Visual cues and rater perceptions of second language comprehensibility, accentedness, and fluency

Aki Tsunemoto, Rachael Lindberg, Pavel Trofimovich, and Kim McDonough

Tsunemoto, A., Lindberg, R., Trofimovich, P., & McDonough, K. (2021). Visual cues and rater perceptions of second language comprehensibility, accentedness, and fluency. *Studies in Second Language Acquisition*. Published online 5 August 2021.
https://doi.org/10.1017/S0272263121000425

Visual Cues and Rater Perceptions of Second Language Comprehensibility, Accentedness, and Fluency

This study examined the role of visual cues (facial expressions and hand gestures) in second language (L2) speech assessment. University students (N = 60) at English-medium universities assessed 2-minute video clips of 20 L2 English speakers (10 Chinese and 10 Spanish speakers) narrating a personal story. They rated the speakers' comprehensibility, accentedness, and fluency using 1,000-point sliding scales. To manipulate access to visual cues, the raters were assigned to three conditions that presented audio along with (a) the speaker's static image, (b) a static image of a speaker's torso with dynamic face, or (c) dynamic torso and face. Results showed that raters with access to the full video tended to perceive the speaker as more comprehensible and significantly less accented compared to those who had access to less visually informative conditions. The findings are discussed in terms of how the integration of visual cues may impact L2 speech assessment.

Keywords accentedness, comprehensibility, fluency, speech assessments, visual cues

Visual Cues and Rater Perceptions of Second Language Comprehensibility, Accentedness, and Fluency

Introduction

Nonverbal behaviors accompanying second language (L2) speech, including a speaker's facial cues (e.g., eyebrow raises, blinks), head movements, and hand gestures, have been shown to enhance speech perception (e.g., Li et al., 2020; Zheng & Samuel, 2019) and improve listening comprehension (e.g., Batty, 2014; Sueyoshi & Hardison, 2005) for the listener. Considering their sensitivity to visual information, listeners with access to a speaker's visual cues may also evaluate the speaker more favorably. However, prior research has primarily relied on audio recordings of speaker performance, and no work to date has focused on how visual cues can contribute to such global listener-based measures of L2 speech as comprehensibility, accentedness, and fluency, which is the goal of this study.

When listeners evaluate comprehensibility (i.e., how easily listeners understand a speaker), they primarily rely on various linguistic dimensions in L2 speech, including phonology, lexis, grammar, fluency, and discourse (e.g., Isaacs & Trofimovich, 2012). Listener perceptions of accentedness, which captures how closely a speaker approximates the expected language variety (Munro & Derwing, 1995), are narrower in scope and are mostly determined by a speaker's segmental and suprasegmental accuracy (Hayes-Harb & Hacking, 2015; Saito et al., 2017). In turn, a listener-rated measure of fluency, which typically captures various aspects of utterance flow, can be largely explained through temporal characteristics of speech, including pausing and articulation speed (Bosker et al., 2013; Kahng, 2018). In research settings, these global dimensions of speech have typically been operationalized in terms of listeners' intuitive

judgements, through Likert-type scales (e.g., Munro & Derwing, 1995) or continuous sliding scales (e.g., Saito et al., 2017).

Although listener-based global dimensions of L2 speech, such as comprehensibility, accentedness, and fluency, are popular measures of L2 performance in research and assessment settings (Saito & Plonsky, 2019), most previous research has focused on these dimensions through evaluation of audio recordings. However, there are certainly situations where L2 speakers are judged when a rater (listener or interlocutor) has access not only to aural but also visual information. For instance, a speaker's appearance (e.g., in images) or various dynamic visual cues (e.g., in video clips or face-to-face interaction) can impact how accented or comprehensible the speaker sounds to a listener (Kang & Rubin, 2009; Kutlu, 2020). Moreover, in high-stakes assessments, such as IELTS, test-takers' performance is often evaluated when the rater not only observes L2 speakers but also interacts with them in real time (Nakatsuhara et al., 2021). In fact, outside laboratory settings or language tests, L2 speakers most often interact with their interlocutors in person as part of daily communication. However, to date, very little is known about how listeners' evaluations of L2 speech vary based on their access to a speaker's visual cues.

Because visual and verbal information are intertwined in human communication, they make a joint contribution to listeners' perception, interpretation, and comprehension of speech (Gullberg, 2006; Kelly et al., 2008). Neurocognitive evidence suggests that people integrate verbal and visual information while processing speech (Bates & Dick, 2002) and that visual information enhances speech processing for listeners (Beattie & Shovelton, 1999). Seeing a speaker's facial expressions (e.g., sadness, happiness, confusion), for example, may help a listener anticipate what kind of information will be shared (Wagner, 2008). Similarly, a speaker

may gesture toward an object to clarify a referent, which can help listeners understand the speaker's intended message (Kendon, 1994).

Paying attention to a speaker's facial expressions (especially articulatory configurations involving lip and jaw movement) can help listeners process speech, such that listeners perceive segmental and suprasegmental information more accurately when they can lipread (Swerts & Krahmer, 2008) or when a speaker's head movement is aligned with prominence markers, such as when a head nod co-occurs with a stressed word (Pelachaud et al., 1996). For example, when speech was masked in noise to make perception more challenging, listeners identified verbs significantly better when they saw a speaker's lip movements and hand gestures, compared to when only auditory information was available (Drijvers & Özyürek, 2017, 2020). Listeners have been shown to identify words more accurately when they saw a speaker closely, as opposed to from a distance (which made lip reading difficult), especially if the speaker had a stronger foreign accent (Zheng & Samuel, 2019).

In addition to facial cues, hand gestures may impact how listeners perceive L2 speech (Gullberg & McCafferty, 2008). For instance, when raters observed a speaker using gestures in face-to-face assessment, they more positively evaluated the speaker in terms of fluency, lexis, grammar, pronunciation, accuracy, range, and effectiveness compared to evaluating the speaker from audio recordings (Nakatsuhara et al., 2021; Nambiar & Goon, 2016; Neu, 1990). Similarly, when listeners transcribed monosyllabic action verbs (with and without vowel errors) while watching video clips, their transcription accuracy was highest when they observed a speaker use iconic gestures illustrating the actions (Wheeler, 2019). Besides the presence or absence of gestures, their frequency of occurrence may be important, such that L2 speakers were evaluated higher in oral proficiency when they used more hand gestures, compared to peers who used

fewer gestures (Gullberg, 1998; Jenkins & Parra, 2003; McCafferty, 2002).

Although facial expression and gestures can aid listeners in various ways, some cues or their combinations can be more beneficial than others. Focusing on visual cues in L2 listening comprehension, Sueyoshi and Hardison (2005) presented L2 English learners with an academic lecture in one of three conditions: audio only, audio plus face, and full video. Those who had access to the full video showed greater comprehension than those exposed to audio only, but there was no difference between the face only and the full video conditions, implying that various facial cues rather than gestures benefitted listener comprehension, although gestures may have been useful for lower-proficiency learners. In a detailed analysis of video-recorded university lectures, Hardison (2018) found that instructors used various combinations of cues (e.g., moving head, raising eyebrows, blinking) when emphasizing words that they felt important to highlight, in which beat gestures (hand or finger movements co-occurring with a rhythmic pulse) were particularly salient for listeners as carrying significance.

Even if a speaker's facial expressions and hand gestures are useful for listeners or raters evaluating L2 performance, the incidence of these cues might depend on individual differences in speaker proficiency (Gregersen et al., 2009) or cultural background (Gullberg, 2006; Kita, 2009), which might moderate the extent to which visual cues are useful for speech assessment. For instance, when engaged in dyadic conversations, advanced L2 speakers used gestures more frequently than lower-proficiency speakers (Gregersen et al., 2009). Similarly, in comparisons of gesture use across speakers from different linguistic and cultural backgrounds, English–Spanish bilinguals produced more gestures than English monolinguals (Pika et al., 2006), native English speakers gestured more frequently than Chinese speakers (So, 2010), and gesture use was more prevalent for L2 English speakers from Romance backgrounds than speakers from Asian

backgrounds (Nicoladis et al., 2018). Thus, to determine if visual cues are relevant to listener-based evaluations of L2 speakers' performance, it would be important to account for speakers' language proficiency and their language background.

Finally, although prior research has generally found positive effects for visual cues, visual information can have few or even negative effects on listeners. For example, when listeners identified L2 phonemes, they were generally more accurate in the audiovisual than the audio condition (Kawase et al., 2014). However, this benefit decreased when listeners had access to non-target articulatory configurations, such as English /ɹ/ produced without lip rounding. Inceoglu (2019) found no difference between the audio and audiovisual conditions for the accuracy of L2 learners' vowel identification, which implied that having access to a speaker's visual cues did not aid learners in a phoneme-focused task. At a global level, Ockey (2007) examined test-takers' engagement with visual materials in a listening comprehension test. Whereas more proficient test-takers found access to videos useful, less proficient test-takers reported videos to be distracting, suggesting that processing visual and verbal information simultaneously might increase listeners' processing load (Mayer & Moreno, 2003; Wagner, 2008). In light of these conflicting findings, it remains unclear whether listeners' evaluations of such global dimensions of L2 speech as comprehensibility, accentedness, and fluency differ as a function of their access to a speaker's visual cues.

The Current Study

Although prior research has revealed potential links between a speaker's visual cues and listener perception, these findings remain tentative because the impact of visual cues has been examined for specific segmental and suprasegmental targets (Inceoglu, 2019; Li et al., 2020; Scarborough et al., 2009) rather than listener-assessed measures of L2 speaking, including

comprehensibility, accentedness, and fluency. Furthermore, the role of visual cues has often been examined through rehearsed output, such as speakers reading words aloud (Li et al., 2020; Wheeler, 2019), and in monologic performances (Sueyoshi & Hardison, 2005), as opposed to spontaneous conversations where an interlocutor's presence may make visual cues more prevalent (Alibali et al., 2000; Bavelas et al., 2002). When extemporaneous speech was targeted (e.g., in university lectures), such research primarily focused on listeners' comprehension (Sueyoshi & Hardison, 2005) instead of their global assessments of speakers' performance. Gesture use also appears to vary across culture (Gullberg, 2006; Iverson et al., 2008; Kita, 2009) and across speaker proficiency (Gregersen et al., 2009). Therefore, it is important to explore the role of visual cues in the evaluation of L2 speech from speakers while controlling for their language proficiency and language background.

To shed light on how rater assessments differ based on access to a speaker's visual information, we elicited evaluations of comprehensibility, accentedness, and fluency while manipulating the type of visual cues (facial expressions, hand gestures) available to raters. Because raters may evaluate L2 speech more positively (Nakatsuhara et al., 2021; Nambiar & Goon, 2016; Neu, 1990), show greater comprehension (Sueyoshi & Hardison, 2005), and process speech more easily when given access to facial expressions (Zheng & Samuel, 2019) or hand gestures (Wheeler, 2019; Drijvers & Özyürek, 2017, 2020), videos were manipulated to provide varying access to visual cues while keeping audio consistent. To explore potential cultural differences in visual cue use (Gullberg et al., 2008; Nicoladis et al., 2018), we evaluated speaking performances by L2 speakers from Chinese- and Spanish-language backgrounds, as speakers from these backgrounds tend to vary greatly in their gesture use (Nicoladis et al., 2018). To account for possible variation in the use of visual cues across individual speakers varying in

their L2 skills, we used several measures of L2 speakers' proficiency and use as control covariates. This study was guided by the following research questions:

1. Do raters evaluate the comprehensibility, accentedness, and fluency of L2 English speakers from Chinese and Spanish backgrounds differently based on access to visual cues?
2. Which visual cues are associated with rater assessments of the speakers' comprehensibility, accentedness, and fluency?

Based on prior work, we hypothesized that access to facial expressions would influence ratings positively because raters can rely on lipreading to identify segment- and prosody-specific information in speech. We also reasoned that having access to a speaker's hand gestures may have an added benefit because raters may receive complementary information through gesture. We anticipated that these potential benefits would vary for the two speaker groups, as L2 speakers from Spanish-language backgrounds may use visual cues more frequently than those from Chinese backgrounds (Nicoladis et al., 2018). Based on the lack of systematic investigation of various visual cues in relation to global dimensions of L2 speech, we had no expectation regarding which specific visual cues might be associated with listener-based ratings of comprehensibility, accentedness, and fluency.

Method

Speech Samples

L2 speech samples were drawn from the Corpus of English as a Lingua Franca Interaction (CELFI), in which L2 English speakers from Canadian English-medium universities in Montreal carried out three, 10-minute communicative tasks in pairs (McDonough & Trofimovich, 2019). For this study, the close-call narrative task in which L2 speakers described a

personal story about a narrow escape from trouble or danger was used (see Appendix A). The task was particularly suitable because L2 speakers interacted with their interlocutors for 10 minutes to exchange stories, and thus had the opportunity to engage in natural communicative behaviors (Alibali et al., 2000).

An equal number of first language (L1) Spanish and Chinese (both Mandarin and Cantonese) speakers were sampled from the corpus based on the following inclusion criteria: (a) minimal hand movement crossing the face (so that facial expressions would not be obscured and the conditions with and without access to facial expressions and gestures could be kept separate), (b) speech sample of 1.5–2.0 minutes in length, (c) minimal interruption by partner (see Table 1 for a summary of background information). The 20 speakers were all Canadian university students (M_{age} = 23.90 years, SD = 3.83) enrolled in either undergraduate (15) or graduate (5) degree programs from different disciplines (e.g., business, psychology, marketing, computer science, studio art). As degree-seeking students, they had met the universities' minimum English language requirement for admission, which was a TOEFL iBT score of 75 (or equivalent). Between the two L1 groups, there were no differences for any of the background variables, $t(18) < -1.53$, $p > .144$, $d < 0.69$, except self-rated English proficiency in speaking, $t(18) = -2.29$, $p = .035$, $d = 1.02$, and in listening, $t(18) = -2.43$, $p = .026$, $d = 1.09$, with Chinese speakers evaluating their L2 oral proficiency lower than Spanish speakers.

Table 1. *Speaker Background Characteristics by Language Group*

	Chinese		Spanish	
Background variables	*M*	*SD*	*M*	*SD*
Age (years)	23.30	3.09	24.40	4.55
Length of residence in Canada (years)	5.90	8.33	3.71	5.59
Age English instruction began (years)	13.20	4.02	11.40	6.11
Self-rated English speaking (1–9 scale)	6.40	1.08	7.60	1.27
Self-rated English listening (1–9 scale)	7.20	0.79	8.20	1.03
Use of English at home (0–100%)	25.00	37.12	53.50	34.81
Use of English at school (0–100%)	88.00	13.78	82.50	31.38
Use of English at work (0–100%)	57.00	43.98	84.44	32.45

To ensure that the speech samples did not differ dramatically in the information provided to raters, they were checked for length and compared through lexical profiling (Cobb, 2019). As summarized in Table 2, although the mean length of the narratives was highly comparable, Spanish speakers produced longer stories than Chinese speakers (by an average of 62 tokens); however, no between-group differences reached statistical significance.

Table 2. *Narrative Characteristics by Speaker Group*

	Chinese		Spanish		Comparison		
Variables	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
Speech duration (s)	109.30	9.62	110.00	10.68	0.15	.879	0.07
Tokens	286.60	81.04	348.50	76.85	1.75	.097	0.78
Types	114.90	21.27	128.10	25.74	1.25	.227	0.56
Lexical density	0.46	0.04	0.43	0.04	–1.73	.102	0.75
K1–3 types (%)	93.82	2.49	92.83	2.29	–0.93	.367	0.41

Rating Stimuli

After selecting the speech samples, short clips from the videos (*M* = 110 seconds, *SD* = 9.90) were extracted to show only the speaker's upper body (face, arms, and torso) while communicating with their interlocutor. The video clips were then manipulated to present audio paired with three different visual conditions (Figure 1) using photo and video editing software (VideoPad, PhotoPad): (a) audio only (a static image of the speaker's face and torso), (b) audio with expressions (dynamic face with a static torso image), and (c) full video (full video containing dynamic face and torso).

Figure 1. Screenshots of the three visual conditions.

Raters

The raters were 60 university students (M_{age} = 23.57 years, *SD* = 5.41) enrolled in undergraduate (47) or graduate (13) programs at the same English-medium Canadian universities as the L2 speakers. The raters were recruited from the same speech community as the speakers on the assumption that they would represent their potential interlocutors (e.g., as classmates). They came from diverse language backgrounds, with 62% reporting English or French as their L1s. They all reported having normal hearing, and eight raters had previously taken a phonetics or phonology course. Twenty-five raters (42%) had English teaching experience of varying lengths (M_{years} = 1.32, *SD* = 1.93). Using a percentage scale (0 = *not at all*, 100 = *very familiar*), they self-reported being moderately familiar with Chinese-accented English (*M* = 51.05%, *SD* = 37.34) and Spanish-accented English (*M* = 58.50%, *SD* = 33.50). Raters were randomly assigned to one of the three visual conditions (20 per condition). As shown in Table 3, there were no statistically significant differences for any of the rater background variables. This implied that the raters were comparable in their characteristics, which minimized the possibility that differences in ratings across the conditions could be attributable to rater profiles.

Table 3. *Raters' Background Characteristics by Condition (20 Raters per Condition)*

	Audio only		Audio with expressions		Full video		Comparison		
Background variable	*n*	%	*n*	%	*n*	%	χ^2	*p*	*V*
Teaching experience	10	50	6	30	9	45	1.78	.410	.17
Linguistics coursework	7	35	6	30	5	25	0.48	.788	.11
	M	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*	*d*
Age (years)	23.00	5.08	22.40	4.30	25.30	6.46	1.64	.204	.16
Chinese accent familiarity (%)	65.25	33.66	37.40	35.15	50.50	39.40	2.97	.059	.59
Spanish accent familiarity (%)	63.20	33.20	53.40	33.72	58.90	34.57	0.42	.658	.21
Daily English listening (%)	67.80	23.51	81.10	15.60	82.60	17.02	2.22	.118	.57
Daily English speaking (%)	72.10	29.86	83.20	18.14	83.85	23.97	2.41	.099	.60

Procedure

The raters participated in small-group sessions (1–3 people) with a researcher in a laboratory setting (1.5 hours), using a laptop computer to access an online interface through LimeSurvey (https://www.limesurvey.org). The interface contained embedded videos presented with three 1,000-point sliding scales (Saito et al., 2017) for each speech dimension (comprehensibility, accentedness, fluency), all available below each video (see Appendix B for a screenshot of the interface). To maintain comparability of findings across studies, comprehensibility, accentedness, and fluency were defined to the raters through established definitions, as used in previous work targeting these dimensions in audio recordings only (Derwing & Munro, 2015), although the raters in this study had access to both aural and visual information across the three rating conditions. Comprehensibility was introduced as the degree

of effort required by the listener to understand the speaker. Accentedness was described as the extent to which the speech differed from a production pattern expected of a native English speaker, with "heavy accent" describing the speech that departs heavily from a native speaker's production and "no accent" referring to nativelike production. Fluency concerned the pace of speech, with fluent performance characterized by few pauses and hesitations and an optimal speaking rate (not too slow and not too fast). The scales contained no numerical markings to capture the raters' impressionistic judgments, but the endpoints were clearly labelled with a negative anchor point (on the left) and a positive anchor point (on the right). For comprehensibility, the endpoints were *hard to understand* and *easy to understand*; for accentedness, they were *heavily accented* and *no accent at all*; for fluency, they were *not fluent at all* and *very fluent*. The initial slider position was always in the middle.

The raters first received a paper-based training manual which introduced the three target constructs along with several other dimensions which are not discussed here (e.g., emotionality, story richness, interest in story). After resolving any questions, the raters independently assessed two practice videos. After rating the practice videos and confirming that they understood the rating task, they proceeded to assess the 20 target videos. All videos were automatically played only once, and the raters assigned ratings after the entire video was played. The videos were presented to each rater in a unique random order. After completing the session, the raters answered debrief questions about their ratings and filled out background questionnaires.

Data Analysis

All speech ratings were first checked for internal consistency using two-way, consistency, average-measure intraclass correlations (ICCs), separately by speaker group (L1 Chinese, L1 Spanish). As summarized in Table 4, ICC values were very high.

Table 4. *Interrater Reliability for Speech Ratings Across 60 Raters by Speaker Group*

	Chinese		Spanish	
Rated variable	ICC	95% CI	ICC	95% CI
Comprehensibility	.97	[.93, .99]	.97	[.93, .99]
Accentedness	.98	[.97, .99]	.99	[.98, .99]
Fluency	.97	[.93, .99]	.98	[.95, .99]

To examine L2 speakers' use of visual cues, the videos were analyzed for facial expressions and hand gestures. The coding followed a bottom-up, data-driven approach because there were no a priori expectations for which visual cues may be related to global L2 speech ratings. Initially, the first and second researchers independently watched four sample videos to identify facial expressions and gestures that occurred in each video. Based on their occurrence in the dataset, facial expressions and head movements were classified, through discussion, into six major categories, as summarized in Table 5. To identify hand gestures, the researchers watched each video without volume to avoid influence from speech (Gullberg, 2010). Hand gestures were defined in reference to a gesture phrase that happened between major resting positions around the stroke, which refers to the most effortful movement expressing a meaning (Stam & Buescher, 2018). Because of low incidence of iconic gestures (illustrating the shape of an object or the motion of an action), metaphoric gestures (representing a concrete image to communicate an abstract idea), and deictic gestures (pointing a finger to indicate an object), they were merged into one category (referential gestures), but beat gestures were kept as a stand-alone category. After establishing the categories for coding facial expressions and hand gestures, the second researcher independently watched all videos and recorded raw frequency counts per category. The first researcher then reviewed all coding decisions, and any disagreement was resolved

through discussion. Finally, a trained research assistant independently coded all videos to assess reliability. Two-way mixed, agreement, average-measure ICCs all exceeded .80, so the second researcher's frequency counts were used in all further analyses. Because the videos for L1 Chinese and L1 Spanish speakers did not differ in length (see Table 2) and because all raters experienced the same materials with or without access to visual cues (depending on the condition), analyses of visual cues were based on raw (non-normalized) frequency counts.

Table 5. *Interrater Reliability for Coding of Visual Cues*

Category	Included features	ICC	95% CI
Head movement	Tilts, shakes, nods	.81	[.53, .93]
Eyebrow movement	Both eyebrows raised, frowns	.95	[.87, .98]
Looking away	Glancing away, looking up, looking aside, looking down	.84	[.60, .94]
Blinking	Blinks	.95	[.86, .98]
Smiling and laughing	Smiles, laughs	.90	[.74, .96]
Lip movement	Pursed lips (e.g., pursing and curling of lips, tongue touching lips)	.95	[.88, .98]
Referential gestures	Iconic, metaphoric, and deictic gestures	.80	[.51, .92]
Temporal highlighting	Beat gestures	.82	[.48, .93]

Statistical Modeling

To address the first research question, which asked whether access to visual cues contributed to the variance in L2 speech ratings, we computed linear mixed-effects models in *R* (version 4.0.2, R Core Team, 2020) using the lme4 package (version 1.1-23, Bates et al., 2015). In each set of models, comprehensibility, accentedness, and fluency served as the outcome variables while condition (audio only, audio with expressions, full video), speakers' L1 (Chinese,

Spanish), and their interaction served as fixed-effects predictors, and raters (60) and speakers (20) were entered as random-effects predictors, which yielded a total of 1,200 observations per model. In addition, each model included three fixed effects as control covariates to account for potential between-speaker differences in L2 proficiency and use: (a) length of residence in Canada, (b) self-assessed L2 speaking and listening proficiency[1] (mean across separate self-ratings for speaking and listening on a 9-point scale), (c) self-reported amount of daily L2 use (mean across separate estimates for English use at home, work, and school on a 0–100% scale). All continuous fixed-effects predictors were *z*-transformed to improve the interpretability of results, so that estimates for predictor variables could be interpreted in relation to the intercept (i.e., the grand mean) of the outcome variable. Among categorical predictors, the audio only (image) condition and the Chinese group were designated as the reference groups. However, because the condition variable included three levels, to obtain model estimates for the final comparison (audio with expressions vs. full video), the model was releveled, such that the audio with expressions condition was designated as the reference group. To perform multiple comparisons for the condition variable (audio only vs. audio with expressions vs. full video), a Tukey correction for *p* values was applied using the glht package in *R* (version 1.5.1, Lenth, 2020).[2]

To address the second research question, which focused on visual cues associated with rater assessments of L2 speakers' comprehensibility, accentedness, and fluency, we computed another set of linear mixed-effects models to explore the relationship between L2 speech ratings and the coded visual cues in the full video condition, where all visual cues were available to the raters. In each set of models, comprehensibility, accentedness, and fluency served as the outcome variables while speakers' L1 (Chinese, Spanish), eight visual cues (head movement, eyebrow

movement, looking away, blinking, smiling and laughing, lip movement, referential gestures, temporal highlighting) and their interaction with speakers' L1 served as fixed-effects predictors, and raters (60) and speakers (20) were entered as random-effects predictors (for a total of 400 observations per model). The Chinese group was designated as the reference group, and the raw counts of visual cues were *z*-transformed. Because these analyses were conducted using a smaller dataset, we did not include speaker-level covariates in these models to ensure they were sufficiently simple to converge (Baayen et al., 2008).

All models for both research questions were fit using the maximum likelihood method, and model fit was evaluated through pairwise likelihood ratio tests (Barr et al., 2013), comparing simpler to more complex models. Random slope models were examined, separately for raters and speakers. However, the inclusion of random slopes did not improve model fit for any outcome variable or the models did not converge, so only the random intercepts of speakers and raters were included in the final models. For selecting fixed-effects predictors, we took an exploratory approach by forward-testing the predictors, and then tested the interactions only when the inclusion of a predictor improved model fit. Although there are no agreed-upon criteria for estimating adequate sample sizes for mixed-effects models (Maas & Hox, 2005; McNeish & Stapleton 2016), the estimates provided by Scherbaum and Ferreter (2008) suggested that a sample of 20 speakers and 60 raters was sufficient to achieve power of .80 with a medium effect size. Therefore, the current data sample was deemed sufficiently large for mixed-effects modeling. Similarly, because there is no consensus regarding the criteria for considering statistical significance in mixed-effects modelling, for example, with some researchers using *t* values greater than 2 to imply statistical significance (Linck & Cunnings, 2015), we obtained *p* values using MuMIn package in *R* (version 1.43.17, Bartoń, 2020) but examined 95% confidence

intervals (CIs) to check the statistical significance of each parameter (interval does not cross zero).

Results

Raters' Access to Visual Cues

The first research question asked whether there was a difference in comprehensibility, accentedness, and fluency ratings based on the raters' access to an L2 speaker's visual cues. As shown in Table 6, descriptively the Chinese speakers generally elicited lower ratings from the raters than the Spanish speakers, and comprehensibility and accentedness (but not fluency) ratings tended to increase from the visually least informative condition (audio only) to the most informative condition (full video).

Table 6. *Summary Statistics for Speech Ratings by Condition and Speaker Group*

	Audio only		Audio with expressions		Full video	
Rated variable	*M*	*SD*	*M*	*SD*	*M*	*SD*
Comprehensibility						
Chinese	493.13	281.08	534.08	248.52	578.35	256.17
Spanish	653.75	235.88	676.34	244.07	702.62	229.88
Accentedness						
Chinese	332.27	280.09	324.27	245.09	433.28	277.51
Spanish	492.77	295.42	475.61	312.91	570.95	303.57
Fluency						
Chinese	500.77	272.67	465.22	226.70	500.06	270.17
Spanish	665.99	225.10	638.52	243.95	690.51	256.56

Table 7 summarizes the final mixed-effects model for comprehensibility. The interaction between condition and speakers' L1 was not significant, *Estimate* = –36.36, *SE* = 27.19, t = –1.34, p = .182, and it did not improve model fit, $\chi^2(2) = 1.79$, p = .409, so the interaction term was excluded from the final model. With respect to the role of visual cues, there were no statistically significant differences in comprehensibility ratings across the three conditions, suggesting that comprehensibility did not vary dramatically based on the raters' access to a speaker's visual cues. However, the raters who had access to the full video tended to provide higher ratings than those with access to audio only (on average +70 points on a 1,000-point scale). Although this difference was not statistically significant after a conservative (Tukey) adjustment for multiple comparisons (p = .101), this finding is nevertheless noteworthy, inasmuch as the 95% CI for this effect did not cross zero and the t value was greater than 2

(Linck & Cunnings, 2015). At minimum, the raters appeared to demonstrate an upward trend in their comprehensibility ratings (see Table 6) from the least visually informative condition (audio only) to the most visually informative situation (full video). In terms of speakers' L1, although the magnitude of the estimate was large, the raters assigned similar comprehensibility ratings to the Spanish and Chinese speakers after we controlled for speaker-level covariates capturing various aspects of the speakers' self-assessed L2 proficiency and use (length of residence, speaking and listening proficiency, daily L2 use). None of the covariates explained any additional model variance in comprehensibility ratings.

Table 7. *Summary of Final Mixed-Effects Model for Comprehensibility*

Parameter	*Estimate*	*SE*	*95% CI*	*t*	*p*
(Intercept)	573.44	33.67	[506.17, 640.71]	17.03	< .001
Condition					
Audio with expressions vs. Audio only	30.54	32.75	[–33.47, 97.01]	0.97	.596
Full video vs. Audio only	69.73	32.75	[1.80, 132.28]	2.05	.101
Full video vs. Audio with expressions	35.27	32.75	[–29.97, 100.52]	1.08	.528
Spanish vs. Chinese	82.91	58.56	[–37.63, 203.44]	1.42	.172
Speaker-level covariates					
Length of residence	49.79	26.01	[–3.75, 103.32]	1.91	.089
L2 speaking and listening	44.77	34.20	[–25.62, 115.16]	1.31	.210
Daily L2 use	–3.13	31.06	[–67.06, 60.80]	–0.10	.901
Random effects	*Variance*	*SD*	*Criterion*	*Estimate*	
Rater (intercept)	8875	94.21	Log-likelihood	–7704.30	
Speaker (intercept)	11951	109.32	AIC	15428.50	
			BIC	15478.90	
			Marginal R^2	0.15	
			Conditional R^2	0.46	

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, marginal R^2 = variance explained by fixed factors, conditional R^2 = variance explained by fixed and random factors. Final model formula: Comprehensibility ~ condition + speakers' L1 + length of residence + L2 speaking and listening + daily L2 use + (1|speaker) + (1|rater).

Table 8 summarizes the final mixed-effects model for accentedness. As with comprehensibility, the interaction between condition and speakers' L1 was not statistically significant, *Estimate* = –22.82, *SE* = 25.36, *t* = –0.90, *p* = .368, and it did not improve model fit, $\chi^2(2) = 0.82$, *p* = .663; therefore, the interaction term was excluded from the final model. With respect to the role of visual cues, accentedness ratings in the full video condition were significantly higher than those in the audio only condition (*p* = .051) and in the audio with expressions condition (*p* = .021), which did not differ between each other. Thus, the raters who had access to both facial expressions and gestures perceived the speakers as being significantly less accented than those seeing only static images (+90 points) and those exposed only to the speakers' facial expressions (+102 points). For speakers' L1, as with comprehensibility, the effect of speaker group was not significant after controlling for speaker-level covariates. However, the speakers' length of residence in Canada accounted for additional variance in accentedness ratings (*p* = .001), such that the speakers with longer residence in Canada (regardless of the visual condition they were evaluated in) were rated as less accented (+110 points) than those with shorter residence in Canada.

Table 8. *Summary of Final Mixed-Effects Model for Accentedness*

Parameter	*Estimate*	*SE*	*95% CI*	*t*	*p*
(Intercept)	412.52	38.39	[335.94, 489.09]	10.75	< .001
Condition					
Audio with expressions vs. Audio only	–12.58	38.29	[–88.85, 63.69]	–0.33	.942
Full video vs. Audio only	89.59	38.29	[13.32, 165.86]	2.34	.051
Full video vs. Audio with expressions	102.29	38.29	[25.90, 178.44]	2.67	.021
Spanish vs. Chinese	20.65	64.72	[–112.55, 153.84]	0.32	.753
Speaker-level covariates					
Length of residence	110.02	28.74	[50.86, 169.18]	3.83	.001
L2 speaking and listening	65.78	37.79	[–12.01, 143.57]	1.74	.097
Daily L2 use	38.22	34.33	[–32.43, 108.87]	1.11	.279

Random effects	*Variance*	*SD*	*Criterion*	*Estimate*
Rater (intercept)	13050	114.20	Log-likelihood	–8029.20
Speaker (intercept)	14810	121.70	AIC	16078.30
			BIC	16129.20
			Marginal R^2	0.33
			Conditional R^2	0.64

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, marginal R^2 = variance explained by fixed factors, conditional R^2 = variance explained by fixed and random factors. Final model formula: Accentedness ~ condition + speakers' L1 + length of residence + L2 speaking and listening + daily L2 use + (1|speaker) + (1|rater).

Table 9 summarizes the final mixed-effects model for fluency. Again, the interaction between condition and speakers' L1 was not significant, *Estimate* = 25.23, *SE* = 24.49, $t = -1.03$, $p = .303$, and it did not improve model fit, $\chi^2(2) = 1.11$, $p = .575$, so the interaction term was not included in the final model. As for the role of visual cues, there were no significant differences in fluency ratings across the three conditions, suggesting that the raters' perceptions of speaker fluency did not vary dramatically based on access to a speaker's visual cues. After controlling for speaker-level covariates, although the magnitude of the estimate was large, there were again no differences between the fluency ratings assigned to the two L1 speaker groups. However, as with accentedness, the speakers' length of residence in Canada accounted for additional variance in fluency ($p = .027$), such that the speakers who resided in Canada longer elicited higher fluency ratings (+ 56 points) than the speakers with shorter residence in Canada.

Table 9. *Summary of Final Mixed-Effects Model for Fluency*

Parameter	*Estimate*	*SE*	*95% CI*	*t*	*p*
(Intercept)	583.38	37.06	[509.76, 657.00]	15.74	< .001
Condition					
Audio with expressions vs. Audio only	–31.51	42.16	[–115.49, 52.46]	–0.75	.735
Full video vs. Audio only	11.90	42.16	[–72.08, 95.88]	0.28	.957
Full video vs. Audio with expressions	43.41	42.16	[–40.56, 127.39]	1.03	.558
Spanish vs. Chinese	100.99	52.73	[–7.57, 209.54]	1.92	.070
Speaker-level covariates					
Length of residence	55.80	23.42	[7.59, 104.02]	2.34	.027
L2 speaking and listening	59.50	30.80	[–3.89, 122.90]	1.93	.068
Daily L2 use	–4.76	27.97	[–62.34, 52.82]	–0.17	.867
Random effects	*Variance*	*SD*	*Criterion*	*Estimate*	
Rater (intercept)	16271	127.56	Log-likelihood	–7991.80	
Speaker (intercept)	9689	98.43	AIC	16003.60	
			BIC	16054.50	
			Marginal R^2	0.21	
			Conditional R^2	0.58	

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, marginal R^2 = variance explained by fixed factors, conditional R^2 = variance explained by fixed and random factors. Final model formula: Fluency ~ condition + speakers' L1 + length of residence + L2 speaking and listening + daily L2 use + (1|speaker) + (1|rater).

Specific Visual Cues and L2 Speech Ratings

The second research question explored the relationship between L2 speakers' use of individual visual cues (e.g., facial expressions, head movement, hand gestures) and rater perceptions of these speakers' comprehensibility, accentedness, and fluency. This question also examined whether this relationship varied by speaker background (Chinese vs. Spanish) because the use of visual cues might differ across different cultures (Kita, 2009; Nicoladis et al., 2018). Table 10 summarizes descriptive statistics for the occurrence of visual cues in the full video condition (i.e., where all visual cues were presumably accessible to the raters). The Chinese speakers generally tended to move their eyebrows less frequently (–6.35 counts per video), produced fewer head movements (–3.75 counts), and made fewer referential gestures (–2.65 counts) than the Spanish speakers, but no differences across any coded categories reached statistical significance, implying that (at least in this dataset) both speaker groups produced comparable frequencies of visual cues in the videos.

Table 10. *Descriptive Statistics for Occurrence of Visual Cues in Full Videos by Speaker Group*

	Chinese		Spanish		Comparison		
Visual cue	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
Head movement	4.60	4.03	8.10	7.53	–1.30	.212	0.58
Eyebrow movement	8.00	8.47	15.70	13.58	–1.52	.146	0.68
Looking away	27.20	11.01	28.00	7.23	–0.19	.850	0.09
Blinking	52.70	19.49	60.40	17.88	–0.92	.369	0.41
Smiling and laughing	4.70	3.13	4.50	2.92	0.15	.884	0.07
Lip movement	0.70	1.06	1.20	3.12	–0.48	.637	0.21
Referential gestures	5.70	4.74	9.10	7.13	–1.26	.225	0.56
Beat gestures	11.70	5.52	13.90	7.89	–0.72	.479	0.32

For comprehensibility (with the final model summarized in Table 11), among the eight visual cues examined, looking away was the only fixed-effects predictor that improved model fit compared to the baseline model ($p = .014$). The frequency of looking away did not interact with speakers' L1, *Estimate* = 37.01, *SE* = 51.61, $t = 0.72$, $p = .482$, and did not improve model fit, $\chi^2(1) = 0.51$, $p = .476$, so the interaction term was excluded from the model. The speakers who tended to look away from their interlocutor more frequently elicited higher comprehensibility ratings from the raters (+65 points), compared to the speakers who engaged in this visual behavior less frequently. Although the Spanish speakers were generally rated as more comprehensible (+118 points) than the Chinese speakers, this result should be interpreted with caution because the 95% CI crossed zero. In addition, to minimize model complexity and maximize result interpretation, mixed-effects modeling for the second research question did not

include speaker-level covariates, which tended to explain between-speaker differences in previous analyses.

Table 11. *Summary of Final Mixed-Effects Model for Comprehensibility and Visual Cues*

	Fixed effects					Random effects	
Parameter	*Estimate*	*SE*	*t*	*95% CI*	*p*	*Source*	*SD*
(Intercept)	640.48	31.76	20.17	[573.47, 707.50]	< .001	Speaker	98.00
Speaker L1	118.41	47.89	2.47	[–23.29, 202.34]	.023	Rater	93.41
Looking away	64.87	23.98	2.71	[–4.95, 108.01]	.014		

Note. Final model formula: Comprehensibility ~ speakers' L1 + looking away + (1|speaker) + (1|rater).

For accentedness (Table 12), only eyebrow movement emerged as a significant predictor (p = .011), and the frequency of eyebrow movement did not interact significantly with speakers' L1, *Estimate* = 127.01, *SE* = 72.77, t = 1.75, p = .097, or improved model fit, $\chi^2(1) = 2.83$, p = .092. The speakers who moved their eyebrows more frequently were rated as less accented (+99 points) than the speakers who engaged in this visual behavior less often. Speakers' L1 (Chinese vs. Spanish) did not explain additional variance in accentedness ratings.

Table 12. *Summary of Final Mixed-Effects Model for Accentedness and Visual Cues*

	Fixed effects					Random effects	
Parameter	*Estimate*	*SE*	*t*	*95% CI*	*p*	*Source*	*SD*
(Intercept)	502.11	46.35	10.83	[408.79, 595.43]	< .001	Speaker	141.50
Speaker L1	70.91	70.14	1.01	[–73.51, 215.33]	.324	Rater	145.50
Eyebrow movement	99.02	35.11	2.82	[26.72, 171.32]	.011		

Note. Final model formula: Accentedness ~ speakers' L1 + eyebrow movement + (1|speaker) + (1|rater).

For fluency (Table 13), again, eyebrow movement was the only significant fixed-effects predictor (p = .012), and the frequency of eyebrow movement did not interact with speakers' L1, *Estimate* = 56.20, *SE* = 62.52, t = 0.90, p = .380, or improved model fit, $\chi^2(1) = 0.79$, p = .374. The speakers who moved their eyebrows more frequently elicited higher fluency ratings from the raters than the speakers who showed less eyebrow movement (+79 points). The Spanish speakers were also generally rated higher than the Chinese speakers (+137 points); however, to minimize model complexity and maximize result interpretation, mixed-effects modeling for this research question did not include speaker-level covariates, which tended to explain between-speaker differences in previous analyses.

Table 13. *Summary of Final Mixed-Effects Model for Fluency and Visual Cues*

	Fixed effects					Random effects	
Parameter	*Estimate*	*SE*	*t*	*95% CI*	*p*	*Source*	*SD*
(Intercept)	595.28	41.89	14.21	[510.84, 679.72]	< .001	Speaker	113.90
Speaker L1	137.07	57.22	2.40	[19.21, 254.93]	.027	Rater	143.50
Eyebrow movement	79.18	28.65	2.76	[20.17, 138.18]	.012		

Note. Final model formula: Fluency ~ speakers' L1 + eyebrow movement + (1|speaker) + (1|rater).

Discussion

This study's main objective was to examine whether rater assessments of L2 comprehensibility, accentedness, and fluency vary as a function of visual cues (facial expressions, hand gestures) available to raters. An additional goal was to explore possible relationships between individual visual cues (e.g., blinking, eyebrow movements, beat gestures) and raters' evaluations of comprehensibility, accentedness, and fluency for speakers from different L1s (Chinese, Spanish), on the assumption that speakers from different backgrounds might vary in their use of visual cues. Visual information appeared to impact raters in their evaluations of L2 accentedness (with a similar trend for comprehensibility) but not fluency, such that the raters who had access to the full, dynamic videos (full video condition) rated L2 speakers as less accented and also tended to judge them as more comprehensible than the raters who evaluated the same speakers while looking at their static images (audio only condition). Although the Chinese speakers generally elicited lower evaluations from the raters than the Spanish speakers, these differences could be accounted through individual differences in the speakers' L2 proficiency and use (used as control covariates). Most importantly, there were few

differences across the Chinese and Spanish speakers in their production of various visual cues, and the speakers' language background did not interact with visual condition, implying that the impact of visual cues on the raters was similar for the speakers from the Chinese and Spanish backgrounds. In terms of specific visual cues associated with ratings, frequent eyebrow movements were associated with less accented and more fluent speech, while frequent looks away were linked to higher comprehensibility.

Role of Visual Cues in L2 Speech Ratings

For comprehensibility, although only the difference between the static and the fully dynamic viewing conditions approached significance after a conservative adjustment for multiple comparisons, the raters tended to progressively enhance their evaluations as more visual information was available to them, assessing L2 speakers on average at 573 (on a 1,000-point scale) when looking at static images, at 605 when having access to facial expressions but not gestures, and at 640 while watching the entire video (facial expressions along with gestures). An incremental, additive effect of visual cues on comprehensibility is consistent with raters' using the aural input in addition to various cues available from a speaker's face, with additional support from gestures (Sueyoshi & Hardison, 2005). In terms of facial expressions, in addition to a speaker's looks away from the interlocutor, the raters may have generally benefitted from seeing a speaker's articulatory configurations (e.g., lip and jaw movement) and facial cues (e.g., eyebrow raises co-occurring with speech rhythm), which highlighted linguistic information (Chui, 2005; Swerts & Krahmer, 2008) and made it easier to understand the speaker. Similarly, a speaker's visual display of emotional reactions (e.g., surprise, confusion), particularly in an emotion-laden task such as exchanging close-call narratives, may have helped the raters

anticipate what kind of information was going to be shared (Wagner, 2008), thereby decreasing their processing effort.

For accentedness, there was no evidence of an incremental, additive effect, in the sense that the raters perceived a speaker as less accented only when the full dynamic video was available to them, assessing the speakers on average at 502 (on a 1,000-point scale), compared to the situation when the raters had no access to visual information (412) or when they observed a speaker's dynamic face but no gestures (400). This finding, which implies that the visual cues available in a speaker's face presented alone (with no hand gestures) were largely inconsequential to rater evaluations of accent, aligns well with prior work on visual cues in phoneme-focused tasks, where listeners did not benefit from seeing a speaker's face in their identification of L2 vowels (Inceoglu, 2019) or in their transcription of monosyllabic verbs spoken with or without vowel errors (Wheeler, 2019).[3]

The raters seemed to benefit the most from having access to a speaker's gestures, as only the full dynamic condition resulted in significantly enhanced accentedness (i.e., with speakers rated as less accented) and in a similar trend for increased comprehensibility (i.e., with speakers rated as more comprehensible), relative to the other viewing situations. Although the visual behaviors with the strongest links to comprehensibility and accentedness in the full video condition were a speaker's eyebrow movements and looks away from the interlocutor (see Tables 11 and 12), it is possible that a speaker's hand gestures—and especially beat gestures which were the most prevalent gesture category—provided complementary visual information to the raters, leading them to perceive the speaker as being easier to understand and less accented. At minimum, beat gestures highlighted prosodic structure for the listener, which simplified speech segmentation, and emphasized particularly important content, which aided listener

comprehension (Drijvers & Özyürek, 2017, 2020; Hardison, 2018; Sueyoshi & Hardison, 2005; Wheeler, 2019). Alternatively, the use of gestures may have allowed speakers to project a greater level of speaking proficiency or confidence (Neu, 1990), which may have been captured in less accented (and more comprehensible) speech ratings from the raters who had access to gestures.

The positive impact of the full audiovisual viewing condition on rater evaluations of accentedness may have also stemmed from other visual cues available in the full video but absent in the other conditions. Apart from a speaker's hand gestures, the full video condition allowed the raters to observe a speaker's body posture and body movement, which may have influenced the ratings. For example, a speaker's relaxed body position might signal confidence to raters, leading them to evaluate the speaker more favorably in job interviews (Jenkins & Parra, 2003) and perhaps also to upgrade the speaker in their ratings. Similarly, natural body movements (e.g., leaning forward, turning, raising shoulders), particularly if they are congruent with discourse content and are synchronized with the social cues provided by the interlocutor, might be perceived as showing greater engagement (e.g., Hardison, 2018) and might therefore contribute to enhanced ratings. Put differently, access to the full, embodied representation of the speaker telling a story may have assisted the raters in predicting story content, resolving potential ambiguities, and ultimately arriving at mutual understanding (Hostetter & Alibali, 2008), which may have been captured in rater evaluations of speakers' accentedness (and possibly comprehensibility).

Finally, when it comes to fluency ratings, there were no differences across the three viewing conditions in assessments of L2 speakers' fluency, suggesting that fluency ratings are not susceptible to the effects of input modality, at least in this study (but see Nakatsuhara et al., 2021). Fluency ratings can largely be accounted for by temporal measures of speech, such as

articulation speed, pausing, and syllable length (Bosker et al., 2013; Kahng, 2018). For instance, when raters assessed L2 speakers' audios, as much as 84% of the total variance in fluency ratings could be attributed to temporal measures, including mean length of syllables and frequency and duration of pausing (Bosker et al., 2013). Because temporal dimensions of speech would likely be salient to the listener from the audio channel alone, with or without access to visual cues, it is unsurprising that visual information made little contribution to rater-assessed fluency.

Specific Visual Cues and L2 Speech Ratings

An additional goal of this study was to examine which individual visual cues are associated with rater perceptions of L2 comprehensibility, accentedness, and fluency for speakers from different language backgrounds. The L1 Chinese and Spanish speaker groups were examined separately, on the assumption that the use of visual cues might be related to speakers' linguistic or cultural background (Gullberg, 2006; Iverson et al., 2008; Kita, 2009). Contrary to our expectations, there were few differences in the use of visual cues between the two speaker groups (see Table 10), which implies that the occurrence of visual behaviors during conversation may not always be attributable to cultural or linguistic differences, at least in interactions involving two university-level students using English to communicate with each other. Additionally, L2 speakers' language background did not interact with the visual condition to yield a difference in ratings for either group, which suggests that the raters relied on visual information in similar ways as they evaluated speakers from both language backgrounds. Lastly, L1 group effects were either weak or inconsistent (e.g., with 95% CI crossing zero) in the analyses of specific visual cues used by the speakers. Instead, individual differences in speakers' proficiency and use, such as their length of residence in Canada, predicted their accentedness and fluency ratings, highlighting the experiential dimension of pronunciation learning.

Although comparisons across the three visual conditions only implied that some visual cues may have factored into the raters' assessments, focused analyses of individual visual cues provided (preliminary) evidence of the relevance of specific cues to each rating. Briefly, when all visual cues were available to raters (i.e., in the full video condition), eyebrow movements, such as raises and frowns, were positively associated with the raters' assessments of accentedness and fluency. Eyebrow movement is known to highlight speech prosody by signaling important content and marking phrase boundaries (Pelachaud et al., 1996), and the raters may have benefitted from this information (to a greater or lesser extent) because it enhanced prosodic cues to speech segmentation and comprehension (Krahmer & Swerts, 2007). In this dataset, eyebrow movement seemed to co-occur with speakers' use of hand gestures to highlight their production of prosodic (nuclear) stress in an intonation phrase, as illustrated in the examples below (where | marks a phrase boundary, italics designate the word carrying nuclear stress, and ↑ indicates the location of an eyebrow raise along with a speaker's use of a hand gesture).

S200: and in the ↑*back* | because it's a vehicle that's a little bit ↑*bigger* | it has a fire extinguisher that's ↑*attached* | just in case there's a ↑*fire*

S101: I feel very like ↑*relaxed* | and I want to hang *out* | and I went to ↑*shopping* mall

For some speakers, such as S200, many nuclear stresses coincided with an eyebrow movement and a beat (hand) gesture, whereas for others, like S101, only some stresses were accompanied by such visual cues. Yet even infrequent visual signals (especially when they involve dual cues, such as an eyebrow raise and an up-down hand gesture) may have enhanced L2 speakers' use of prosody for the raters, highlighting important content (Hardison, 2018; Krahmer & Swerts, 2007) and leading to more favorable evaluations of L2 speakers. That such favorable assessments extended to accentedness and fluency is not altogether surprising, given strong links between

prosody and accentedness (Kang, 2010) and fluency (Préfontaine & Kormos, 2016). What must be clarified in future work, however, is whether temporal highlighting of prosody through facial expressions (e.g., eyebrow movement) is more salient and thus more useful for raters as a cue, with and without temporal highlighting through hand gesture.

Looking away (breaking eye contact) was the only visual cue associated positively with comprehensibility. Looking away may have been particularly beneficial for the raters because speakers tend to look away during speech planning or when processing complex information (Knapp & Hall, 2001). Continuous eye contact with an interlocutor appears to interfere with a speaker's production of spontaneous speech (Beattie & Hughes, 1987), whereas gaze aversion, particularly during complex tasks, seems to aid a speaker in cognitive functioning, leading to improvement in performance, such as when answering complex arithmetic and reasoning questions (Glenberg et al., 1998). Inspection of the data indeed showed that the L2 speakers most frequently looked away while pausing, presumably when they were planning their next utterance (Beattie & Hughes, 1987) or dealing with an increased arousal while describing personally-relevant emotional experiences (MacPherson et al., 2017). As illustrated in the examples below, for some speakers, such as S52, looking away coincided with filled pauses or hesitations (enclosed in brackets), while for others, like S269, looking away occurred at phrase boundaries.

S52: I almost got robbed uh… on the street [by uh by uh… like] three or two kids [and uh… it was uh…] it was the night, and I walk with my friends [and uh we were walking uh] in a dark street

S269: I was spending the whole week doing this project [look away] so I almost lost it [look away] but then I remember I send it in my iCloud [look away] so when I open my computer I had to start my computer again

Also, the frequency of L2 speakers looking away from their interlocutor likely captured the attested tendency for listeners to look at their interlocutor's face for sustained periods of time but for speakers to alternate between brief eye gazes to the interlocutor and looks away (Bavelas et al., 2002; Turkstra, 2005). It may well be, then, that those L2 speakers who engaged in such expected behaviors (i.e., alternating their eye gaze between looking at their partner and looking away, without staring at the interlocutor while speaking) tended to be evaluated higher by the raters. An interim conclusion, until confirmed in future work, is that looking away not only was the behavior associated with speech planning, which may have promoted cohesive storytelling, but also one that was expected by an external observer, with the consequence that this visual behavior alleviated at least some processing burden for the raters.

In this dataset, the coded categories of blinking, lip and head movement, instances of laughter and smiling, and referential gestures were not associated with any rater-assessed speech dimensions. Although blinking as well as lip and head movement have been shown to facilitate listeners' processing of linguistic information in speech, typically in phoneme-focused tasks (Drijvers & Özyürek, 2017, 2020; Munhall et al., 2004; Zheng & Samuel, 2019), this work generally targeted specific articulatory configurations (e.g., lip rounding in production of particular vowels) through fine-grained measures, such as movement velocity, distance between lips, or angles of head movement. By contrast, in this study, L2 speakers' blinking as well as lip and head movements were coded broadly (e.g., pursing and curling of lips, tongue touching lips), without any reference to specific articulatory configurations, which may explain why these categories were not strongly linked to L2 speech assessments. As for instances of positive emotion (smiling, laughter) and referential gestures, our findings (at least with respect to gestures) may seem to contradict prior work showing positive contribution of gesture to

improving word-level intelligibility (Drijvers & Özyürek, 2017, 2020; Wheeler, 2019) and utterance-level comprehension (Sueyoshi & Hardison, 2005). However, the lack of strong associations involving positive emotion and referential gestures could be partly explained by the relatively infrequent use of these cues in this dataset (see Table 10), which made it challenging for meaningful associations to emerge. Also, it is plausible that raters may attend to not only frequency but also function (Gullberg et al., 2008) and appropriateness (Jenkins & Parra, 2003) of referential gestures and displays of positive emotion when assessing L2 speech. Therefore, these issues may need to be revisited in a targeted future investigation using a larger, more representative dataset.

Implications, Limitations, and Future Work

The current findings offer several implications for the use of visual stimuli in L2 speech assessment. In the context of a global health crisis, people are increasingly taking advantage of the multimedia, such as videoconferencing tools, with visual stimuli now routinely used for language training and testing. Because having access to facial expressions and hand gestures appears to enhance perceptions of L2 speech for the listener (particularly for accentedness), whenever possible, L2 speakers may benefit from choosing video over an audio format, for instance, for communication involving work, online learning, or job interviews. In terms of instruction, teachers might wish to raise L2 speakers' awareness of the potential relevance of visual cues (e.g., eyebrow raises) to listener perceptions (Hardison, 2018), particularly for such listener-centered constructs as accentedness and comprehensibility. In terms of language assessment, our results are in agreement with Nakatsuhara et al.'s (2021) findings, where test takers' speaking performances elicited higher IELTS scores in terms of fluency, lexis, grammar, and pronunciation in the video and live (face-to-face) assessment conditions than in the audio

condition. Taken together, these findings suggest that both trained test examiners and untrained raters are sensitive to visual cues when assessing L2 speech for a variety of rated dimensions. However, to understand the nature and scope of modality effects on rater-based assessments of L2 speech, researchers may need to examine whether specific speech ratings, such as fluency, might be more or less susceptible to modality effects as a function of rater training or experience. Researchers might also need to continue exploring modality effects for L2 comprehensibility and accentedness, with a focus on a potential threshold beyond which raters' access to audiovisual information may no longer enhance their perceptions but instead may detract them from L2 speech (e.g., Mayer & Moreno, 2003; Ockey, 2007), causing variability in performance.

To close, a few limitations to this exploratory work should be acknowledged. First, to provide a more nuanced understanding of language and culture differences in the impact of visual cues on the listener, future research needs to examine L2 speakers from more diverse language and culture backgrounds. Because only a select few visual cues emerged as relevant to L2 speech ratings (and only in quantitative analyses), it would be important to investigate rater perception of various visual cues through targeted qualitative investigations. Although we captured the incidence of some visual cues through frequency counts, future studies also need to capture not just the frequency of visual cues but their function and appropriateness, which was outside the scope of this initial, exploratory work. Similarly, as shown by large differences between marginal and conditional R^2 values in mixed-effects models, a substantial amount of variance in the ratings was attributable to random effects (i.e., individual differences) across speakers and raters. Although we have controlled for a few speaker-level covariates, researchers might wish to isolate such sources of variance targeting, for example, differences in L2 proficiency and personality for speakers (Gregersen et al., 2009; Hostetter & Potthoff, 2012;

O'Carroll et al., 2015) and variation in cognitive ability for raters (Chu et al., 2014; Smithson & Nicoladis, 2013). Finally, because this study's findings are associational in nature, we do not know if the raters actually noticed, attended to, or in any way processed specific visual cues available to them. While informal observations of the raters' behaviors during the rating sessions suggested that they remained engaged with video stimuli, other methodologies must be employed to show whether and to what degree the listener attends to a speaker's visual cues (e.g., Gullberg & Holmqvist, 2006). Thus, we call for future research using eye-tracking accompanied by stimulated recall to investigate raters' cognitive processing while assessing L2 speech.

Notes

1. Standardized proficiency scores were not available for all speakers, so it was impossible to include TOEFL or IELTS scores as a control covariate.
2. We would like to thank an anonymous reviewer for suggesting various excellent improvements to mixed-effects modeling and directing us to appropriate R resources.
3. Given that numerically lowest accentedness ratings occurred in the audio with expressions condition, as suggested by an anonymous reviewer, it may be that visual information, such as non-target lip and jaw movements, is detrimental to rater perceptions of accent (Kawase et al., 2014).

References

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes, 15,* 593–613.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278.

Bartoń, K. (2020). *MuMIn: Multi-Model Inference*. R package version 1.43.17. https://CRAN.R-project.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Bates, E., & Dick, F. (2002). Language, gesture, and the developing brain. *Developmental Psychobiology, 40,* 293–310.

Batty, A. O. (2014). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing, 32,* 3–20.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication, 52,* 566–580.

Beattie, G. W., & Hughes, M. (1987). Planning spontaneous speech and concurrent visual monitoring of a televised face: Is there interference? *Semiotica, 65,* 97–106.

Beattie, G. W., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica, 123*, 1–30.

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contribution of pauses, speed and repairs. *Language Testing, 30,* 159–175.

Chu, M., Meyer, A., Foulkes, L., & Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, *143*, 694–709.

Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse, *Journal of Pragmatics, 37,* 871–887.

Cobb, T. (2019). VocabProfilers [computer program]. https://www.lextutor.ca/vp

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation Fundamentals*. John Benjamins.

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research, 60*, 212–222.

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

Glenberg, A. M., Shroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition*, *26*, 651–658.

Gregersen, T., Cuhat, G. O., & Storm, J. (2009). An examination of L1 and L2 gesture use: What role does proficiency play? *The Modern Language Journal*, *93*, 195–208.

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

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *International Review of Applied Linguistics in Language Teaching*, *44*, 103–124.

Gullberg, M. (2010). Methodological reflections on gesture analysis in second language acquisition and bilingualism research, *Second Language Research, 26*, 75–102.

Gullberg, M., & Holmqvist, K. (2006). What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition, 14*, 53–82.

Gullberg, M., & McCafferty, S. G. (2008). Introduction to gesture and SLA: Toward an integrated approach. *Studies in Second Language Acquisition, 30*, 133–146.

Gullberg, M., De Bot, K., & Volterra, V. (2008). Gestures and some key issues in the study of language development. *Gesture, 8*, 149–179.

Hardison, D. M. (2018). Visualizing the acoustic and gestural beats of emphasis in multimodal discourse: Theoretical and pedagogical implications. *Journal of Second Language Pronunciation*, *4*, 232–259.

Hayes-Harb, R., & Hacking, J. F. (2015). Beyond rating data: What do listeners believe underlies their accentedness judgments? *Journal of Second Language Pronunciation*, *1*, 43–64.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*, 495–514.

Hostetter, A. B., & Potthoff, A. L. (2012). Effects of personality and social situation on representational gesture production. *Gesture, 12*, 63–83.

Inceoglu, S. (2019). Individual differences in L2 speech perception: The role of phonological memory and lipreading ability. *The Modern Language Journal*, *103*, 782–799.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*, 475–505.

Iverson, J. M., Capirci, O., Volterra, V., & Goldin-Meadow, S. (2008). Learning to talk in a gesture-rich world: Early communication in Italian vs. American children, *First Language, 28*, 164–181.

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*, 90–107.

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*, 569–591.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*, 301–315.

Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, *28*, 441–456.

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

Kelly, S. D., Manning, S. M., & Rodak, S. (2008). Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass*, *2*, 569–588.

Kendon, A. (1994). Do gestures communicate? A review. *Research on Language and Social Interaction*, *27*, 175–200.

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

Knapp, M. L., & Hall, J. A. (2001). *Nonverbal communication in interaction*. Holt, Rinehart and Winston.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*, 396–414.

Kutlu, E. (2020). Now you see me, now you mishear me: Raciolinguistic accounts of speech perception in different English varieties. *Journal of Multilingual and Multicultural Development*. Advance Online Publication.

Lenth, R. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.5.1. https://CRAN.R-project.org/package=emmeans

Li, P., Baills, F., & Prieto, P. (2020). Observing and producing durational hand gestures facilitates the pronunciation of novel vowel-length contrasts. *Studies in Second Language Acquisition*. Advance Online Publication.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*(S1), 185–207.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample size for multilevel modeling. *Methodology*, *1*, 86–92.

MacPherson, D., Abur, D., & Stepp, C. (2017). Acoustic measures of voice and physiologic measures of autonomic arousal during speech as a function of cognitive load. *Journal of Voice*, *31*, 504.e1–504.e9.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*, 43–52.

McCafferty, S. G. (2002). Gesture and creating zones of proximal development for second language learning. *The Modern Language Journal*, *86*, 192–203.

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*, 295–314.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, *15*, 133–137.

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

Nambiar, M. K., & Goon, C. (2016). Assessment of oral skills: A comparison of scores obtained through audio recordings to those obtained through face-to-face evaluation. *RELC Journal*, *24*, 15–31.

Nakatsuhara, F., Inoue, C., & Taylor, L. (2021). Comparing rating modes: Analysing live, audio, and video ratings of IELTS speaking test performances. *Language Assessment Quarterly*, *18*, 83–106.

Neu, J. (1990). Assessing the role of nonverbal communication in the acquisition of communicative competence in L2. In R. C. Scarcella, E. S. Andersen, & S. D. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 121–138). Newbury House.

Nicoladis, E., Nagpal, J., Marentette, P., & Hauer, B. (2018). Gesture frequency is linked to story-telling style: Evidence from bilinguals. *Language and Cognition*, *10*, 641–664.

O'Carroll, S., Nicoladis, E., & Smithson, L. (2015). The effect of extroversion on communication: Evidence from an interlocutor visibility manipulation. *Speech Communication, 69,* 1–8.

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

Pelachaud, C., Badler, N. I., & Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science, 20,* 1–46.

Pika, S., Nicoladis, E., & Marentette, P. F. (2006). A cross-cultural study on the use of gestures: Evidence for cross-linguistic transfer? *Bilingualism: Language and Cognition, 9,* 319–327.

Préfontaine, Y, & Kormos, J. (2016). A qualitative analysis of perceptions of fluency in second language French. *International Review of Applied Linguistics in Language Teaching, 54,* 151–169.

R Core Team. (2020). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning, 69,* 652–708.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics, 38,* 439–462.

Scarborough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech, 52,* 135–175.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12*, 347–367.

Smithson, L., & Nicoladis, E. (2013). Verbal memory resources predict iconic gesture use among monolinguals and bilinguals. *Bilingualism: Language and Cognition, 16*, 934–944.

So, W. C. (2010). Cross-cultural transfer in gesture frequency in Chinese–English bilinguals. *Language and Cognitive Processes, 25*, 1335–1353.

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.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*, 661–699.

Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics, 36*, 219–238.

Turkstra, L. S. (2005). Looking while listening and speaking: Eye-to-face gaze in adolescents with and without traumatic brain injury. *Journal of Speech, Language, and Hearing Research, 48,* 1429–1441.

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

Wheeler, P. (2019). *The effect of vowel accuracy, visual speech, and iconic gesture on intelligibility.* Unpublished master's thesis, University College London, UCL Institute of Education. Retrieved from

https://www.teachingenglish.org.uk/sites/teacheng/files/Page%20Wheeler_University%20College%20London.pdf

Zheng, Y. I., & Samuel, A. G. (2019). How much do visual cues help listeners in perceiving accented speech? *Applied Psycholinguistics*, *40*, 93–109.

Appendix A

Instructions for the Close-Call Narrative Task

Close-Call Story

Tell about a close call or a near-miss incident that happened to you in as much detail as you can.

A close call is something that happened where you were almost hurt, or something scary or bad almost happened, but in the end everything turned out okay.

Make sure that you tell something that you are comfortable sharing.

To give you some ideas, people have told stories about skiing accidents, nearly losing a term paper on the computer, or getting lost.

Don’t hesitate to ask questions or exchange comments while your partner is sharing their story.

Appendix B

A Screenshot of the Rating Interface