10.1037/1082-989X.1.2.170

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423–425. https://doi.org/10.1044/jshr.1202.423

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*(3), 399–423. https://doi.org/10.2307/3588487

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Harvard University Press.

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*(5), 2677–2690. https://doi.org/10.1121/1.422788

Graziano, M., & Gullberg, M. (2018). When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in Psychology*, *9*, Article 879. https://doi.org/10.3389/fpsyg.2018.00879

Gregersen, T., & MacIntyre, P. D. (2017). *Optimizing language learners nonverbal behavior: From tenet to technique*. Channel View Publications. https://doi.org/10.21832/9781783097371

Gullberg, M. (1998). *Gesture as a communication strategy in second language discourse: A study of learners of French and Swedish* (Vol. 35). Lund University.

Gullberg, M. (2008). Gestures and second language acquisition. In P. Robinson & N. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 276–305). Routledge.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*(2), 201–223. https://doi.org/10.2307/3588378

Hardison, D. M. (1999). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. *Language Learning*, *49*, 213–283. https://doi.org/10.1111/0023-8333.49.s1.7

Hardison, D. M. (2005). Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment. *Applied Psycholinguistics*, *26*(4), 579–596. https://doi.org/10.1017/s0142716405050319

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*(3), 1740–1751. https://doi.org/10.1121/1.2166611

Holler, J., Shovelton, H., & Beattie, G. (2009). Do iconic hand gestures really contribute to the communication of semantic information in a face-to-face context? *Journal of Nonverbal Behavior*, *33*(2), 73–88. https://doi.org/10.1007/s10919-008-0063-9

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, *137*(2), 297–315. https://doi.org/10.1037/a0022128

Hughes, R., & Reed, B. S. (2016). *Teaching and researching speaking* (3rd ed.). Routledge. https://doi.org/10.4324/9781315692395

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics*, *23*(1), 83–103. https://doi.org/10.1093/applin/23.1.83

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. *The Modern Language Journal*, *87*(1), 90–107. https://doi.org/10.1111/1540-4781.00180

Jongman, A., Wang, Y., & Kim, B. H. (2003). Contributions of semantic and facial information to perception of nonsibilant fricatives. *Journal of Speech, Language, and Hearing Research, 46*(6), 1367–77. https://doi.org/10.1044/1092-4388(2003/106)

Kang, O., Thomson, R. I., & Moran, M. (2018a). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*(1), 115–146. https://doi.org/10.1111/lang.12270

Kang, O., Thomson, R. I., & Moran, M. (2018b). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics*. Advance online publication. https://doi.org/10.1093/applin/amy053

Kang, T., Arvizu, M. N. G., Chaipuapae, P., & Lesnov, R. O. (2019). Reviews of academic English listening tests for non-native speakers. *International Journal of Listening*, *33*(1), 1–38. https://doi.org/10.1080/10904018.2016.1185210

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*(3), 1352–1362. https://doi.org/10.1121/1.4892770

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

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, *24*(2), 313–334. https://doi.org/10.1080/01690960802365567

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207–227). Mouton. https://doi.org/10.1515/9783110813098.207

Kendon, A. (2000). Language and gesture: Unity or duality. In D. McNeill (Ed.), *Language and gesture*, *2* (pp. 47–63). Cambridge University Press. https://doi.org/10.1017/cbo9780511620850.004

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

Kerr, A. W., Hall, H. K., & Kozub, S. A. (2002). *Doing statistics with SPSS*. Sage.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145–167. https://doi.org/10.1080/01690960802586188

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. https://doi.org/10.1016/s0749-596x(02)00505-3

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS.* 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

Levis, J., & Im, J. (2015). Judgments of non-standard segmental sounds and international teaching assistants' spoken proficiency levels. In G. Gorsuch (Ed.), *Talking matters: Research on talk and communication of international teaching assistants* (pp. 113–142). New Forums Press.

Lewis, T. N., & Kirkhart, M. (2018). Effect of iconic gestures on second language vocabulary retention in a naturalistic setting. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2016-0125

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PLoS One*, *4*(3), Article e4638. https://doi.org/10.1371/journal.pone.0004638

Macedonia, M., & Klimesch, W. (2014). Long-term effects of gestures on memory for foreign language words trained in the classroom. *Mind, Brain, and Education*, *8*(2), 74–88. https://doi.org/10.1111/mbe.12047

Macedonia, M., & Knösche, T. R. (2011). Body in mind: How gestures empower foreign language learning. *Mind, Brain, and Education*, *5*(4), 196–211. https://doi.org/10.1111/j.1751-228x.2011.01129.x

Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping*, *32*(6), 982–998. https://doi.org/10.1002/hbm.21084

Macedonia, M., & von Kriegstein, K. (2012). Gestures enhance foreign language learning. *Biolinguistics*, *6*, 393–416.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131–141. https://doi.org/10.3109/03005368709077786

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*(9), 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*(2), 127–145. https://doi.org/10.1080/15434303.2011.565845

McCafferty, S. G. (2002). Gesture and creating zones of proximal development for second language learning. *The Modern Language Journal*, *86*(2), 192–203. https://doi.org/10.1111/1540-4781.00144

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

McNeill, D. (2005). *Gesture and thought*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226514642.001.0001

McNeill, D. (2016). *Why we gesture*. Cambridge University Press. https://doi.org/10.1017/cbo9781316480526

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746. https://doi.org/10.1038/264746a0

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (1995b.) Processing time, accent, and comprehensibility in the perception of foreign-accented speech. *Language and Speech*, *38*(3), 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*(4), 520–531. https://doi.org/10.1016/j.system.2006.09.004

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning*, *58*(3), 479–502. https://doi.org/10.1111/j.1467-9922.2008.00448.x

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. 377–396). Wiley Blackwell. 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

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*(1), 4–12. https://doi.org/10.1007/s00426-005-0031-5

Nielsen, K. (2004). Segmental differences in the visual contribution to speech intelligibility. *Journal of the Acoustical Society of America*, *115*(5), 2606. https://doi.org/10.1121/1.4784676

Neri, A., Cucchiarini, C., & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL*, *20*(2), 225–243. https://doi.org/10.1017/s0958344008000724

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*(4), 517–537. https://doi.org/10.1177/0265532207080771

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), Article 20130296. https://doi.org/10.1098/rstb.2013.0296

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Plough, I., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing*, *35*(3), 427–445. https://doi.org/10.1177/0265532218772325

Quené, H., & Van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication*, *52*(11-12), 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*(3), 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*(1), 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

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). Russell Sage Foundation.

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*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do L2 listeners perceive the comprehensibility of foreign-accented speech? Roles of first language profiles, second language proficiency, age, experience, familiarity and metacognition. *Studies in Second Language Acquisition*, *41*(5), 1133–1149. https://doi.org/10.1017/s0272263119000226

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception and Psychophysics*, *59*(1), 73–80. https://doi.org/10.3758/bf03206849

Sekiyama, K., & Tohkura, Y. I. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, *90*(4), 1797–1805. https://doi.org/10.1121/1.401660

Sewell, A. (2017). Functional load revisited. *Journal of Second Language Pronunciation*, *3*(1), 57–79. https://doi.org/10.1075/jslp.3.1.03sew

Sherman, J., & Nicoladis, E. (2004). Gestures by advanced Spanish-English second-language learners. *Gesture*, *4*(2), 143–156. https://doi.org/10.1075/gest.4.2.03she

Shin, D. (1998). Using videotaped lectures for testing academic listening proficiency. *International Journal of Listening*, *12*(1), 57–80. https://doi.org/10.1080/10904018.1998.10499019

Sime, D. (2006). What do learners make of teachers' gestures in the language classroom? *IRAL-International Review of Applied Linguistics in Language Teaching*, *44*(2), 211–230. https://doi.org/10.1515/iral.2006.009

Smith, L. E., & Nelson, C. L. (1985). International intelligibility of English: Directions and resources. *World Englishes*, *4*(3), 333–342. https://doi.org/10.1111/j.1467-971X.1985.tb00423.x

Sommers, M., Tye-Murray, N., & Spehar, B. (2005). Audio-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing*, *26*(3), 263–275. https://doi.org/10.1097/00003446-200506000-00003

Stam, G., & McCafferty, S. G. (2008). Gesture studies and second language acquisition: A review. In S. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research* (pp. 15–36). Routledge. https://doi.org/10.4324/9780203866993

Stam, G., & Buescher, K. (2018). Gesture research. In A. Phakiti, P. DeCosta, P. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 793–809). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-59900-1_36

Stringer, L., & Iverson, P. (2019). Accent intelligibility differences in noise across native and nonnative accents: Effects of talker–listener pairing at acoustic–phonetic and lexical levels. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2213–2226. https://doi.org/10.1044/2019_JSLHR-S-17-0414

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 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*(2), 212–215. https://doi.org/10.1121/1.1907309

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/1362168819858246

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

Team, A. (2016). *Audacity* (Version 2.1.2) [Computer software]. http://www.audacityteam.org

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, *8*(2), 219–235. https://doi.org/10.1075/gest.8.2.06tel

Thomson, R. I., & Derwing, T. M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326–344. https://doi.org/10.1093/applin/amu076

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification*, *11*(4), 233–241. https://doi.org/10.1177/1084713807307409

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, *5*(3), 218–243. https://doi.org/10.1080/15434300802213015

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*(4), 493–513. https://doi.org/10.1177/0265532209355668

Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics*, *37*(3), 344–356. https://doi.org/10.1016/j.wocn.2009.04.002

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

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

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*(5), 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*(1), 93–109. https://doi.org/10.1017/S0142716418000462

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*(1), 69–84. https://doi.org/doi:10.1016/j.system.2007.11.004

Appendices

Appendix A

Stimuli List (organised by condition)

Frequency from lextutor BNC-COCA-25; 1k = within top 1,000 word families; 2k = within top 2,000
GA = General American; phonetic transcriptions from *The Longman Pronunciation Dictionary* (Wells, 2008).

audio-only + standard pronunciation condition

word	frequency	GA pronunciation
tell	1k	/tel/
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	/bend/
bump	2k	/bʌmp/

audio-only + vowel error condition

word	frequency	vowel change
put	1k	/pʊt/ → /pu:t/
rest	1k	/rest/ → /ri:st/
teach	1k	/ti:tʃ/ → /tetʃ/
move	1k	/mu:v/ → /mʊv/
pass	1k	/pæs/ → /pes/
guess	1k	/ges/ → /gɪs/
let	1k	/let/ → /læt/
tap	2k	/tæp/ → /tep/
match	2k	/mætʃ/ → /metʃ/
cheat	2k	/tʃi:t/ → /tʃet/

visual speech + standard pronunciation 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	/dræg/
sink	2k	/sɪŋk/
risk	2k	/rɪsk/

visual speech + vowel error condition

word	frequency	vowel change
feed	1k	/fi:d/ → /fɪd/
laugh	1k	/læf/ → /lef/
fish	1k	/fɪʃ/ → /feʃ/
step	1k	/step/ → /stɪp/
send	1k	/send/ → /si:nd/
lose	1k	/lu:z/ → /lʊz/
check	1k	/tʃek/ → /tʃæk/
lend	2k	/lend/ → /lɪnd/
spread	2k	/spred/ → /spri:d/
spell	2k	/spel/ → /spæl/

iconic gesture + standard pronunciation 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

word	frequency	GA pronunciation	iconic gesture
shut	1k	/ʃʌt/	both hands touching at a 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 upwards) move upwards
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	/tʃɑ:p/	right hand (open palm facing left) strikes left hand (open palm facing up)

iconic gesture + vowel error condition

word	frequency	vowel change	iconic gesture
give	1k	/gɪv/ → /gev/	hand (open palm, facing up) extends forward
catch	1k	/kætʃ/ →/ketʃ/	right hand and left hand (curved) clasp together
push	1k	/pʊʃ/ → /pu:ʃ/	both hands (open palms, facing out) extend away from body
sleep	1k	/sli:p/ → /slep/	hands (palms together) press to side of face
smell	1k	/smel/ → /smɪl/	open hand makes “whiff” motion near nose
press	1k	/press/ → /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 is 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/ → /flep/	hand (open palm, facing up) turns over (open palm, facing down)
twist	2k	/twɪst/ → /twest/	both hands (clenched position) rotate in opposite directions

Appendix B

Intelligibility Task

Name: ____________________________________

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 actress is trying to communicate? If you are not sure, please try to guess.*

Trial

1. ________________________

2. ________________________

3. ________________________

4. ________________________

5. ________________________

6. ________________________

Experiment

1. ________________________

2. ________________________

3. ________________________

4. ________________________

5. ________________________

6. ________________________

7. ________________________

8. ________________________

9. ________________________

10. ________________________

11. ________________________

12. ________________________

13. ________________________

14. ________________________

15. ________________________

16. ________________________

17. ________________________

18. ________________________

19. ________________________

20. ________________________

21. ________________________

22. ________________________

23. ________________________

24. ________________________

25. ________________________

26. ________________________

27. ________________________

28. ________________________

29. ________________________

30. ________________________

31. ________________________

32. ________________________

33. ________________________

34. ________________________

35. ________________________

36. ________________________

37. ________________________

38. ________________________

39. ________________________

40. ________________________

41. ________________________

42. ________________________

43. ________________________

44. ________________________

45. ________________________

46. ________________________

47. ________________________

48. ________________________

49. ________________________

50. ________________________

51. ________________________

52. ________________________

53. ________________________

54. ________________________

55. ________________________

56. ________________________

57. ________________________

58. ________________________

59. ________________________

60. ________________________

APPENDIX C

Visual Cue Preference Questionnaire – 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
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 – L2 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.	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
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
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)

Appendix D

Biographical Information Interview – L1 Participants

[administered orally] Before you leave, I'd like to ask you some questions about your language background and experience. You can skip any question that you wish not to answer. Any information that you provide will be kept confidential.

1. Gender: □ Male □ Female □ Non-binary □ Prefer not to answer

2. Age: __________ □ Prefer not to answer

3. Nationality: ________________________________

4. First language(s) : ____
If Chinese, please specify (e.g. Mandarin, Cantonese, etc).

5. Do you speak any additional languages? Give your proficiency level (*e.g.* beginner, intermediate, advanced).

Language	**Proficiency level**
_________________	________________________
_________________	________________________
_________________	________________________

6. What are the first language(s) of your parent(s)/guardian(s)?
Parent/Guardian 1: ___________________________
Parent/Guardian 2: ___________________________

7. What language(s) do you speak **at home** (generally)? Give a percentage of time (*e.g.* English 70%, Spanish 30%)

Language	**percentage of time this language is spoken at *home***
English	_______%
_________________	_______%
_________________	_______%

8. What language(s) do you speak **at school** (generally)? Give a percentage of time (*e.g.* English 70%, Spanish 30%)

Language	**percentage of time this language is spoken at *school***
English	_______%
_________________	_______%
_________________	_______%

9. What language(s) do you speak **at work** (generally)? Give a percentage of time (*e.g.* English 70%, Spanish 30%)

Language	**percentage of time this language is spoken at *work***
English	_______%
_________________	_______%
_________________	_______%

□ I don't have a job

10. Have you ever had any hearing problems?

❑ Yes ❑ No

11. Do you have normal vision or corrected to normal vision?

❑ Yes ❑ No

12. Do you have any experience in teaching English?

❑ Yes ❑ No

If yes, how long have you taught English? ________________________ (years)

13. Have you ever taken any linguistics classes? If yes, what kinds of classes?

❑ No ❑ Yes → explain: ____

____

THANK YOU FOR YOUR PARTICIPATION!

Biographical Information Interview – L2 Participants

[administered orally] Before you leave, I'd like to ask you some questions about your language background and experience. You can skip any question that you wish not to answer. Any information that you provide will be kept confidential.

1. Gender: □ Male □ Female □ Non-binary □ Prefer not to answer

2. Age: __________ □ Prefer not to answer

3. Nationality: ________________________________

4. First language(s) : ____
If Chinese, please specify (e.g. Mandarin, Cantonese, etc).

5. Do you speak any additional languages? Give your proficiency level (*e.g.* beginner, intermediate, advanced).

Language	Proficiency level
_________________	________________________
_________________	________________________
_________________	________________________

6. Have you taken the IELTS, TOEFL, or other English proficiency exam in the last 3 years? If so, what was you score?

□ IELTS → score: _________ □ TOEFL → score: _________ □ other → explain: __________

7. What are the first language(s) of your parent(s)/guardian(s)?
Parent/Guardian 1: ___________________________
Parent/Guardian 2: ___________________________

8. What language(s) do you speak **at home** (generally)? Give a percentage of time (*e.g.* English 70%, Spanish 30%)

Language	percentage of time this language is spoken at ***home***
English	_______%
_________________	_______%
_________________	_______%

9. What language(s) do you speak **at school** (generally)? Give a percentage of time (*e.g.* English 70%, Spanish 30%)

Language	percentage of time this language is spoken at ***school***
English	_______%
_________________	_______%
_________________	_______%

10. What language(s) do you speak **at work** (generally)? Give a percentage of time (*e.g.* English 70%, Spanish 30%)

Language **percentage of time this language is spoken at *work***

English _______%

_________________ _______%

_________________ _______%

□ I don't have a job

11. Have you ever had any hearing problems?

□ Yes □ No

12. Do you have normal vision or corrected to normal vision?

□ Yes □ No

13. Do you have any experience in teaching English?

□ Yes □ No

If yes, how long have you taught English? ________________________ (years)

14. Have you ever taken any linguistics classes? If yes, what kinds of classes?

□ No □ Yes → explain: ____

____

15. How long have you been living in the UK? ________________________ (months)

16. Have you ever lived in another English-speaking country?

□ No □ Yes → how long: ____________________________ (months)

THANK YOU FOR YOUR PARTICIPATION!

APPENDIX E

Information Sheet Provided to Participants

My name is Page Wheeler and I am inviting you to take part in my research project on the effect of facial cues and gesture on intelligibility. I am a MA TESOL (Teaching English to Speakers of Other Languages) student at UCL's Institute of Education. I will be conducting this research for my dissertation project, supervised by Dr. Kazuya Saito. In this project, I am hoping to investigate factors which might affect the intelligibility of speech, including the presence/absence of facial cues, gestures, and pronunciation errors. I very much hope that you would like to take part. This information sheet will try to answer any questions you might have about the project, but please don't hesitate to contact me if there is anything else you would like to know.

Who is carrying out the research?
This research is being conducted by Page Wheeler at the Institute of Education, University College London.

Why are we doing this research?
The main aim of this study is to investigate factors that influence the intelligibility of speech. In other words, I am interested in the factors that affect a listener's ability to accurately understand the words that are spoken. The study examines a variety of factors, including pronunciation, the availability of visual (facial) information, the use of gesture, and the language background of the listener. A secondary aim of this study is to explore the attitudes and preferences of L1 and L2 users regarding the use of visual information and gesture.

Why am I being invited to take part?
You are being invited to take part as either a first or second language speaker of English in order to aid us in identifying factors that influence intelligibility as well as potential differences between L1 and L2 users.

What will happen if I choose to take part?
If you choose to take part, you will be asked to perform a transcription task and fill in a questionnaire at the Institute of Education. In the transcription task, you will watch a series of short video clips. In each video, an actress says an action verb, sometimes accompanied by a gesture, and you will have to write the verbs you hear. Some words may be mispronounced. In some video clips, the speakers' lips will be covered. In the questionnaire, you will answer a few questions about your attitudes and preferences regarding visual information and gestures. You will also be asked to provide some biographical information about your language background and experience. The whole session should take approximately 45 minutes.

Will anyone know I have been involved?
No one will know that you have been involved. All data will be coded to ensure that your identity remains anonymous, and this data will only be accessible to myself. The data reported in the dissertation will not include any information that would suggest your identity.

Could there be problems for me if I take part?
There are no anticipated problems or risks for you if you take part in this research.

What will happen to the results of the research?
The results will be reported in my dissertation project. Your identity will be kept anonymous throughout the process and no individuals will be identifiable in the reported results. If you would like me to share the results of my research with you, I am happy to do so. All data will be stored securely on a hard drive that only I have access to. After I have completion of the research, all data will be permanently deleted from the hard drive.

Do I have to take part?
It is entirely up to you whether or not you choose to take part. I hope that if you do choose to be involved then you will find it a valuable experience. If you choose not to take part, there will be no repercussions. You may also withdraw your participation at any point.

Data Protection Privacy Notice
The data controller for this project will be University College London (UCL). The UCL Data Protection Office provides oversight of UCL activities involving the processing of personal data, and can be contacted at data-protection@ucl.ac.uk. UCL's Data Protection Officer can also be contacted at data-protection@ucl.ac.uk. Further information on how UCL uses participant information can be found here:
https://www.ucl.ac.uk/legal-services/privacy/ucl-general-research-participant-privacy-notice

Contact for further information
If you have any further questions before you decide whether to take part, you can reach me at page.wheeler.18@ucl.ac.uk.

Thank you very much for taking the time to read this information sheet.

APPENDIX F

Consent Form

If you are happy to participate in this study, please complete this consent form.

	Yes	**No**
I have been informed about the nature of this study and willingly consent to take part in it.	□	□
I understand that my identity will be kept confidential.	□	□
I understand that I can withdraw from the study at any time.	□	□
I understand that I can contact Page Wheeler at any time and request for my data to be removed from the project data base.	□	□

Name ________________________________ Signed ____________________________________

Date _________________________________

Page Wheeler
UCL Institute of Educational
20 Bedford Way, London WC1H 0AL
page.wheeler.18@ucl.ac.uk