Electrophysiological signatures of second language multimodal comprehension

Ye Zhang (y.zhang.16@ucl.ac.uk)
Department of Experimental Psychology
University College London, UK

Rong Ding (rong.ding@mpi.nl)
Language and Computation in Neural Systems Group,
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Diego Frassinelli (diego.frassinelli@uni-konstanz.de)
Department of Linguistics,
University of Konstanz, Konstanz, Germany

Jyrki Tuomainen (j.tuomainen@ucl.ac.uk)
Department of Speech, Hearing and Phonetic Science,
University College London, UK

Sebastian Klavinskis-Whiting (sebastian.klavinskis-whiting@chch.ox.ac.uk)
Experimental Psychology,
Oxford University, UK

Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)
Department of Experimental Psychology,
University College London, UK

Abstract

Language is multimodal: non-linguistic cues, such as prosody, gestures and mouth movements, are always present in face-to-face communication and interact to support processing. In this paper, we ask whether and how multimodal cues affect L2 processing by recording EEG for highly proficient bilinguals when watching naturalistic materials. For each word, we quantified surprisal and the informativeness of prosody, gestures, and mouth movements. We found that each cue modulates the N400: prosodic accentuation, meaningful gestures, and informative mouth movements all reduce N400. Further, effects of meaningful gestures but not mouth informativeness are enhanced by prosodic accentuation, whereas effects of mouth are enhanced by meaningful gestures but reduced by beat gestures. Compared with L1, L2 participants benefit less from cues and their interactions, except for meaningful gestures and mouth movements. Thus, in real-world language comprehension, L2 comprehenders use multimodal cues just as L1 speakers albeit to a lesser extent.

Keywords: multimodal communication; language; N400; gesture; mouth; prosodic accentuation, bilingualism, L2

Introduction

In face-to-face communication, spoken words are always accompanied by multimodal information, such as prosodic accentuations, gestures, and mouth movements.

Evidence from behavioural, electrophysiological, and neuroimaging studies suggested that these individual cues modulate L1 comprehension. Prosodic accentuation (i.e., prosodic stress characterized as higher pitch, higher amplitude and longer duration) has been found to facilitate word comprehension by making specific words more prominent (e.g., Cutler et al., 1997; Li & Ren, 2012; Kristensen et al., 2013). Meaningful gestures directly provide semantic information and facilitate semantic processing (e.g., Holle & Gunter, 2007; Kelly et al., 1999; Skipper, 2014); whereas, beat gestures (i.e., rhythmic hand movement with no direct meaning; McNeill, 1992) make words more prominent (Krahmer & Swerts, 2007; Wang & Chu, 2013; Hubbard et al., 2009). Mouth movements mainly provide sensory information (e.g., Pilling, 2009; Sumby & Pollack, 1954), although some studies found facilitatory effects at a semantic level (Brunellière et al., 2013; Hernández-Gutiérrez et al., 2018).

A recent study by Zhang and colleagues (2020) with native English speakers investigated the pattern of interaction between multimodal cues when they co-occur in naturalistic contexts. In two EEG studies, participants watched videos of an actress producing naturalistic passages (taken from the BNC and BBC programmes). Zhang et al. (2020) quantified linguistic surprisal per each word, and the informativeness of each cue. They established that words' surprisal (based on prior linguistic context) predicts N400 effects and they assessed how different multimodal cues modulate this effect. They found that multimodal cues always modulate N400: presence of prosodic accentuation and meaningful gestures reduce the N400 but beat gestures enhance it. Moreover, the

N400 modulation is dynamic and actively depending on the co-occurring cues: prosodic accentuation enhances the facilitatory effect of meaningful gestures (indexed by a N400 reduction); while the facilitatory effect of informative mouth movements is only observed when gestures (both meaningful and beats) are present.

Multimodal Cues in L2 Speakers

How do L2 speakers process multimodal cues? Two previous studies compared verbal reports after watching audio/audiovisual stimuli, showing that multimodal information is in general used by participants (Gruba, 2004; Seo, 2002). Other studies focused on individual cues (either gestures, prosody or mouth) and reported processing differences in L2 and L1. Behavioural and eye-tracking studies suggested that although both L1 and L2 participants respond to prosodic accentuation (e.g., faster phoneme detection, Akker & Cutler, 2003), L2 participants may be less capable of mapping prosodic with semantic information (Akker & Cutler, 2003; Perdomo & Kaan, 2019; Lee et al., 2019). Similarly, mouth movements improve language perception in both L1 and L2 (Drijvers & Özyürek, 2019; Navarra & Soto-Faraco, 2007), but while L2 users look more at the speakers' mouth than L1 users (Birulés et al., 2020), their behavioural improvement in terms of word recognition is smaller (Drijvers, Vaitonytė, et al., 2019; Drijvers & Özyürek, 2019). Some studies suggested that meaningful gestures improve word recognition to a smaller extent in L2 than in L1 (Drijvers, Vaitonytė, et al., 2019; Drijvers & Özyürek, 2019), while other studies suggested that L2 users look more at the hands compared with L1 (Drijvers, Vaitonytė, et al., 2019), and meaningful gestures improve L2 comprehension to a larger extent than L1 (Dahl & Ludvigsen, 2014), especially for low proficiency participants (Sueyoshi & Hardison, 2005). Beat gestures facilitate the learning of L2 words, and the effect is larger when prosodic accentuation is present (Kushch et al., 2018). However, this effect is modulated by the naturalness of the gesture (Rohrer et al., 2020). Taken together, these results suggest that L2 comprehenders benefit from each multimodal cue in comprehension, but not as much as L1 speakers do, possibly because L2 comprehension is computationally more demanding thus resulting in insufficient cognitive resources (e.g., Hopp, 2010) or because they are simply less familiar with the cues in L2 (e.g., Ortega-Llebaria and Colantoni, 2014). Further, meaningful gestures and mouth movement are more clearly linked to facilitatory effect than other cues (e.g., Birulés et al., 2020, Dahl & Ludvigsen, 2014, Sueyoshi & Hardison, 2005).

Current Study

Here, we present an electrophysiological study of how L2 Mandarin-English speakers process naturalistic-style audiovisual materials, including a comparison of the results obtained (L2 speakers) with data from a group of L1 speakers tested on the same materials. In contrast to most previous work, we do not isolate single cues, rather we investigate how the different cues are processed and how they interact in naturalistic-style materials. Non-native English speakers watched videos of an actress producing passages chosen from TV scripts (same material as Zhang et al., 2020 Exp.2) while their EEG was recorded.

Based on Zhang et al., (2020), we predicted that prosodic accentuation, meaningful gesture, and mouth movement would make words easier to process, indexed by smaller N400, while beat gestures would make words stand out more. Based on previous literature, we further predict that overall L2 participants would show a smaller N400 change than L1 users.

Methods

The stimuli, procedure and part of the analyses are identical to Zhang et al. (2020), Exp 2 (see https://www.biorxiv.org/content/10.1101/2020.01.08.896712v3).

Participants

Twenty (16 female, aged 18-40) students were recruited from University College London. All participants are highly proficient L2 English speakers (Mandarin-English; >7.5/9 in IELTS listening tests; >2 years in English-speaking country; use English daily). All participants had normal hearing, vision, and no known neurological disorder. Participants gave written consent approved by the local ethics committee and were paid £7.5/hour for participation.

Materials

Materials were the same as in Zhang et al. (2020) (see Figure 1 for an example). In the original study, to better approximate real-life language use, 79 passages were chosen from BBC script library (https://www.bbc.co.uk/writersroom/scripts). A native English-speaking actress produced them with natural prosody and facial expression. Two versions (one with and another without gestures) were created, resulting in 158 video clips (duration 10s-34s). A comparison of the same word across with/without gesture videos was performed to avoid confounds in "gesturability" if comparing different words (due to semantic differences between more and less gesturable words; e.g., *combing* v.s. *pleasing*). Four additional passages were used as practice trails.

Participants of the EEG study rated the difficulty of each passage after the experiment on a 1-5 scale. The average difficulty score of the 79 passages was not significantly different across L1 and L2 participants (L1: M=2.53, S.D.=.53; L2: M=2.58, S.D.=.76; paired-sample t-test p=0.46), with all values staying within ±3S.D. Therefore, all the 79 passages were included in further analyses.

Procedures

Participants sat ~1m facing a computer and wearing earphones. After practice trials, participants were presented with 79 video clips (gesture/no-gesture was randomized and counterbalanced across participants). Videos were displayed

with an intertrial interval of 1000ms. Forty videos were followed by yes/no questions to ensure that participants paid attention to the stimuli (mean accuracy=0.82, p<.001 in one sample t-test comparing against chance level). Participants were instructed to watch the videos carefully and answer as quickly and accurately as possible. The whole EEG experimental session took ~60 mins.

Quantification of Cues

For each video, we annotated the onset and offset of each word (mean duration=508ms, SD=306), and then quantified the informativeness of each cue per content word as below.

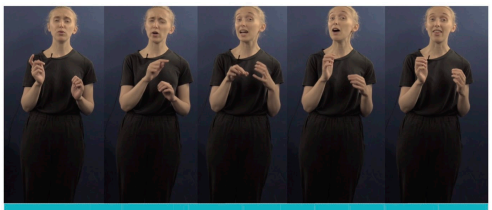
Surprisal (mean surprisal=8.17, SD=1.92) was obtained using a bigram language model. The surprisal of each word in the passages was computed based on previous content words using the following formula:

$$\text{Surprisal}(w_{t+1}) = -\log P(w_{t+1}|w_{1\ldots t})$$

Prosodic accentuation (mean F0=288Hz, SD=88) was quantified as the mean F0 per word, extracted using Praat.

Gestures were coded as meaningful gestures or beats by two expert coders (reliability coding was carried out by a third coder; intercoder reliability >95%, kappa>0.90, p<.001). Meaningful gestures (N=457) included iconic gestures (e.g., drawing movements for the word "drawing") and deictic gestures (e.g., pointing to the hair for "hair"). Beat gestures (N=340) comprised rhythmic hand movements without clear meaning. Each word was then linked either with a meaningful gesture (if a meaningful gesture associated with its meaning is present), a beat gesture (if a beat gesture overlaped with it) or no gesture.

Mouth informativeness (mean informativeness=0.67, SD=0.29) was quantified by Krason, Zhang & Vigliocco (in prep). Participants guessed the identity of words based on mouth movements. Then, averaged phonological distances between responses and target answer was calculated to measure the mouth informativeness.



	Mary	thought	of	joining	Clara.	They	might	pick	flowers	together.
Surprisal	-	6.73	-	9.02	-	-	-	8.63	9.41	7.40
Prosody	-	389.29	-	342.19	-	-	-	432.18	364.69	482.95
Meaningful Gesture	-	0	-	1	-	-	-	1	0	0
Beat Gesture	-	1	-	0	-	-	-	0	0	0
Mouth Movement	-	0.39	-	0.02	-	-	-	0.1	0.21	0.32

Figure 1: Stimuli and informativeness of cues.

Preprocessing of EEG data

The data was pre-processed with EEGLAB and ERPLAB running under MATLAB. EEG files were referenced to mastoids, down-sampled to 256Hz, separated into -100 to 1200ms epochs time-locked to word onset and filtered with a 0.05-100Hz band-pass filter. Artefacts (e.g., eye movements and muscle noise) were first corrected with ICA, and the remaining ones were rejected using a moving window peak-to-peak analysis and step-like artifact analysis (mean artefact rejection=8.69%, SD=14.12). Then, an additional 30Hz low-pass filter was applied to the data. Due to likely overlap between any baseline period (-100 to 0ms) and the EEG signal elicited by the previous word, we did not perform baseline correction, but instead extracted the mean EEG amplitude in this time interval and later used it as a control variable in the analysis (Frank et al., 2015).

Hierarchical Linear Modeling Analysis

First, we established the precise time window in which linguistic surprisal has an effect (following Zhang et al., 2020). We first performed hierarchical LInear MOdeling (LIMO toolbox) rather than specifying a N400 window a priori. This regression-based EEG analysis decomposes ERP signal into time-series of beta coefficient waveforms associated with each continuous variable. A variable is considered significant if its beta coefficient waveform is significantly different from zero (a flat line). We focused on the 0-1200ms time window, and carried out a one-sample t-test to compare the group level response with 0 (bootstrap set at 1000, clustering corrected against spatial and temporal multiple comparison).

Linear Mixed Effect Regression Analysis (LMER)

LMER analysis was conducted using the lme4 package. For each participant, mean ERPs in the 500-800ms time window (determined by the LIMO analysis above) were extracted from 32 electrodes for all content words and were used as dependent variables. In all the models below, we only included the words with gestures (in with gesture videos) and the corresponding words without gestures (in without gesture video) to balance the number of observations between groups.

Analysis 1: The independent variables included were: 1) main effect of log-transformed surprisal, mean F0, meaningful gesture, beat gesture and mouth movements; 2) two-way interactions between these cues; 3) three-way interactions involving surprisal and any two multimodal cues; and 4) control variables including baseline (-100 to 0ms ERP), word length, word order in the passage, passage order in the experiment, x, y and z coordinates of electrode. Frequency was omitted from the model due to multiple collinearity with surprisal. No main effect or interaction showed multicollinearity (VIF<2.4, kappa=5.63). Continuous variables were standardized and categorical variables were sum coded. We further included the highest interaction (three-way interactions between surprisal and

cues) as random slopes for participants (Barr, 2013). We did not include lemma as random intercept or other interactions as random slopes due to convergence issues.

Analysis 2: Here we compared results from L2 participants to those of L1 participants who were tested with the same materials (Zhang et al., 2020). The EEG responses within 500-800ms from the 20 L1 participants reported in Zhang et al. (2020, Exp 2) were combined with the L2 data described above. Native status and the interaction between native status and the multimodal cues were added to the LMER model presented in Analysis 1. No main effect or interaction showed multicollinearity (VIF<2.5, kappa=5.76).

Results

Hierarchical Linear Modeling Analysis

As shown in Figure 2, EEG responses for words with higher surprisal were significantly more negative in the 500-800ms time window post-stimulus, in line with the N400 in previous studies. We focused on the 500-800ms window in all following analyses.

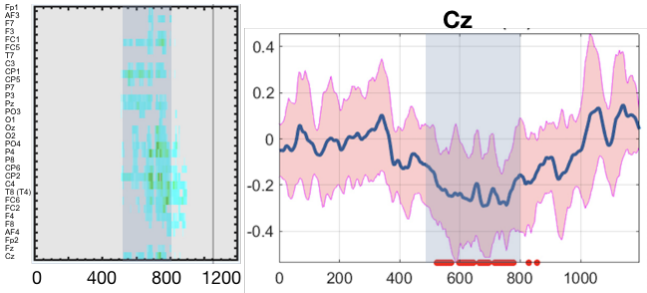


Figure 2: surprisal elicit negative ERP at 500-800ms.

Linear Mixed Effect Regression Analysis (LMER)

Below, we report only the significant effects. Full results at https://osf.io/zk47n/?view_only=e7d847fab90945c5bfd69dc1a59dc887.

Analysis 1: How do multimodal cues affect L2 processing? We found a main effect of surprisal: more surprising words induced more negative N400. Crucially, multimodal cues modulated ERP amplitude (Figure 3). We found significant positive main effects of: mean F0 (β=0.004, SE=0.002, p=.011) and mouth informativeness (β=0.007, SE=0.001, p<.001), indicating that words with higher pitch or more informative mouth movement elicited less negative N400 overall. While both informative mouth movements (β=0.010, SE=0.002, p<.001) and meaningful gestures (β=0.019, SE=0.001, p<.001) showed a positive interaction with surprisal, indicating that less predictable words showed less negative N400 when accompanied by informative mouth movements and meaningful gestures, mean F0 showed a negative interaction with surprisal (β=-0.006, SE=0.002, p<.001), indicating that less predictable words showed larger N400 with prosodic accentuation.

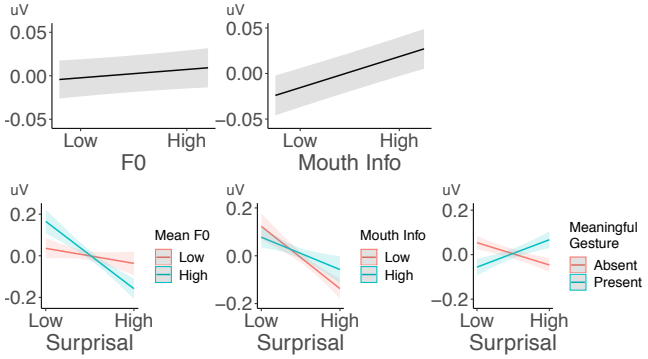


Figure 3: Multimodal cues each modulate L2 processing.

In addition, we found a number of interactions between multimodal cues (Figure 4). We found a negative interaction between F0 and mouth informativeness (β=-0.003, SE=0.001, p=.009), such that N400 was more negative for high mouth informativeness and high pitch words. Conversely, there was a positive interaction between F0 and meaningful gesture (β=0.003, SE=0.001, p=.002), mediated by an interaction with surprisal (β=0.012, SE=0.005, p=.031), indicating that meaningful gestures elicited even less negative N400 when co-occurring with high pitch, especially for high surprisal words. While the interaction between mouth and meaningful gestures was positive (β=0.004, SE=0.001, p<.001), the interaction between mouth and beat gestures was negative (β=-0.006, SE=0.001, p<.001), indicating that meaningful gestures induced less negative N400 while beat gestures induced more negative N400 for words with informative mouth movement.

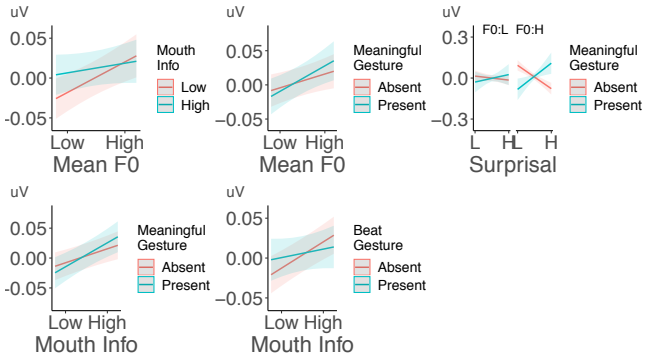


Figure 4: multimodal cues interact in L2 processing.

Analysis 2: Do multimodal cues show the same effects in L1 and L2? Overall, L2 participants showed smaller effects: surprisal had a smaller negative effect in L2 than L1 (β=-0.009, SE=0.001, p<.001); Compared with L1 users, L2 participants showed a smaller reduction of negative N400 with high pitch (β=0.004, SE=0.001, p<.001) especially for high surprisal words (β=0.003, SE=0.001, p=.007); high mouth informativeness (β=0.004, SE=0.001, p<.001); meaningful gestures (β=0.002, SE=0.001, p=.012); and a smaller negative effect of beat gestures (β=-0.006, SE=0.001, p<.001). The only exceptions were that L2 participants showed a larger reduction in N400 than L1 speakers for high

surprisal words with meaningful gestures (β=-0.008, SE=0.001, p<.001) or informative mouth movements (β=-0.007, SE=0.001, p<.001).

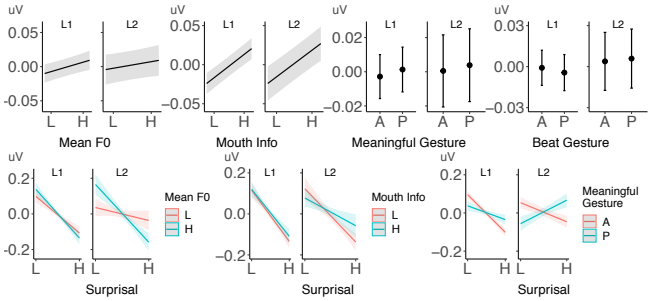


Figure 5: Effect of each cue in L1 and L2.

L2 participants were also less affected by the interaction between cues. There were a number of 3-way interactions: between native status, pitch and beat gestures (β=0.005, SE=0.001, p<.001); native status, mouth informativeness and meaningful gestures (β=0.003, SE=0.001, p<.001); native status, mouth informativeness and beat gestures (β=0.006, SE=0.001, p<.001). These all indicated that L2 users were less sensitive to the multimodal cues and their combinations. However, 4-way interactions between native status and surprisal, prosody, meaningful gestures (β=-0.009, SE=0.003, p=.014) and surprisal, mouth informativeness, meaningful gestures (β=-0.006, SE=0.003, p=0.026) indicated that L2 users benefited more than L1 users from the combination of higher pitch and meaningful gestures as well as more informative mouth movement and meaningful gestures when words are less predictable.

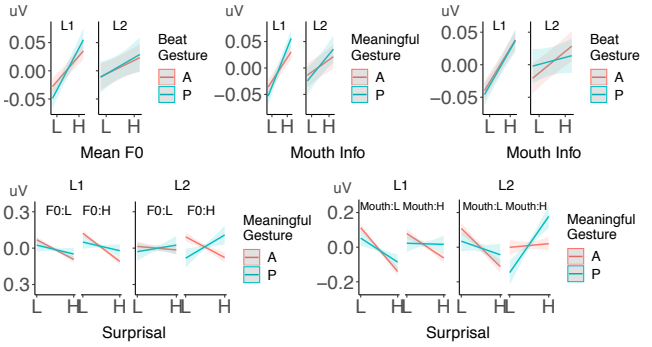


Figure 6: Effects of cue interactions in L1 and L2.

Discussion

We characterised how highly proficient L2 speakers use multimodal cues in naturalistic audio-visual comprehension. First, we established that L2 users are sensitive to linguistic predictability (surprisal). We then characterised how multimodal cues such as prosodic accentuation, gestures and mouth movements modulate linguistic processing. As predicted based on L1 performance reported in Zhang et al. (2020), we found that words with higher pitch induce less negative N400 overall but especially for more predictable words, while informative mouth movement and meaningful gestures elicit less negative N400 for less predictable words. As in L1, we further found a number of interactions among the cues: higher pitch enhances the facilitatory effect (N400 reduction) of meaningful gestures (especially for high surprisal words) but decreases the same effect for mouth movement. The co-occurrence between mouth informativeness and meaningful gestures induce less negative N400 while the co-occurrence between mouth informativeness and beat gestures induce more negative N400. Compared with L1, L2 users show overall reduced facilitatory effects of multimodal cues and their interactions, in line with previous studies. However, when words are less predictable based on their linguistic context, L2 users do benefit more than L1 from meaningful gestures (especially when co-occurring with prosodic accentuation) and informative mouth movement (especially when co-occurring with meaningful gestures).

The first main finding of our study is that different multimodal cues impact L2 processing. In line with previous behavioural studies, we found that prosodic accentuation facilitates L2 comprehension, indexed by smaller N400 (Akker & Cutler, 2003; Takahashi et al., 2018), although this effect is smaller for less predictable words. While new and less predictable words tend to be produced with accentuation (Cruttenden, 2006) and the presence of prosodic accentuation has been shown to facilitate processing of these words in L1 (e.g., Bock and Mazzella, 1983; Zhang et al., 2020), L2 users did not show the same effect. This suggest that while L2 users are sensitive to prosodic information, their ability to map it with semantic newness or predictability is limited (Akker & Cutler, 2003; Perdomo & Kaan, 2019; Lee et al., 2019). This is potentially due to the limited cognitive resources available (e.g., Hopp, 2010; Sorace, 2011) or because the users encountered problems identifying prosodic prominence in online processing (e.g., Rosenberg et al., 2010). We also found a facilitatory effect of meaningful gestures. Previous L2 studies found that incongruent meaningful gestures induced larger N400 (Drijvers & Özyürek, 2018; Ibáñez et al., 2010). In line with previous behavioural studies (Dahl & Ludvigsen, 2014; Sueyoshi & Hardison, 2005), our finding further indicated that naturally occurring congruent meaningful gestures make comprehension easier (smaller N400). This effect is especially strong for high surprisal words, suggesting that the semantic information conveyed by meaningful gestures is used when linguistic information is difficult. We report for the first time that informative mouth movements also facilitate L2 comprehension. While previous studies found that seeing the mouth leads to better recognition of words in noise (Drijvers, Vaitonytė, et al., 2019; Drijvers & Özyürek, 2019), we show that mouth movement can also improve comprehension of clear speech, possibly by enhancing the recognizability of words.

Because we did not manipulate any multimodal cue but we investigated them in their natural context, our study also allows us to assess how these multimodal cues interact. Prosodic accentuation enhances the facilitatory effect of meaningful gestures (especially for less predictable words). This may come about because higher pitch enhances attention

to other co-present cues, or because of "local" binding of the cues that can arise as accentuation often co-occur with gestures (Holler & Levinson, 2019). Interestingly, while co-occurrence of meaningful gestures and more informative mouth movement induces less negative N400, co-occurrence between beat gestures and informative mouth movements induces more negative N400. It is possible that the presence of hand movements draws participants' visual attention away from the mouth. While this shift of attention can yield additional semantic information when the gestures are meaningful, if the gestures are beat, the processing is more difficult.

Compared with L1 users, L2 users show smaller effects of the multimodal cues (in isolation and in combination), in line with previous studies (Akker & Cutler, 2003; Drijvers & Özyürek, 2019). Coupling of multimodal cues sometimes induces even larger N400 (e.g., co-occurrence of mouth and beat), indicating that multimodal communication in L2 may be more easily penalised, potentially because L2 users are less capable of accessing and integrating multimodal information. This may be associated with their cognitive resources being more limited due to the computationally demanding nature of the L2 processing (e.g., Hopp, 2010). Alternatively, they may be less familiar with the naturally occurring pattern of cues in a non-native language.

On the other hand, when words are less predictable based on linguistic information only, L2 users benefit more than L1 users from some multimodal cues (namely meaningful gesture, especially when prosodically stressed, or informative mouth movement, especially when co-occurring with meaningful gestures). In comparison with prosodic accentuation or beat gestures (both showing smaller effect in L2 than L1), meaningful gestures and mouth movements provide semantic or sensory information that is independent from linguistic input, and thus can be especially helpful for L2 users when linguistic information is hard and less predictable. It is possible that L2 users are capable of regulating their attentional resources in online processing, and pay more attention to informative multimodal cues to compensate for their relatively lower linguistic proficiency. Indeed, L2 users are more likely to look at the hands (Drijvers, Vaitonytė, et al., 2019) and mouth (Birules et al., 2020) and benefit more from meaningful gestures than L1 speakers (Dahl & Ludvigsen, 2014; Sueyoshi & Hardison, 2005). Note that previous studies reporting smaller gestural and mouth enhancement in L2 were mostly measuring single word recognition (Drijvers, Vaitonytė, et al., 2019; Drijvers & Özyürek, 2019; Drijvers, van der Plas, et al., 2019), which may not provide sufficient information for such adjustment to occur.

Our results provide key constraints to theories of L2 processing. Current theories in L2 comprehension typically focus on linguistic processing (e.g., Clahsen and Felser, 2006; Hopp, 2010; Kaan, 2014), thus cannot accommodate our findings of how L2 users actively use multimodal cues in comprehension. Some domain general theories may better capture our findings, such as Holler and Levinson's (2019) proposal that multimodal cues are bonded together and dynamically modulate language processing, or Skipper's (2015) proposal, according to which multimodal information is processed in different but partially overlapping sub-networks that constantly communicate with each other. To conclude, our study provides the first electrophysiological investigation of natural L2 processing. We characterise how multimodal cues jointly modulate L2 comprehension, and highlight those cues that can be most useful for L2 comprehenders.

Acknowledgments

The work reported here was supported by a European Research Council Advanced Grant (ECOLANG, 743035) and Royal Society Wolfson Research Merit Award (WRM\R3\170016) to GV, as well as an Experimental Psychology Society Undergraduate Research Bursary to SK. The authors declare no competing financial interests.

References

Akker, E., & Cutler, A. (2003). Prosodic cues to semantic structure in native and nonnative listening. *Bilingualism: Language and Cognition*, *6*(2), 81–96.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*.

Birulés, J., Bosch, L., Pons, F., & Lewkowicz, D. J. (2020). Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech. *Language, Cognition and Neuroscience*, *35*(10), 1314–1325.

Brunellière, A., Sánchez-García, C., Ikumi, N., & Soto-Faraco, S. (2013). Visual information constrains early and late stages of spoken-word recognition in sentence context. *International Journal of Psychophysiology*, *89*(1), 136–147.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, *40*(2), 141–201.

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.

Drijvers, L., & Özyürek, A. (2018). Native language status of the listener modulates the neural integration of speech and iconic gestures in clear and adverse listening conditions. *Brain and Language*, *177–178*, 7–17.

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*, 0023830919831311.

Drijvers, L., Vaitonytė, J., & Özyürek, A. (2019). Degree of Language Experience Modulates Visual Attention to Visible Speech and Iconic Gestures During Clear and Degraded Speech Comprehension. *Cognitive Science*, *43*(10), e12789.

Drijvers, L., van der Plas, M., Özyürek, A., & Jensen, O. (2019). Native and non-native listeners show similar yet distinct oscillatory dynamics when using gestures to access speech in noise. *NeuroImage*, *194*, 55–67.

Gruba, P. (2004). Understanding Digitized Second Language Videotext. *Computer Assisted Language Learning*, *17*(1), 51–82.

Hernández-Gutiérrez, D., Abdel Rahman, R., Martín-Loeches, M., Muñoz, F., Schacht, A., & Sommer, W. (2018). Does dynamic information about the speaker's face contribute to semantic speech processing? ERP evidence. *Cortex*, *104*, 12–25.

Holle, H., & Gunter, T. C. (2007). The Role of Iconic Gestures in Speech Disambiguation: ERP Evidence. *Journal of Cognitive Neuroscience*, *19*(7), 1175–1192.

Holler, J., & Levinson, S. C. (2019). Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences*, *23*(8), 639–652.

Hubbard, A. L., Wilson, S. M., Callan, D. E., & Dapretto, M. (2009). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping*, *30*(3), 1028–1037.

Ibáñez, A., Manes, F., Escobar, J., Trujillo, N., Andreucci, P., & Hurtado, E. (2010). Gesture influences the processing of figurative language in non-native speakers: ERP evidence. *Neuroscience Letters*, *471*(1), 48–52.

Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, *40*(4), 577–592.

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*(3), 396–414.

Krason, A., Zhang, Y., & Vigliocco, G., (in prep). Mouth informativeness norms for 1,743 English words.

Kristensen, L. B., Wang, L., Petersson, K. M., & Hagoort, P. (2013). The Interface Between Language and Attention: Prosodic Focus Marking Recruits a General Attention Network in Spoken Language Comprehension. *Cerebral Cortex*, *23*(8), 1836–1848.

Kushch, O., Igualada, A., & Prieto, P. (2018). Prominence in speech and gesture favour second language novel word learning. *Language, Cognition and Neuroscience*, *33*(8), 992–1004.

Lee, A., Perdomo, M., & Kaan, E. (2019). Native and second-language processing of contrastive pitch accent: An ERP study. *Second Language Research*, 0267658319838300.

Li, X., & Ren, G. (2012). How and when accentuation influences temporally selective attention and subsequent semantic processing during on-line spoken language comprehension: An ERP study. *Neuropsychologia*, *50*(8), 1882–1894.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

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.

Perdomo, M., & Kaan, E. (2019). Prosodic cues in second-language speech processing: A visual world eye-tracking study. *Second Language Research*, 0267658319879196.

Pilling, M. (2009). Auditory Event-Related Potentials (ERPs) in Audiovisual Speech Perception. *Journal of Speech, Language, and Hearing Research*, *52*(4), 1073–1081.

Rohrer, P., Delais-Roussarie, E., & Prieto, P. (2020). Beat Gestures for Comprehension and Recall: Differential Effects of Language Learners and Native Listeners. *Frontiers in Psychology*, *11*.

Seo, K. (2002). Research Note: The Effect of Visuals on Listening Comprehension: A Study of Japanese Learners' Listening Strategies. *International Journal of Listening*, *16*(1), 57–81.

Skipper, J. I. (2014). Echoes of the spoken past: How auditory cortex hears context during speech perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130297.

Sueyoshi, A., & Hardison, D. M. (2005). The Role of Gestures and Facial Cues in Second Language Listening Comprehension. *Language Learning*, *55*(4), 661–699.

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.

Takahashi, C., Kao, S., Baek, H., Yeung, A. H., Hwang, J., & Broselow, E. (2018). Native and non-native speaker processing and production of contrastive focus prosody. *Proceedings of the Linguistic Society of America*, *3*(1), 35-1–13.

Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: An ERP study. *Neuropsychologia*, *51*(13), 2847–2855.

Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J. I., & Vigliocco, G. (2020). More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension. *BioRxiv*.