More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension

Ye Zhang[1], Diego Frassinelli[2], Jyrki Tuomainen[3], Jeremy I Skipper[1], Gabriella Vigliocco[1]

[1]Experimental Psychology, University College London, UK
[2]Institute of Natural Language Processing, University of Stuttgart, Germany
[3]Speech, Hearing and Phonetic Sciences, University College London, UK

Abbreviated title: Electrophysiology of multimodal comprehension

Corresponding Author
Ye Zhang
Experimental Psychology
University College London
26 Bedford Way
London WC1H OAP
United Kingdom
E-mail: y.zhang.16@ucl.ac.uk

Number of pages: 30
Number of Figures: 7
Number of words: 226 abstract, 646 introduction, 1,469 discussion

Acknowledgements: 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.

The authors declare no competing financial interests.

Abstract

The natural ecology of human language is face-to-face interaction, comprising cues, like co-speech gestures, mouth movements and prosody, tightly synchronized with speech. Yet, this rich multimodal context is usually stripped away in experimental studies as the dominant paradigm focuses on speech alone. We ask how these audio-visual cues impact brain activity during naturalistic language comprehension, how they are dynamically orchestrated and whether they are organized hierarchically. We quantify each cue in video-clips of a speaker and we used a well-established electroencephalographic marker of comprehension difficulties, an event-related potential, peaking around 400ms after word-onset. We found that multimodal cues always modulated brain activity in interaction with speech, that their impact dynamically changes with their informativeness and that there is a hierarchy: prosody shows the strongest effect followed by gestures and mouth movements. Thus, this study provides a first snapshot into how the brain dynamically weights audiovisual cues in real-world language comprehension.

Introduction

Language originated, is learned and most often used in face-to-face settings. In these contexts, linguistic information such as discourse, is accompanied by other multimodal ("non-linguistic") cues like speech intonation (prosody), hand gestures and mouth movements. Behavioural, neuroimaging and electrophysiological research has shown that these cues taken individually (in experimental studies in which the other cues are controlled) can improve speech perception and language comprehension[1–4]. However, most theoretical accounts of language comprehension are grounded in studies focusing only on a single (usually linguistic) cue. This limits their ecological validity.

There is increasing evidence for prediction across distributed brain regions as a general account of brain functioning and (audiovisual) speech perception and language comprehension in particular [5–7]. Predictions matter in speech and language because they provide a constraint on interpretation of notoriously variant acoustic signals and other ambiguities at the word and higher linguistic levels. Most previous studies have addressed prediction based on prior linguistic material (e.g., prior discourse), however, multimodal cues such as prosody, gestures and mouth movements do also provide information useful in making predictions concerning upcoming words or sounds [8,9]. Yet, the mechanisms underscoring how the brain processes these cues during comprehension are not known. In particular, there are - at least - three key questions that we need to answer in order to develop comprehensive theories of natural language comprehension. First, we need to understand to what extent the processing of multimodal cues is central (or marginal) in natural language processing (e.g., whether a cue is only used when the speech is unclear, or in experimental tasks that force attention to it). Answering this question is necessary in order to properly

frame theories of natural language processing because if some multimodal cues (e.g., gesture or prosody) always contribute to processing, this would imply that our current speech-only focus is too narrow, if not misleading. Second, we need to understand the dynamics of online multimodal comprehension. In particular, to provide mechanistic accounts of language comprehension, it is necessary to establish how the weight of a certain cue dynamically changes depending upon the context (e.g., whether meaningful hand gestures are weighted more when prior linguistic context is less informative and/or when mouth movements are less informative). Finally, it is important to establish whether there is a stable hierarchical organization of cues (e.g., prior linguistic context may always be weighted more than gestures, which are in turn weighted more than mouth movements).

Prosody, gesture and mouth movements as predictors of upcoming words: the state of the art

Accentuation (i.e., prosodic stress characterized as higher pitch that makes words acoustically prominent) marks new information [10]. Many behavioural studies have revealed that comprehension is facilitated with appropriate accentuation (new information is accentuated, and old information de-accentuated [11,12]. Incongruence between the presence of prosodic accentuation and newness of information increases processing difficulty, inducing increased activation in left inferior frontal gyrus, interpreted as increased phonological and semantic processing difficulty [13]. In electrophysiological (EEG) studies, such mismatch elicits more negative N400 (an event-related-potential (ERP) peaking negatively 400ms after word presentation around central-parietal areas [14], that has been argued to mark prediction in language comprehension [2]) than appropriate accentuation [15–20].

Meaningful co-speech gestures make upcoming words more predictable. Behavioural studies have shown that they improve comprehension [21]. In line with behavioural studies, EEG studies have shown that activating the less predictable meaning of the homonymous word "ball" using a "dancing" gesture, reduces the N400 response to a later mention of "dance" [22–24]. Incongruence between meaningful gestures and linguistic context triggers more negative N400 compared with congruent gestures, suggesting that meaningful gestures are involved in semantic comprehension [25–28]. Meaningful gestures are linked to activation in temporal and inferior frontal regions, which are associated with meaning processing [9,29–32]. Moreover, the presence of meaningful gestures results in a significant reduction in cortical activity in auditory language regions (namely posterior superior temporal regions), a hallmark of prediction [33].

Fewer studies have investigated beat gestures (meaningless gestures time-locked to the speech rhythm) [34]. Some argued that beats enhance saliency of associated speech in a similar manner as prosodic accentuation [35], and activate the same regions as prosody in auditory cortex [36]. One study reported that beat gestures induce less negative N400, similar to prosodic accentuation [37]. Other EEG studies, however, reported that beat gestures modulated brain responses in a later window (around 600ms [38,39]).

Finally, while many studies have focused on the sensory mechanisms underscoring the use of mouth movements in speech, less is known about whether the informativeness of mouth movements affects word predictability. Behavioural and fMRI studies have shown only a small facilitatory effect of seeing mouth movements when meaningful gestures are also present [9,31,40]. Two electrophysiological studies, however, reported conflicting findings. While Brunellière and colleagues linked more informative mouth movements to more negative

N400 amplitude [41], generally indicating increased processing difficulty, Hernández-Gutiérrez and colleagues failed to find any N400 effect associated with mouth movements [42].

Thus, previous studies indicate that at least when taken one by one, multimodal non-linguistic cues interact with speech, modulating the predictability of upcoming words. They report, however, such interactions in controlled settings where only the investigated cues are manipulated while the others are kept constant to ensure experimental control because of the challenges of doing experimental research with naturalistic stimuli [43]. Thus, for example, prosody is normalised and auditory (rather than audiovisual) presentation is used when studying speech [44]; or only the mouth, rather than the whole body is shown when studying audiovisual speech perception [41]; or the face is hidden when studying gestures [22]. Such a reductionist approach is considered to be necessary to ensure experimental control. However, the materials and tasks used in the studies often do not reflect the conditions in which the brain processes language in real-world face-to-face contexts in which it is simply impossible not to see a person's gestures while they speak, or their mouth movements while we see their gestures. This approach breaks the natural and possibly predictive correlation among cues with unknown consequences on processing [45,46], as the disruption of the relative reliability of cues can affect whether and how much the brain relies on a given cue [47,48].

The present study

Here we address the three key questions about face-to-face multimodal communication outlined above using a design that maintains ecological validity. We asked thirty-six (31 included, mean age=27, 17 women) native English speakers to watch 100 videos in which an actress produced short passages (taken from a naturalistic corpus of British English) with natural prosody, co-speech gestures and mouth movements. One-third of the videos were

followed by yes/no questions about the content of the video to ensure participants paid attention during the experiment and to acquire behavioural responses. Participants were instructed to watch the videos carefully and to answer questions as quickly and accurately as possible. We measured the electrophysiological responses to each word produced and assessed how each cue and their interactions modulate N400 responses to each word. We use the N400 as a biological marker of processing difficulty, associated with word predictability [49]. Crucially as discussed above, prosody, gestures and mouth movements have all been shown to modulate N400 responses to words, rendering this event-related potential especially well-suited for the study of multimodal language.

We quantified the informativeness of each cue for all content words in the passages. Word-predictability was computed using surprisal, a measure of the probability that a word follows from previous words [49,50]. Prosody (prosodic accentuation) was quantified in terms of the mean fundamental frequency (F0) of the words; gestures (meaningful gestures and beats) were coded as present/absent for each word and finally mouth movements associated to each word were quantified in terms of their informativeness (i.e., how easy it is to guess the word just by looking at the mouth movements). Quantification of the different cues word-by-word allows us to address how their dynamic change impacts electrophysiological responses. Figure 1 gives an example of an annotated sentence. We analyzed the impact of these cues and their interactions on the N400 response for each word during continuous speech, using robust and fine-grained measures for surprisal, prosody, gestures and mouth movements.

Figure 1 about here

We assess whether the processing of these multimodal cues is central (or marginal) in natural language processing by measuring whether presence/informativeness of these cues modulate N400 amplitude in natural language processing. Previous work suggests that prosody and meaningful gestures will both reduce the N400 amplitude as both provide meaningful information that makes upcoming less predictable words (based on linguistic context), more predictable thus, easier to process. Here, we go beyond this by asking whether the same pattern will hold when it is not just one, but multiple cues contributing to the process. Second, we evaluate the dynamic nature of multimodal cue processing by analyzing the interaction between cues. If the weight of a certain cue dynamically changes depending upon the context, then its impact on N400 should show modulations as a function of other cues. Finally, we assess whether there is a hierarchical ranking of multimodal cues by investigating the relative magnitude of the effect of each cue as well as the extent to which a cue interacts with other cues.

Results

Behavioural Analysis

This first analysis establishes whether differences in surprisal are associated with difficulties in processing as indicated by how accurate and fast subjects were in answering the 35 comprehension questions. We used generalized mixed effect modeling and for both accuracy and response time models, mean surprisal (averaged across all content words in the video) was included as predictor variable. Participant and sentence pair were added as random intercepts to control for by-participant and by-video variation.

We found that accuracy decreased with an increase in surprisal (Mean=82.1%, SD=0.384, β=-0.784, p<.001). Similarly, we found that sentences with higher averaged surprisal had slower reaction times (Mean=4129.8 ms, SD=2881.3, β=0.089, p=.024). These findings confirm that sentences with higher surprisal were harder to process.

EEG Analyses

Time Window Sensitive to Linguistic Context

The time window in which linguistic context affects processing is an empirical question, given that no previous study has investigated the effect of surprisal in audiovisual multimodal communication. Therefore, rather than specifying a N400 window a priori, we first identified the time window where electrophysiological responses were sensitive to surprisal using hierarchical LInear MOdeling (LIMO toolbox [51]). While traditional ERP analysis compares different conditions and thus may require dichotomization of the predictor variable, this regression based ERP analysis linearly decomposes an ERP into time-series of beta coefficient waveforms elicited by continuous variables. Significant differences between the beta coefficient waveforms and zero (flat line) represent the existence of an effect [52,53].

We found that the beta values of central-parietal electrodes were significantly more negative (compared with 0) in around 300-600ms time window across electrodes (Figure 2). Words with higher surprisal, elicited more negative signals or larger N400 amplitudes. No other time window was significantly sensitive to surprisal. As a result, we focused on the 300-600 interval in our subsequent analysis.

Figure 2 about here

Modulation of word predictability by multimodal cues

After determining the time window in which surprisal has an effect, we performed linear mixed effect analysis (LMER) on the resulting time window. LMER was selected due to the advantage in accommodating both categorical and continuous variables, thus increasing statistical power [54]. Moreover, LMER can account for both by participant and item variance and can better accommodate unbalanced designs, suitable for EEG studies investigating naturalistic language processing [49,55].

Mean ERP in the 300-600ms time window was used as the dependent variable. The independent variables included surprisal, mean F0, meaningful gestures, beat gestures, mouth informativeness and all up to three-way interactions between surprisal and any other two cues, alongside other control variables (see Methods). We further included word types (lemma) and participant as random intercepts. The highest interactions (all three-way interactions) were also included as random slopes for participants [56], as was surprisal as random slope for lemma.

We focus first on the main effects of the multimodal cues and their interactions with surprisal to establish whether multimodal cues mediated the effect of the predictability of linguistic context. Full model results are reported in Table 1.

Table 1 about here

As shown in Figure 3 (panel A), we found a main effect of prosody (mean F0) (β=0.010, $p<.001$). Words produced with higher mean pitch showed less negative EEG, or smaller N400

amplitude in the 300-600 time window, compared with words produced with lower pitch. Figure 3, panel B reports the interaction between surprisal and mean F0 (β=0.017, p<.001). The larger N400 amplitude associated with high surprisal words was modulated by pitch. High surprisal words elicited a larger reduction of N400 amplitude when the pitch was higher, in comparison to low surprisal words.

Figure 3 about here

We found a similar main effect of meaningful gestures (Figure 4). Words accompanied by a meaningful gesture showed a significantly less negative N400 (β=0.006, p<.001). There was also a significant interaction between surprisal and meaningful gesture, indicating that for high surprisal words, the presence of a meaningful gesture makes the N400 less negative (β=0.008, p<.001).

Figure 4 about here

In contrast to meaningful gestures, beat gestures showed a different pattern (Figure 5). We found a significant main effect of beat gestures (β=-0.005, p=.001), suggesting that words accompanied by beats gestures elicited a more negative N400. There was a significant interaction, such that high surprisal words accompanied by beat gestures showed a further increase in negativity compared with low surprisal words (β=-0.012, p<.001).

Figure 5 about here

Dynamics and Hierarchy of Multimodal Cue Processing: Interactions among Multimodal Cues

We found a number of significant interactions between mean F0 and other multimodal cues (Figure 6). First, there was an interaction between mean F0 and meaningful gesture (β=0.004, p<.001). Words with meaningful gestures showed even less negative amplitude of N400 with increased mean F0. Second, there was an interaction between mean F0 and mouth informativeness (β=0.003, p=.040) in which words with higher mouth informativeness elicited less negative N400 when the pitch was high, but more negative N400 when the pitch was low. This interaction was further mediated by a three-way interaction between mean F0, mouth informativeness and surprisal (β=-0.013, p=.011): for words with low F0, low mouth informativeness induced more negative N400 for high surprisal words, while for words with high F0, such effect was reversed. Finally, mean F0 and beat gestures also interacted (β=-0.006, p<.001), but the direction and significance of this effect varies with different measures of prosody (see supplementary material); thus, we refrain from any further discussion of the effect.

Figure 6 about here

Figure 7 (panel A) shows the interaction between mouth informativeness and meaningful gesture (β=-0.06, p<.001). Words with meaningful gestures elicited more negative N400 when mouth informativeness was high. This two-way interaction was further mediated by a significant three-way interaction between surprisal, mouth informativeness and meaningful gesture (β=-0.010, p=.030). When words were not accompanied by meaningful gestures, high surprisal words with low mouth informativeness elicited more negative N400 compared with

high mouth informativeness words. However, when words were accompanied by meaningful gestures, high surprisal words with low mouth informativeness elicited more positive N400 (Figure 7, Panel B).

Figure 7 about here

Discussion

The present study investigated for the first time the electrophysiological correlates of real-world multimodal language processing tracking on-line processing difficulty as indexed by N400 amplitude. First, we confirmed the N400 as a biomarker of prediction during naturalistic audiovisual language comprehension: high surprisal words elicited longer reaction times and lower accuracy behaviourally, and a more negative N400 between 300 and 600ms post-stimulus, strongest in central-posterior electrodes. Crucially, our study provides a first comprehensive picture of how the brain dynamically weights audiovisual cues in real-world language comprehension and it provides first answers to the three key questions we presented in the introduction.

First, we asked whether the processing of these multimodal cues is central (or marginal) in natural language processing. These results provide first answers to the key questions introduced above. Prosodic accentuation and meaningful gestures reduced the N400 amplitude overall, especially for high surprisal words. In contrast, the presence of beat gestures increased the N400 amplitude overall, but especially for high surprisal words. Mouth movements did not modulate surprisal independently, but participated in complex interactions

involving other cues and surprisal. Thus, our results clearly show that language comprehension in its natural ecology (face-to-face communication) involves more than just speech: the predictability of words based on linguistic context is *always* modulated by the multimodal cues thus forcing a reconsideration of theoretical claims strictly based on speech only.

Second, we addressed the dynamic nature of multimodal cue processing showing how the weights given to each cue depend on which other informative cue is present at that moment in processing, as indexed by the presence of interactions between cues. We found a number of novel interactions which provide novel insight into how cue-weight changes depending upon what other cues are present. First, the facilitatory effect of meaningful gestures was reduced for words with high mouth informativeness, especially for high surprisal words. Second, the facilitatory effect of meaningful gestures was greater for accented words. Similarly, a facilitatory effect of mouth informativeness was only observed when words were accented and only for high surprisal words.

Finally, we assessed whether there is some sort of hierarchical ranking of the multimodal cues. Our results suggest that this is the case. Prosodic accentuation (providing information useful to drive attentional and semantic processes) had the most pervasive role in our study: it interacted with meaningful gestures and mouth informativeness in addition to surprisal. The N400 reduction observed for meaningful gestures and mouth informativeness was enhanced for words carrying accentuation; moreover, an N400 reduction for high surprisal words with high mouth informativeness was only observed for words carrying accentuation. This global effect is consistent with the claims by Kristensen and colleagues according to whom, prosodic accentuation engages a domain general attention network [57]. Thus, accentuation may

draw attention to other cues which consequently would be weighted more heavily. Alternatively, or additionally, as argued by Holler and Levinson, listeners are attuned to natural correlations among the cues (e.g., high pitch correlates to larger mouth movements and increased gesture size) and would use cue-bundles for prediction [8]. Meaningful gestures (providing semantic information) and beat gestures (guiding attentional processes) came next in terms of impact. Informative mouth movements (providing sensory-level information about phonetic/phonological make-up of the words) had the smallest effect in the time window investigated here.

Prosody, gesture and mouth movements as predictors of upcoming words: beyond the state of the art

In addition to providing key novel insight into the importance, dynamic engagement and hierarchical organization of the multimodal cues, our study further provides constraints and clarifications to previous studies that have investigated each cue separately.

Prosodic accentuation has been considered to mark 'newness' [1], as speakers are more likely to stress a word if it conveys new information [10]. Previous electrophysiological studies have shown that un-accented new words elicit more negative N400 [15–20]. Our findings complement previous work in showing that in multimodal contexts, presence of accentuation for less predictable words reduces the amplitude of the N400, suggesting that prosodic accentuation can enhance expectation for lower probability continuations, in line with earlier behavioural works [11,12].

We found that meaningful gestures support processing, especially for high surprisal words. This result is in line with studies that showed N400 reduction for the subordinate meaning of

ambiguous words (e.g. “ball” meaning dancing party) in the presence of a corresponding gesture [22–24], and previous work suggesting that words produced with incongruent gestures induce a larger N400 (see review in Özyürek, 2014 [4]). Our results show that gestures play a more general role in face-to-face communication: not only they ease comprehension when semantic processing is difficult (due to incongruence or ambiguity), but they also provide additional semantic information about upcoming words therefore increasing their predictability.

Crucially, *meaningful* gestures, but not beat gestures, increase the predictability of upcoming words. High surprisal words accompanied by beat gestures elicited a larger N400. This effect might be accounted for in terms of beats enhancing the saliency of a specific word [35], and highlighting its lack of fit into the previous context. Alternatively, it is possible that listeners try to extract meaning from all gestures and integrate it with speech by default, and since beats are not meaningful, integration fails, inducing processing difficulties. Wang and Chu failed to find the same effects of beat gestures within the N400 time window [37]. The reasons for the different results are unclear. However, the lack of any meaningful gestures in their study could have discouraged listeners from paying attention to gestures. Shifts in the weight attributed to different multimodal cues depending upon the specific task used are documented in the literature [22,47,48]. Importantly, the dissociation between meaningful and beat gestures further allows us to exclude the possibility that the N400 reduction observed (for meaningful gestures and for prosody) comes about because these multimodal cues share processing resources with speech processing, letting less predictable words go unnoticed.

Beats and prosodic accentuation have been argued to serve the same function in communication, namely, to make specific words more prominent and therefore attract

attention to them [35]. Our results provide evidence against such a claim as their electrophysiological correlates dissociated: beat gestures elicited more negative N400 especially for high surprisal words, while prosodic accentuation elicited less negative N400, especially for high surprisal words (see also Wang and Chu, 2013 [37]). Thus, our work supports the view that only prosodic accentuation is used as a marker of information status supporting prediction for new words.

We did not find any significant main effect of mouth informativeness or any interactions between mouth and surprisal in the N400 time-window. Mouth movements have long been recognized to facilitate speech perception especially in noise [58], and synchronized audiovisual compared with audio only speech showed reduction of N1/P2 amplitude, indicating easier sensory-level processing [59,60]. However, in our study we focused on 300-600ms after word onset in order to capture the effect of surprisal and we did not consider earlier (100-300ms) time windows. Two previous studies have investigated the impact of mouth movements within the N400 time window. Hernández-Gutiérrez and colleagues did not find any N400 difference between audiovisual and audio-only speech [42]; while Brunellière and colleagues found an increase in N400 amplitude for more informative mouth movements [41]. Further research is necessary to clarify these discrepancies, however, our results suggest that mouth informativeness can affect processing in the N400 time window but only in combination with other cues when presented in multimodal context.

Finally, our results further extend the previous literature by showing that when multiple cues from the same (visual) channel - such as meaningful gestures and informative mouth movements - are present, they can compete for attentional resources such that the facilitatory

effect of representational gestures for high surprisal words is reduced when mouth movements are informative.

Toward a neurobiology of natural communication

Our result calls for a new neurobiological model of natural language use that accounts for the effects of multimodal cues on language comprehension, as well as the interactions within multiple multimodal cues. In probabilistic-based predictive accounts, the N400 is taken as an index of the processing demands associated with low predictability [2]. It has been argued that prior to the bottom-up information, a comprehender holds a distribution of probabilistic hypotheses of the upcoming input constructed by combining his/her probabilistic knowledge of events with contextual information. This distribution is updated with new information, and consequently becomes the new prior distribution for the next event. Thus, the N400 is linked to the process of updating the distribution of hypotheses: smaller N400 is associated with more accurate prior distributions/predictions [2]. Our work shows that these mechanisms do not operate only on linguistic information but crucially, they weight in 'non-linguistic' multimodal cues. Prosodic accentuation marks low predictability of the upcoming words, thus more attention and larger weights would be assigned to other cues at both semantic (meaningful gestures) and sensory (mouth movement) levels. Meaningful gestures, could directly impact the prior distribution for the next word (see also discussion in Holler and Levinson, 2019 [8]).

In terms of neuroanatomical models, those in which language comprehension is considered in context and associated with many interconnected networks distributed throughout the whole brain [45,46] can, in principle, accommodate the results reported here. For example, in the Natural Organization of Language and Brain (NOLB) model, each multimodal cue is

proposed to be processed in different but partially overlapping sub-networks [46]. Indeed, different sub-networks have been associated with gestures and mouth movements, with a 'gesture network' being weighted more strongly than a 'mouth network' when gestures are present [9,31]. These distributed sub-networks are assumed to actively predict and provide constraints on possible interpretations of the acoustic signal, thus enabling fast and accurate comprehension [31]. Our finding of multiple interactions between cues is compatible with this view, thus suggesting that multimodal prediction processes are dynamic, re-weighting each cue based on the status of other cues.

Conclusions

To conclude, our study investigated language processing in its naturalistic multimodal environment for the first time, and provided novel evidence that, first, multimodal "non-linguistic' cues have a central role in processing as they always modulate predictions on what is going to be said next; second, they dynamically interact among one another and with linguistic cues to construct these predictions, and finally, cues are not equal but are organised in a hierarchical manner. More generally, our study provides a new, more ecologically valid, way to understand the neurobiology of language, in which multimodal cues are dynamically orchestrated.

Methods

Participants

Thirty-six native English speakers with normal hearing and normal or corrected to normal vision were paid £8 to participate in the present study after giving written consent. Five participants were excluded, three due to technical issues, one for falling asleep, and one for

excessive muscle noise, leaving thirty-one participants. The experimental procedure was approved by the local ethics committee of the university.

Material

Two-hundred and forty-six naturalistic sentence pairs (two consecutive sentences) were extracted from the British National Corpus (BNC, http://www.natcorp.ox.ac.uk/)[61]. Sentences were selected in a semi-random fashion with the only constraints the second sentence had to be at least five words long, and contain at least one verb that could be easily gestured (e.g. turn the pages). If necessary, we edited slightly the first sentence to facilitate readability and resolved all ambiguities (e.g. proper nouns without a clear reference were changed into pronouns), while the second sentence was kept unmodified. Twelve native English speakers were paid £2 each to evaluate the sentence pairs for grammaticality, meaningfulness and gesturability on a 1-5 scale. We selected 103 sentence pairs that had averaged gesturability > 2 (and SD < 2.5); and had no grammatical errors or semantic anomalies. Three sentence pairs were used for practice, and 100 were used as stimuli (Mean gesturability=2.67, SD=0.58).

A native British English-speaking actress produced the 103 sentence pairs. She stood in front of a dark-blue background, wearing black T-shirt and trousers to keep her arms and hands visible, and did not wear glasses to keep her face visible. She was instructed to read out the sentences presented behind the camera at a natural speed, with natural prosody and facial expressions. Each sentence pair was recorded with and without gestures. For videos with gestures, the actress was instructed to gesture as she naturally would. For videos without gestures, she was asked to stand still keeping her arms along her body.

Quantification of Cues

The onset and offset of each word were automatically detected using a word-phoneme aligner based on Hidden Markov Models [62]. The timing was then checked and corrected manually if needed. The mean word duration was 440.32ms (SD=375.69ms). Next, for each content word (i.e., nouns, adjectives, verbs and adverbs) we quantified the informativeness of different multimodal cues. We did not quantify measures of informativeness for function words (i.e., articles, pronouns, auxiliary verbs and prepositions) because Frank and colleagues failed to show any effect of the predictability (measured as surprisal) for such words [49]. Linguistic predictability was measured using surprisal (Mean surprisal=7.92, SD=2.10), defined as the negative log likelihood of the probability of a word to follow a sequence of other words [63]. Previous work has shown that surprisal provides a good measure of predictability, especially for low predictability words [50] and predicts reading times [50,64] and N400 amplitude [49]. Here, surprisal was generated using a bigram language model trained on the lemmatized version of the first slice (~19-million tokens) of the ENCOW14-A corpus (https://corporafromtheweb.org/encow14/), an English web corpus [65]. We chose a bigram model to reduce data sparsity and, consequently, increase the robustness of our surprisal measures. Moreover, Frank and colleagues showed that bigram models perform equally well, if not better than more complex models - trigram, recurrent neural networks (RNN) and probabilistic phrase-structure grammar (PSG) - in fitting N400 data [49]. Once trained, the bigram model was used to calculate the surprisal of each word in the sentence pairs based on previous content words in the two sentences using the following formula:

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

where w_{t+1} indicates the current word, and $w_{1...t}$ stands for previous content words in the two sentences. We also developed models in which the number of content words used in computing surprisal was varied. Given the minor differences observed, we decided to include all previous content words in the two-sentence stimuli (from the first word in the first sentence to the word preceding the target word, in the supplementary material (section 1), we show results for different window sizes to justify this choice).

Prosodic information for each word was quantified as the mean F0 of the word (mean F0=298.39Hz, SD=84.19Hz). We automatically extracted mean F0, maximum F0, minimum F0, mean intensity and F0 change per word using Praat (version 6.0.29, http://www.praat.org/) [66]. A comparison of results obtained with these different pitch measurements showed that they are similar (see supplementary material), therefore we chose to report results for mean F0 because it has been used most often as a measure of prosody in previous work [67].

Gestures were coded in ELAN (version 5.0.0, https://tla.mpi.nl/tools/tla-tools/elan/)[68] as meaningful gestures or beats by an expert coder. Coding was checked for reliability by asking a second expert coder to annotate 10% of the productions, resulting in an interrater agreement of 89.6% (kappa=0.802, p<.001). Meaningful gestures (N=359) comprised iconic gestures (e.g. drawing movements for the word "drawing") and deictic gestures (e.g. pointing to the hair for "hair"). Beat gestures (N=229) comprised rhythmic movements of the hands without clear meaning [34], but regarded to enhance the salience of the speech [35].

Mouth informativeness was quantified per word in a separate online experiment. The video-clips (from the 'without gesture' condition, as mouth movements are occasionally hidden by the hand in the "with gesture" condition) were muted, cropped to leave only the face and

segmented at the word level. The resulting video clips were presented using the Gorilla Experiment Builder (https://gorilla.sc) to 95 native English speaking subjects (Mean age =25, 54 Females) who did not participate in the main experiment. Participants were recruited through Prolific Academic (https://www.prolific.ac) and were paid £2 for participation. Each video clip was presented at the top of the screen with four words presented at the bottom, including the target word and three foils randomly selected from all words in the experiment with the same length as the target. Participants were instructed to select the word matching the video. They were allowed to watch the video for as many times as they needed before making their choice. The four words were not shown until after the participant played the video. Each word received 10 responses, and we calculated the mouth informativeness of each word using its mean accuracy (mean=0.77, SD=0.42) divided by the average number of times the clip was played (mean=2.83, SD=2.52; mean informativeness=0.32, SD=0.15).

Procedure

After the three sentence-pairs given as practice trials, each participant was presented with 50 gesture and 50 no gesture videos in randomized order using Presentation software (V. 18.0, www.neurobs.com). Each sentence pair was separated by a 2000ms interval. The experiment was counterbalanced across every two participants so that they watched the videos in the same order but with counterbalanced gesture/no-gesture conditions. One-third of the videos (35) were followed by yes/no questions about the content of the video to ensure participants paid attention during the experiment and to acquire behavioural responses. For example, for the video “Emma screamed and swore at them. She was especially angry if the girls dared to eat any of her food or drink her coffee”, the question was “Is Emma going to share her sweets with the other girls?”. Participants sat comfortably one meter away from the screen with a

resolution of 1024*768, wearing 50Ω headphones, and were instructed to watch the videos carefully and to answer questions as quickly and accurately as possible (prioritizing accuracy) by pressing the left ("Yes") or right ("No") control key. Participants were asked to avoid moving, keep their facial muscles relaxed and reduce blinking, but they were also told that it is better to blink occasionally than to avoid blinking because of potential discomfort due to e.g., drying of the eyes. Similar instructions were written on the screen. The recording took thirty minutes with three breaks.

EEG Recording

A 32-channel BioSimi system with silver-silver chloride electrodes and 24 bit resolution was used for the EEG recording, following a 10-10 international system layout. A common reference included the CMS electrode (serving as the online reference) and DRL electrode (serving as the ground electrode). Elastic head caps were used to keep the electrodes in place. Two external electrodes were attached under the left and right mastoids for off-line reference, while two other external eye electrodes were attached below the left eye and on the right canthus to detect blinks and eye movements. Electrolyte gel was inserted on each electrode to improve connectivity between the skull and the electrode. To check for relative impedance differences, the electrode offsets were kept between +/-25mV. The recording was carried out in a shielded room with the temperature kept at 18 °C.

Behavioural Analysis

We used generalized mixed effect modeling to test whether surprisal had an effect on the accuracy and response time for questions following 35 sentences. The analysis was conducted using LME4 [69] package running under R Studio (version 3.4.1, http://www.rstudio.com/). We used logistic regression in the accuracy analysis and linear regression

for response time. In both, mean surprisal (averaged across all content words in the video) was the predictor variable, participant and sentence pair were added as random intercept to control for by participant and by video variation. All continuous variables (response time and surprisal) per sentence were standardized (centered and scaled) using the "scale" function built in R so that each coefficient represents the effect size of the variable.

EEG pre-processing

The raw data were pre-processed with EEGLAB (version 14.1.1)[70] and ERPLAB (version 7.0.0)[71] running under MATLAB (R2017b, https://www.mathworks.com/products/matlab.html). All electrodes were included. Triggers were sent at the onset of each video, and word onset was subsequently calculated from the word boundary annotation. Any lag between trigger and stimuli presentation was also measured and corrected (Mean=0.21s, SD=0.07). The EEG file was re-referenced to average of the left and right mastoids (M1 and M2), down-sampled from 2048Hz to 256Hz to speed up preprocessing, and separated into epochs each containing data from -100 to 924ms around word onset [49]. The data was filtered with a second order Butterworth 0.05-100Hz band-pass filter. Due to the 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 regression analysis [49]. We conducted independent component analysis based artifact correction (ICA). Two independent experts manually labelled eye movement and other noise (e.g. heart beat, line noise) components that were subsequently removed from the data. Further artifact rejection was conducted by first using a moving window peak-to-peak analysis (Voltage Threshold=100 µV, moving window full width=200 ms, window step=20 ms) and then step-like artifact analysis (Voltage

Threshold=35 μV, moving window full width=400 ms, window step=10 ms). This resulted in an average rejection of 12.43% (SD=12.49) of the data. The ERP files were then computed from pre-processed data files, and were additionally filtered with a 30Hz low-pass filter.

EEG Analysis: Hierarchical Linear Modeling

We used the LIMO (hierarchical LInear MOdeling) toolbox [51] working under MATLAB (R2017b, https://www.mathworks.com/products/matlab.html). For each participant, we created a single-trial file from the EEG file, and a continuous variable containing surprisal of each word. In the first level analysis for each participant, a regression was performed for each data point in 0-924ms time window per electrode per word, with EEG voltage as the dependent variable and word surprisal as the independent variable, thus generating a matrix of beta values, which indicate whether and when surprisal has an effect for each participant. In the second level analysis across all participants, the averaged beta matrix was compared with 0 using a one-sample t-test (bootstrap set at 1000, clustering corrected against spatial and temporal multiple comparison) [72].

EEG Analysis: Linear Mixed Effect Regression Analysis

After determining the time window where surprisal has an effect, we performed linear mixed effect analysis (LMER) on the resulting time window. We excluded from analyses: (a) all function words, modal words, and proper names; (b) words without a surprisal value (26 words, due to the lack of occurrence of the combination between the word and its context in the corpus); (c) words without a mean F0 score (4 words, due to insufficient data points when calculating the average); (d) words associated with both beat and meaningful gestures (3 words); (e) words occurring without any gesture in the “with gesture” condition, and the corresponding words in without gesture videos. This was done to avoid data unbalance as

there were three times more words with no gestures (combining the videos with and without gestures). Mean ERP in the 300-600ms time window (as determined in the prior hierarchical linear modeling step) was extracted from 32 scalp electrodes for each word and was used as the dependent variable. Mean ERP in the -100 to 0ms time window was extracted as the baseline. The independent variables included 1) predictors: log-transformed surprisal, mean F0, meaningful gestures, beat gestures, mouth informativeness, and all up to three-way interactions between surprisal and any two cues, excluding interactions containing meaningful gesture*beat gestures (as the three instances were removed from the data), 2) control variables: baseline extracted between -100 to 0ms, word frequency, word length, word order in the sentence, sentence order in experiment, relative position of each electrode measured by its X, Y and Z coordinate position [73] acquired from BioSemi website (https://www.biosemi.com/download.htm). No main or interaction effects showed multicollinearity, with variance inflation factor (VIF) less than 2, kappa=4.871. All continuous variables, including ERP, surprisal, mean F0, mouth informativeness, baseline, frequency, word length, word order, sentence order and X, Y, Z position of electrodes were standardized (centered and scaled) so that each coefficient represents the effect size of the variable. Surprisal and frequency were log transformed to normalize the data. All categorical variables were sum coded so that each coefficient represents the size of the contrast from the given predictor value compared with the grand mean (intercept) [55].

We further included word types and participant as random intercept in the random structure. We attempted to construct a maximal random structure by entering all main and interactions as a random slope of participants, but the model failed to converge. As a result, we included the highest interaction (three-way interactions) as random slope for participants [56], and

surprisal was included as random slope for lemma. Our analysis included 31 participants, 381 word type lemmas and 480,212 data points.

References

1. Cutler, A., Dahan, D. & van Donselaar, W. Prosody in the Comprehension of Spoken Language: A Literature Review. *Lang. Speech* **40**, 141–201 (1997).
2. Kuperberg, G. R. & Jaeger, T. F. What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* **31**, 32–59 (2016).
3. Massaro, D. W. & Jesse, A. Audiovisual speech perception and word recognition. *Oxf. Handb. Psycholinguist.* (2007) doi:10.1093/oxfordhb/9780198568971.013.0002.
4. Özyürek, A. Hearing and seeing meaning in speech and gesture: insights from brain and behaviour. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130296 (2014).
5. Arnal, L. H., Wyart, V. & Giraud, A.-L. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nat. Neurosci.* **14**, 797–801 (2011).
6. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
7. Friston, K. & Kiebel, S. Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. B Biol. Sci.* **364**, 1211–1221 (2009).
8. Holler, J. & Levinson, S. C. Multimodal Language Processing in Human Communication. *Trends Cogn. Sci.* **23**, 639–652 (2019).
9. Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C. & Small, S. L. Gestures Orchestrate Brain Networks for Language Understanding. *Curr. Biol.* **19**, 661–667 (2009).
10. Cruttenden, A. The de-accenting of given information: A cognitive universal. in *Pragmatic Organization of Discourse in the Languages of Europe* 311–355 (Walter de Gruyter, 2006).
11. Bock, J. K. & Mazzella, J. R. Intonational marking of given and new information: Some consequences for comprehension. *Mem. Cognit.* **11**, 64–76 (1983).

12. Terken, J. & Nooteboom, S. G. Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Lang. Cogn. Process.* **2**, 145–163 (1987).

13. van Leeuwen, T. M. *et al.* Phonological markers of information structure: An fMRI study. *Neuropsychologia* **58**, 64–74 (2014).

14. Kutas, M. & Federmeier, K. D. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annu. Rev. Psychol.* **62**, 621–647 (2011).

15. Baumann, S. & Schumacher, P. B. (De-)Accentuation and the Processing of Information Status: Evidence from Event-Related Brain Potentials. *Lang. Speech* **55**, 361–381 (2012).

16. Bögels, S., Schriefers, H., Vonk, W. & Chwilla, D. J. Pitch accents in context: How listeners process accentuation in referential communication. *Neuropsychologia* **49**, 2022–2036 (2011).

17. Heim, S. & Alter, K. Prosodic pitch accents in language comprehension and production: ERP data and acoustic analyses. *Acta Neurobiol. Exp. (Warsz.)* **66**, 55 (2006).

18. Li, W., Deng, N., Yang, Y. & Wang, L. Process focus and accentuation at different positions in dialogues: an ERP study. *Lang. Cogn. Neurosci.* **33**, 255–274 (2018).

19. Magne, C. *et al.* On-line Processing of "Pop-Out" Words in Spoken French Dialogues. *J. Cogn. Neurosci.* **17**, 740–756 (2005).

20. Schumacher, P. B. & Baumann, S. Pitch accent type affects the N400 during referential processing: *NeuroReport* **21**, 618–622 (2010).

21. Hostetter, A. B. When do gestures communicate? A meta-analysis. *Psychol. Bull.* **137**, 297–315 (2011).

22. Holle, H. & Gunter, T. C. The Role of Iconic Gestures in Speech Disambiguation: ERP

Evidence. *J. Cogn. Neurosci.* **19**, 1175–1192 (2007).

23. Obermeier, C., Holle, H. & Gunter, T. C. What Iconic Gesture Fragments Reveal about Gesture–Speech Integration: When Synchrony Is Lost, Memory Can Help. *J. Cogn. Neurosci.* **23**, 1648–1663 (2011).
24. Obermeier, C., Dolk, T. & Gunter, T. C. The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex* **48**, 857–870 (2012).
25. Bernardis, P., Salillas, E. & Caramelli, N. Behavioural and neurophysiological evidence of semantic interaction between iconic gestures and words. *Cogn. Neuropsychol.* **25**, 1114–1128 (2008).
26. Kelly, S. D., Kravitz, C. & Hopkins, M. Neural correlates of bimodal speech and gesture comprehension. *Brain Lang.* **89**, 253–260 (2004).
27. Kelly, S. D., Ward, S., Creigh, P. & Bartolotti, J. An intentional stance modulates the integration of gesture and speech during comprehension. *Brain Lang.* **101**, 222–233 (2007).
28. Özyürek, A., Willems, R. M., Kita, S. & Hagoort, P. On-line Integration of Semantic Information from Speech and Gesture: Insights from Event-related Brain Potentials. *J. Cogn. Neurosci.* **19**, 605–616 (2007).
29. Green, A. *et al.* Neural integration of iconic and unrelated coverbal gestures: A functional MRI study. *Hum. Brain Mapp.* **30**, 3309–3324 (2009).
30. Holle, H., Gunter, T. C., Rüschemeyer, S.-A., Hennenlotter, A. & Iacoboni, M. Neural correlates of the processing of co-speech gestures. *NeuroImage* **39**, 2010–2024 (2008).
31. Skipper, J. I., van Wassenhove, V., Nusbaum, H. C. & Small, S. L. Hearing Lips and Seeing Voices: How Cortical Areas Supporting Speech Production Mediate Audiovisual Speech Perception. *Cereb. Cortex* **17**, 2387–2399 (2007).

32. Willems, R. M., Özyürek, A. & Hagoort, P. Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *NeuroImage* **47**, 1992–2004 (2009).

33. Skipper, J. I. Echoes of the spoken past: how auditory cortex hears context during speech perception. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130297 (2014).

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

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

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

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

38. Biau, E., Fromont, L. A. & Soto−Faraco, S. Beat Gestures and Syntactic Parsing: An ERP Study. *Lang. Learn.* **68**, 102–126 (2018).

39. Dimitrova, D., Chu, M., Wang, L., Özyürek, A. & Hagoort, P. Beat that Word: How Listeners Integrate Beat Gesture and Focus in Multimodal Speech Discourse. *J. Cogn. Neurosci.* **28**, 1255–1269 (2016).

40. Drijvers, L. & Özyürek, A. Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension. *J. Speech Lang. Hear. Res.* **60**, 212–222 (2017).

41. Brunellière, A., Sánchez-García, C., Ikumi, N. & Soto-Faraco, S. Visual information constrains early and late stages of spoken-word recognition in sentence context. *Int. J.*

Psychophysiol. **89**, 136–147 (2013).

42. Hernández-Gutiérrez, D. *et al.* Does dynamic information about the speaker's face contribute to semantic speech processing? ERP evidence. *Cortex* **104**, 12–25 (2018).
43. Alday, P. M. M/EEG analysis of naturalistic stories: a review from speech to language processing. *Lang. Cogn. Neurosci.* **34**, 457–473 (2019).
44. Hagoort, P. & Brown, C. M. ERP effects of listening to speech: semantic ERP effects. 13 (2000).
45. Hasson, U., Egidi, G., Marelli, M. & Willems, R. M. Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition* **180**, 135–157 (2018).
46. Skipper, J. I. The NOLB model: a model of the natural organization of language and the brain. in *Cognitive Neuroscience of Natural Language Use* (ed. Willems, R. M.) 101–134 (Cambridge University Press, 2015). doi:10.1017/CBO9781107323667.006.
47. Obermeier, C., Kelly, S. D. & Gunter, T. C. A speaker's gesture style can affect language comprehension: ERP evidence from gesture-speech integration. *Soc. Cogn. Affect. Neurosci.* **10**, 1236–1243 (2015).
48. Gunter, T. C. & Weinbrenner, J. E. D. When to Take a Gesture Seriously: On How We Use and Prioritize Communicative Cues. *J. Cogn. Neurosci.* **29**, 1355–1367 (2017).
49. Frank, S. L., Otten, L. J., Galli, G. & Vigliocco, G. The ERP response to the amount of information conveyed by words in sentences. *Brain Lang.* **140**, 1–11 (2015).
50. Smith, N. J. & Levy, R. The effect of word predictability on reading time is logarithmic. *Cognition* **128**, 302–319 (2013).
51. Pernet, C. R., Chauveau, N., Gaspar, C. & Rousselet, G. A. LIMO EEG: A Toolbox for Hierarchical LInear MOdeling of ElectroEncephaloGraphic Data. *Comput. Intell.*

Neurosci. **2011**, 831409 (2011).

52. Smith, N. J. & Kutas, M. Regression-based estimation of ERP waveforms: I. The rERP framework: rERPs I. *Psychophysiology* **52**, 157–168 (2015).
53. Smith, N. J. & Kutas, M. Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations: rERPS II. *Psychophysiology* **52**, 169–181 (2015).
54. MacCallum, R. C., Zhang, S., Preacher, K. J. & Rucker, D. D. On the practice of dichotomization of quantitative variables. *Psychol. Methods* **7**, 19 (2002).
55. Alday, P. M., Schlesewsky, M. & Bornkessel-Schlesewsky, I. Electrophysiology Reveals the Neural Dynamics of Naturalistic Auditory Language Processing: Event-Related Potentials Reflect Continuous Model Updates. *eneuro* **4**, ENEURO.0311-16.2017 (2017).
56. Barr, D. J. Random effects structure for testing interactions in linear mixed-effects models. *Front. Psychol.* **4**, (2013).
57. Kristensen, L. B., Wang, L., Petersson, K. M. & Hagoort, P. The Interface Between Language and Attention: Prosodic Focus Marking Recruits a General Attention Network in Spoken Language Comprehension. *Cereb. Cortex* **23**, 1836–1848 (2013).
58. Sumby, W. H. & Pollack, I. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* **26**, 212–215 (1954).
59. Pilling, M. Auditory Event-Related Potentials (ERPs) in Audiovisual Speech Perception. *J. Speech Lang. Hear. Res.* **52**, 1073–1081 (2009).
60. van Wassenhove, V., Grant, K. W. & Poeppel, D. Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci.* **102**, 1181–1186 (2005).
61. BNC Consortium. The British national corpus, version 3 (BNC XML Edition). *Distrib. Oxf. Univ. Comput. Serv. Behalf BNC Consort.* **5**, 6 (2007).

62. Rapp, S. Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models. An Aligner for German. (1995).

63. Shannon, C. E. Communication theory of secrecy systems. *Bell Syst. Tech. J.* **28**, 656–715 (1949).

64. Levy, R. Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177 (2008).

65. Schäfer, R. & Bildhauer, F. Building large corpora from the web using a new efficient tool chain. in 486–493 (2012).

66. Boersma, P. Praat, a system for doing phonetics by computer. *Glot Int* **5**, 341–345 (2001).

67. Kakouros, S., Salminen, N. & Räsänen, O. Making predictable unpredictable with style – Behavioral and electrophysiological evidence for the critical role of prosodic expectations in the perception of prominence in speech. *Neuropsychologia* **109**, 181–199 (2018).

68. Sloetjes, H. & Wittenburg, P. Annotation by category-ELAN and ISO DCR. in (2008).

69. Bates, D. *et al.* Package 'lme4'. *Convergence* **12**, 2 (2015).

70. Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).

71. Lopez-Calderon, J. & Luck, S. J. ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front. Hum. Neurosci.* **8**, 213 (2014).

72. Pernet, C., Latinus, M., Nichols, T. & Rousselet, G. Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. *J. Neurosci. Methods* **250**, 85–93 (2015).

73. Winsler, K., Midgley, K. J., Grainger, J. & Holcomb, P. J. An electrophysiological megastudy of spoken word recognition. *Lang. Cogn. Neurosci.* **33**, 1063–1082 (2018).

Figures

Table 1. Full result: linear mixed effects regression model with N400 (300-600ms) as dependent variable.

Fixed Effects	β	Std Error	t	p
(Intercept)	0.004	0.010	0.416	0.677
Predictor Variables				
Surprisal	0.011	0.016	0.726	0.468
Mean F0	0.010	0.002	4.710	<.001***
Mouth Informativeness	-0.004	0.003	-1.477	0.140
Meaningful Gesture (Present)	0.006	0.001	4.867	<.001***
Beat Gesture (Present)	-0.005	0.001	-3.386	0.001**
Surprisal:Mean F0	0.017	0.002	7.996	<.001***
Surprisal:Mouth Informativeness	-0.002	0.002	-0.726	0.468
Surprisal:Meaningful Gesture (Present)	0.008	0.001	5.653	<.001***
Surprisal:Beat Gesture (Present)	-0.012	0.001	-9.248	<.001***
Mean F0:Mouth Informativeness	0.003	0.002	2.051	0.040*
Mean F0:Meaningful Gesture (Present)	0.004	0.001	3.710	<.001***
Mean F0:Beat Gesture (Present)	-0.006	0.002	-3.777	<.001***
Mouth Informativeness:Meaningful Gesture (Present)	-0.006	0.001	-5.757	<.001***
Mouth Informativeness:Beat Gesture (Present)	0.003	0.002	1.874	0.061
Surprisal:Mean F0:Mouth Informativeness	-0.013	0.005	-2.554	0.011*
Surprisal:Mean F0:Meaningful Gesture (Present)	0.007	0.006	1.278	0.201
Surprisal:Mean F0:Beat Gesture (Present)	0.002	0.006	0.453	0.650
Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	-0.010	0.005	-2.168	0.030*
Surprisal:Mouth Informativeness:Beat Gesture (Present)	0.003	0.004	0.683	0.495

Control Variables				
Word Order	-0.010	0.002	-4.782	<.001***
Word Length	-0.011	0.004	-2.361	0.018*
Sentence Order	-0.009	0.001	-9.247	<.001***
Baseline	0.788	0.001	876.10	<.001***
Frequency	0.036	0.009	3.834	<.001***
Electrode X	-0.006	0.001	-7.075	<.001***
Electrode Y	0.008	0.001	8.622	<.001***
Electrode Z	0.001	0.001	0.934	0.351

Random Effects		Variance	Std.Dev.
Lemma	(Intercept)	0.012	0.111
	Surprisal	0.045	0.213
Participant ID	(Intercept)	0.001	0.034
	Surprisal:Mean F0:Mouth Informativeness	0.001	0.028
	Surprisal:Mean F0:Meaningful Gesture (Present)	0.001	0.030
	Surprisal:Mean F0:Beat Gesture (Present)	0.001	0.030
	Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	0.001	0.024
	Surprisal:Mouth Informativeness:Beat Gesture (Present)	0.001	0.023

Model				
AIC	BIC	logLik	deviance	df.resid
892158.7	892746	-446026.3	892052.7	480159

Notes. * $p < .05$, ** $p < .01$, *** $p < .001$

	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

Figure1. Example of stimuli and annotations. Annotation was carried out for content words only. Each frame corresponds to an image during each such word.

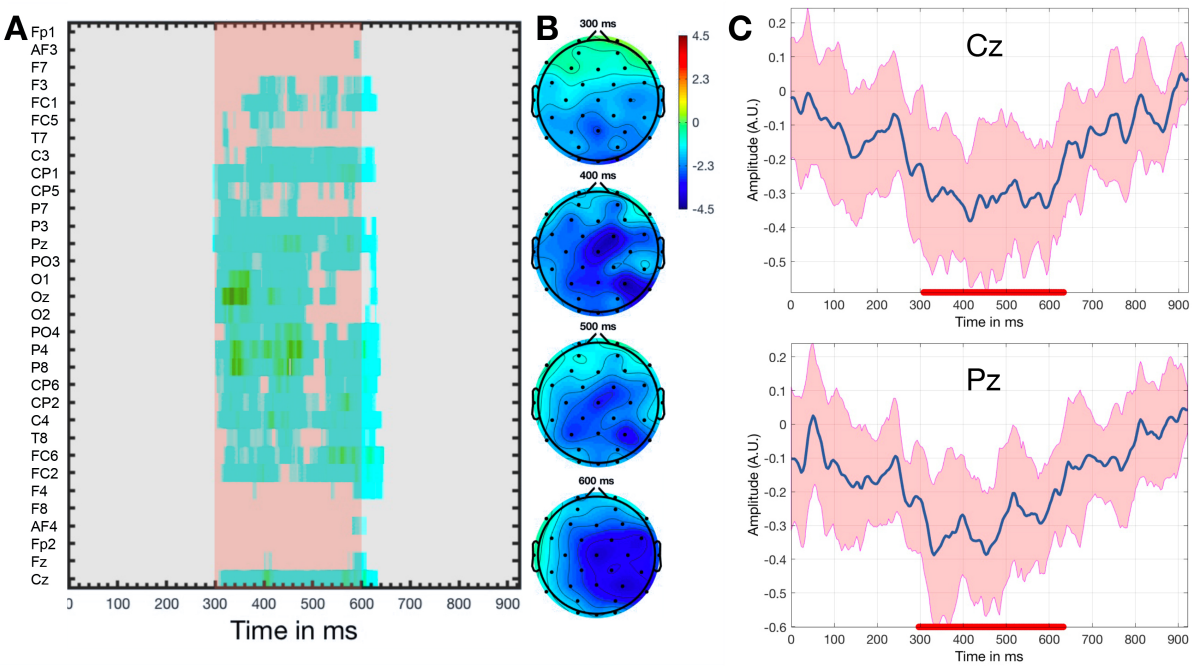

Figure 2. Hierarchical linear modelling results showing the ERP sensitive to surprisal (one-sample t-test P<0.05, cluster-corrected). (A) Time window (300-600ms) showing increased significant negativity associated with surprisal (in pink). Grey areas are not statistically significant. (B) Topographic maps illustrating the scalp distribution for the 300-600 time window. Deeper blue area indicates more negative beta values. (C) Averaged beta plot for electrode Cz and Pz illustrating that beta values for surprisal were significantly negative compared with 0 (flat waveform) in 300-600ms. The blue line indicates the average beta value, while red indicates the confidence interval. The red line underlying the figures indicates the significant time window. Cz and Pz are chosen here because they are most often used to depict N400 effects (that are maximal at central-parietal locations [8])

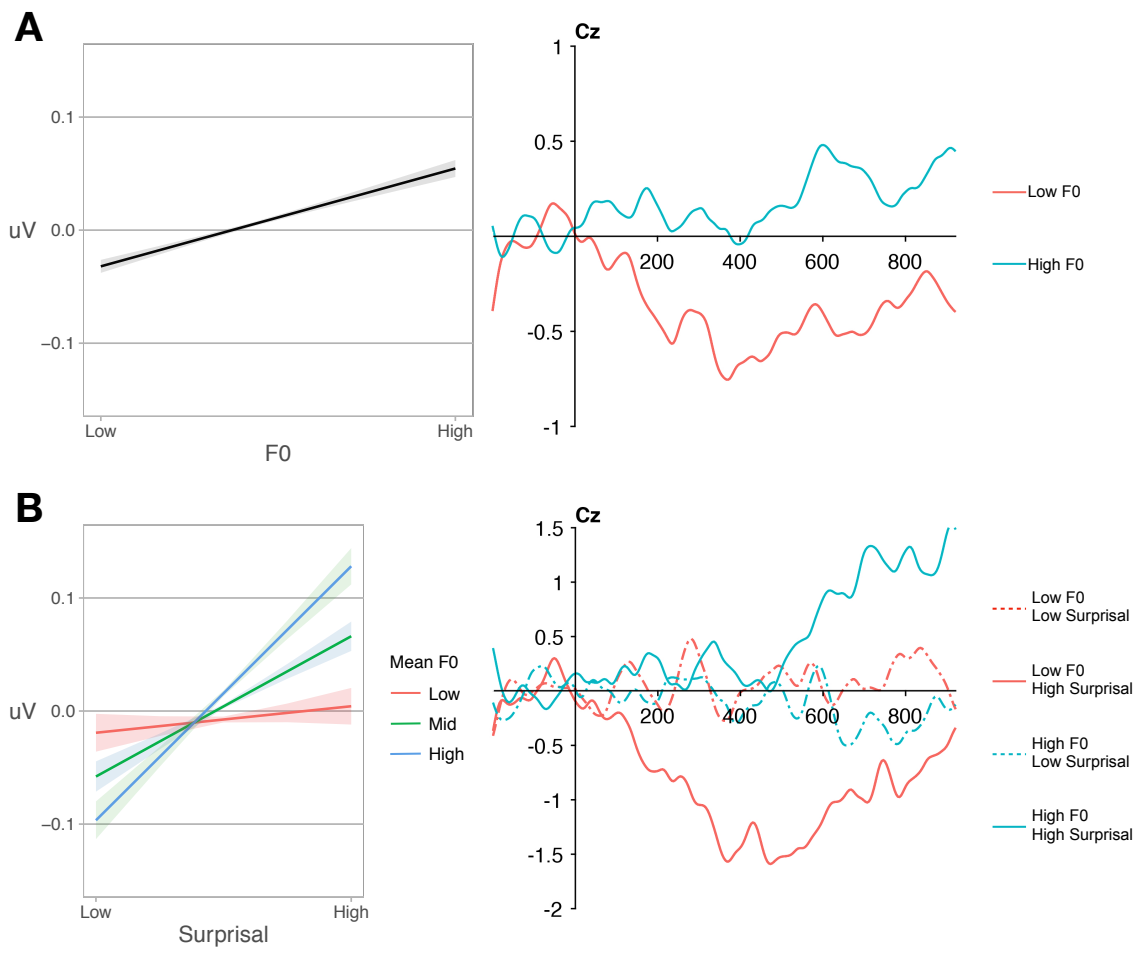

Figure 3. Prosodic Accentuation (mean F0) modulation of N400 amplitude. (A) Main effect of Prosodic Accentuation. (B) Interaction between Prosodic Accentuation and Surprisal. Plots on the left depict the predicted value of expected value of the mean amplitude of the ERP within 300-600ms (grey areas = confidence intervals). Plots on the right show the EEG waveform. For illustrative purposes, in the EEG plots all continuous variables are categorized and the EEG waveform was additionally filtered with 15Hz low pass filter.

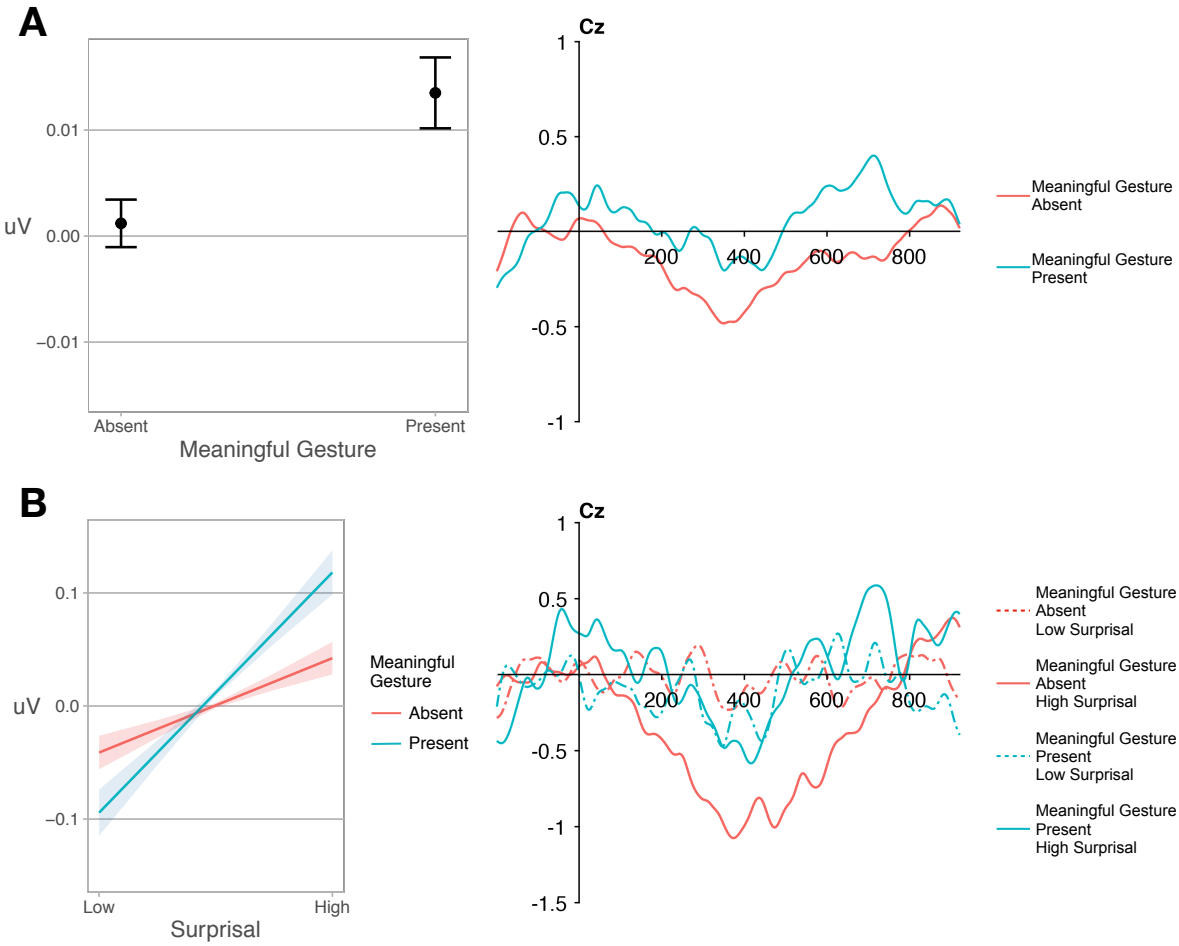

Figure 4. Meaningful gesture modulation of N400 amplitude. (A) Main effect of meaningful gestures. (B) Interaction between Meaningful Gestures and Surprisal. Conventions are the same as in Figure 3.

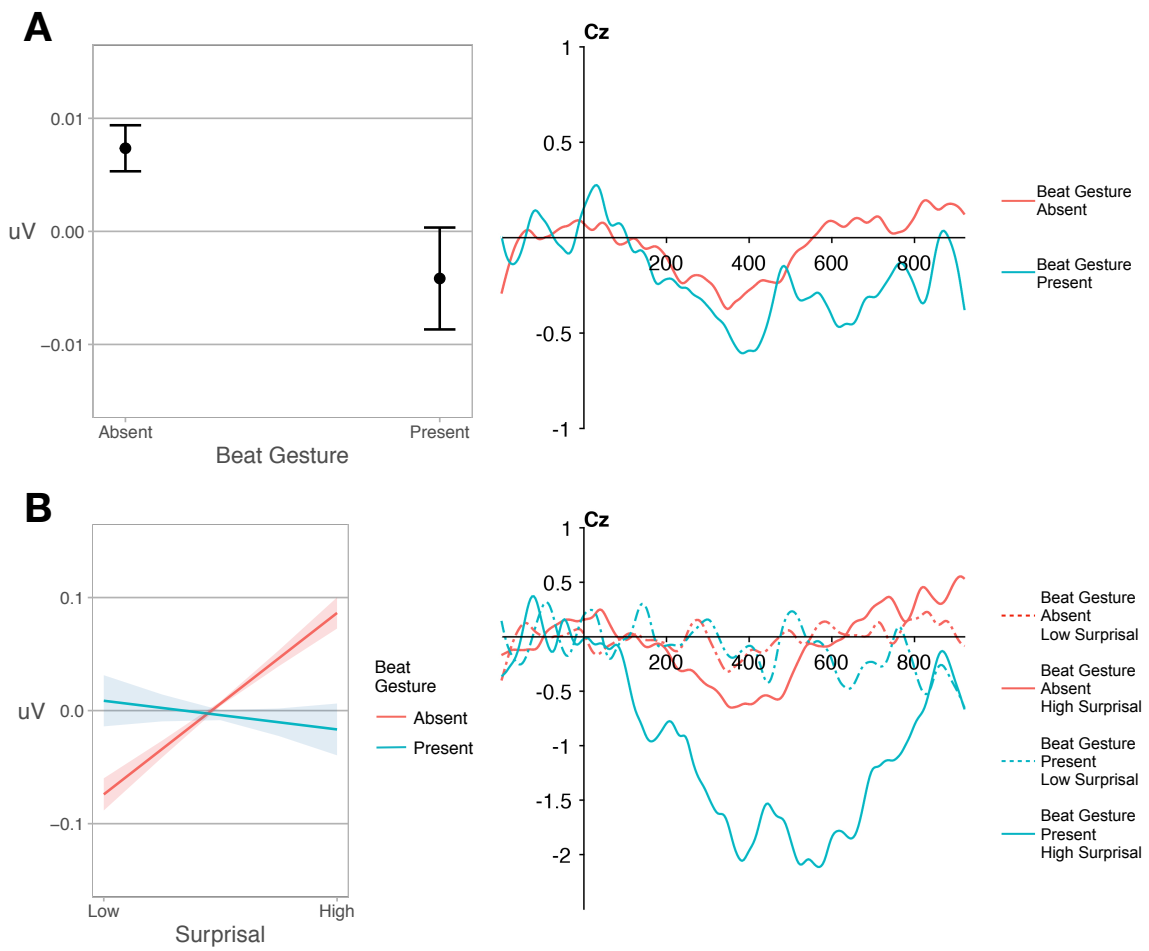

Figure 5. Beat gesture modulation of N400 amplitude. Main effect of beat gestures. (B) Interaction between Beat Gestures and Surprisal. Conventions are the same as in Figure 3.

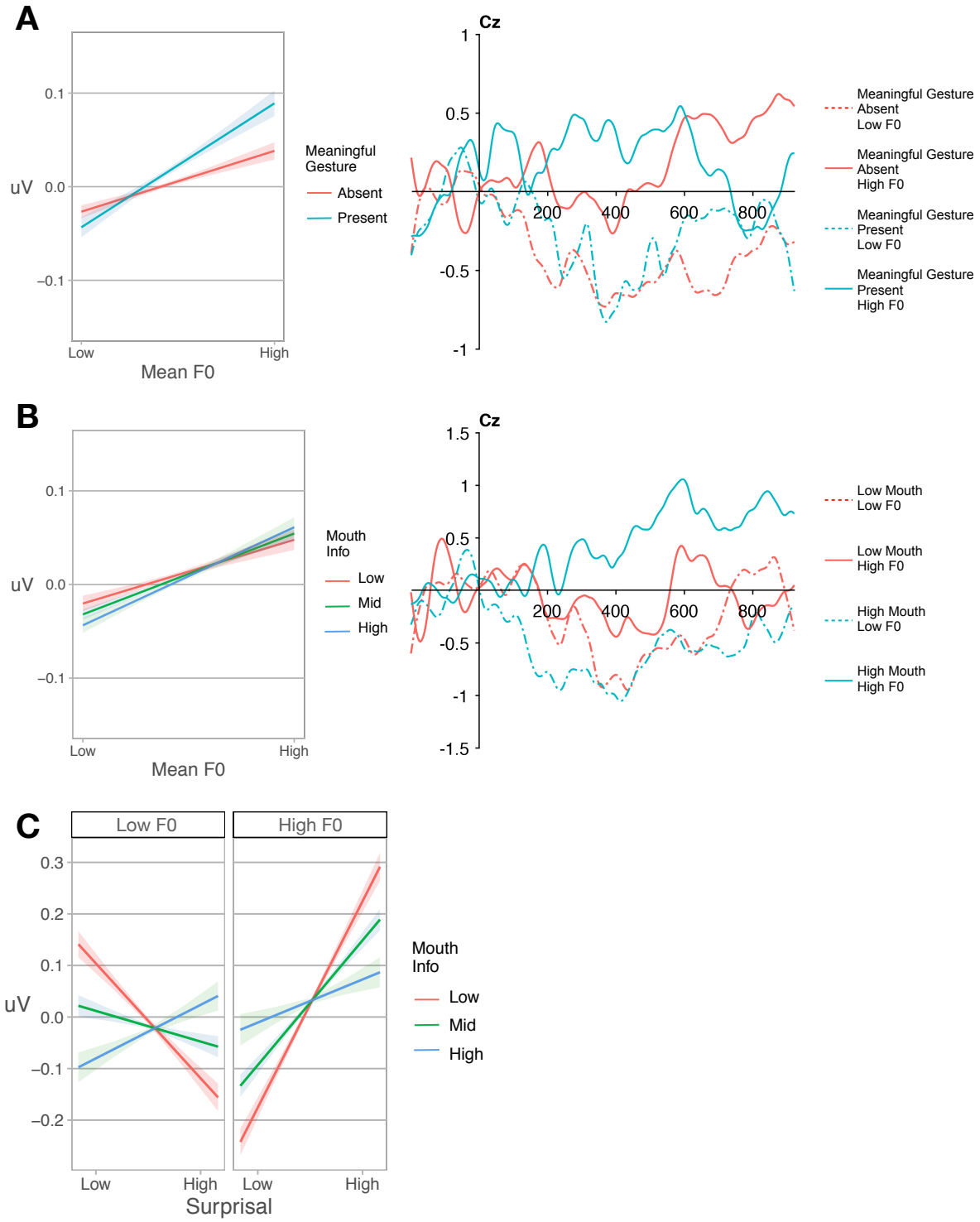

Figure 6. Interactions between Prosodic Accentuation and other Cues. (A) Interaction between Prosodic Accentuation and Meaningful Gestures. (B) Interaction between Prosodic Accentuation and Mouth Informativeness. (C) Interaction between Prosodic Accentuation, Surprisal and Mouth Informativeness. Conventions are the same as in Figure 3.

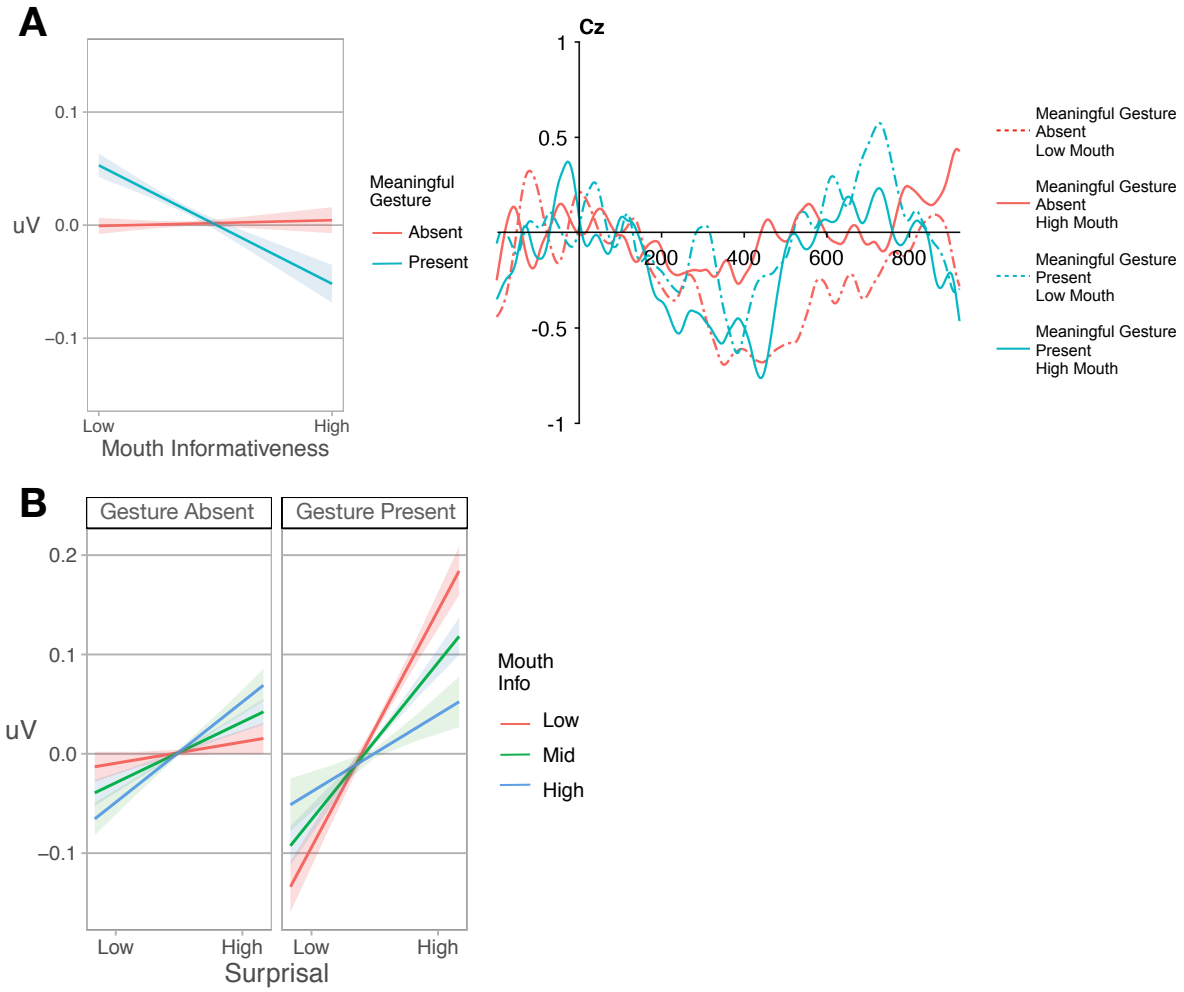

Figure 7. Interactions between Mouth Informativeness and other Cues. (A) Interaction between Mouth Informativeness and Meaningful Gestures. (B) Interaction between Mouth Informativeness, Surprisal and Meaningful Gestures. Conventions are the same as in Figure 3