Comprehension in-situ: how multimodal information shapes language processing

Candidate: Ye Zhang (16006868)

Degree: Cognitive Neuroscience

Institution: Department of Experimental Psychology, Faculty of Brain Sciences,

University College London

**Declaration: I, Ye Zhang, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis*

Abstract

The human brain supports communication in dynamic face-to-face environments where spoken words are embedded in linguistic discourse and accompanied by multimodal cues, such as prosody, gestures and mouth movements. However, we only have limited knowledge of how these multimodal cues jointly modulate language comprehension. In a series of behavioural and EEG studies, we investigated the joint impact of these cues when processing naturalistic-style materials. First, we built a mouth informativeness corpus of English words, to quantify mouth informativeness of a large number of words used in the following experiments. Then, across two EEG studies, we found and replicated that native English speakers use multimodal cues and that their interactions dynamically modulate N400 amplitude elicited by words that are less predictable in the discourse context (indexed by surprisal values per word). We then extended the findings to second language comprehenders, finding that multimodal cues modulate L2 comprehension, just like in L1, but to a lesser extent; although L2 comprehenders benefit more from meaningful gestures and mouth movements. Finally, in two behavioural experiments investigating whether multimodal cues jointly modulate the learning of new concepts, we found some evidence that presence of iconic gestures improves memory, and that the effect may be larger if information is presented also with prosodic accentuation. Overall, these findings suggest that real-world comprehension uses all cues present and weights cues differently in a dynamic manner. Therefore, multimodal cues should not be neglected for language studies. Investigating communication in naturalistic contexts containing more than one cue can provide new insight into our understanding of language comprehension in the real world.

Impact Statement

Language has evolved, is learnt and is mostly used in face-to-face contexts in which comprehenders can take advantage of multiple multimodal cues such as prosody, gestures and mouth movements and prosody in order to support speech and language processing. Yet, prior behavioural and neurobiological work has focused on speech only or on the impact of a single multimodal cue (prosody, or mouth movements, or gestures) on speech processing, showing that each of them can modulate brain activity associated with language comprehension. However, by only manipulating one (speech) or two variables (e.g., speech and gesture or speech and prosody) while the others are controlled, the natural pattern of correlations among the cues is lost with unknown consequences for the processing.

Here we present some of the very first language comprehension studies that uses more ecologically valid experimental paradigms allowing us to assess how the brain dynamically weights speech along with the other audiovisual cues. Instead of removing potentially correlated cues (e.g. blocking mouth movements when studying gestures), we quantified each multimodal cue and analysed their joint impact in a series of behavioural and EEG studies. Our findings provide answers to two key questions concerning language comprehension. First, we asked if the processing of audio-visual multimodal cues is central in natural language processing. We found that multimodal cues, such as gestures and prosody, always affect speech processing. Similar patterns are found across L1 and L2 comprehenders, and across electrophysiological and behavioural tasks, indicating their centrality. Second, we demonstrate the dynamic nature of multimodal cue processing showing that the weight given to each cue (i.e., to what extent a specific cue affects processing at a given moment) depends on which other informative cue is present at that moment in

processing. Thus, for example, meaningful gestures showed larger effects with the presence of prosodic stress.

Thus, we provide direct behavioural and neural evidence that language comprehension cannot be reduced to speech-only processing. These findings have impact on our understanding of language as well as outside academia. First, these findings challenge the traditional linguistic primacy assumed by most existing theoretical approaches. Instead, language comprehension may be better viewed as a process where comprehenders construct meanings based on all information available, both linguistic and multimodal.

Our findings provide experimental evidence to further constrain theories of the cognitive mechanism underlying multimodal comprehension. Taking an approach aligned with the cue combination approaches from sensory and motor neuroscience (e.g. Jacobs, 2022), we show how we can develop more ecologically valid paradigms that do not sacrifice rigour to study language comprehension. Our paradigm can be employed by future studies investigating language comprehension in more naturalistic settings.

Outside academia, our findings that multimodal cues improve comprehension and learning outcome motivate educational practices to attend to multimodal information. More generally, these findings encourage all practices involving communications to include multimodal information in their communicative setting, especially given the current COVID-19 pandemic, which restricted multimodal communication (due to e.g. mask wearing and reduction of face-to-face meetings).

Table of Contents

Table of Figures

Chapter 3

Chapter 1

1 Overview: Multimodal language comprehension

The human brain supports communication in dynamic face-to-face environments where spoken words are embedded in linguistic discourse and accompanied by multimodal cues. While many previous studies have suggested that each of these cues individually can modulate language comprehension (e.g. prosody: Cole, 2015; Gestures: Hostetter, 2011; Mouth: Peelle & Sommers, 2015), most of the investigations focused on individual cues out of their naturalistic context, which usually contains other cues. This dominant paradigm lacks ecological validity, because it can amplify the effect of the individual cue being investigated, and hide away potential interactions between cues.

Therefore, this PhD thesis investigates how multimodal cues individually and crucially jointly modulate language comprehension in more naturalistic settings, where multiple cues co-occur. Chapter 2 offers a literature review of how each individual cue, including prosody, gestures and mouth movements, modulate language comprehension. The review shows that multimodal cues are central to language comprehension, but also highlights the need to investigate their impacts in more naturalistic contexts. In order to carry out such studies, it is necessary first to be able to quantify the cues. While there are objective and agreed upon ways to quantify prosodic and hand-gesture cues, this is not the case for mouth movements. Chapter 3 presents a corpus of mouth informativeness for English words, thus

providing normative data to be used in the following studies to track mouth information. Chapter 4 describes an EEG experiment and a subsequent replication where we investigated how multimodal cues (i.e., prosodic modulation, meaningful and beat gestures, mouth informativeness) dynamically affect language comprehension, as indexed by N400. We presented participants with videos where an actress uttered passages with naturally occurring prosody, gestures and mouth movements. We then measured the joint impact of linguistic predictability and multimodal information on N400 amplitude. Chapter 5 presents another EEG study using the same paradigm but testing non-native speakers, investigating whether and how multimodal cues modulate second language processing. The next planned step would have been to carry out a similar study with fMRI, however this plan had to be abandoned due to the COVID pandemic. Thus, Chapter 6 introduces an online study investigating whether and how gestures and prosody jointly modulate the learning of new concepts. We manipulated the interaction between the two variables while keeping the naturalistic setting. Finally, Chapter 7 discusses the general findings in the context of current theories of communicative efficiency. Overall, this body of work suggests that multimodal information is an integral part of language comprehension, contributing to efficiency and that multimodal comprehension in the naturalistic context is dynamic, actively changing based on linguistic information and other multimodal cues present.

Reference

1. Cole, J. (2015). Prosody in context: a review. *Language, Cognition and Neuroscience*, *30*(1-2), 1-31.

2. Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological bulletin*, *137*(2), 297.

3. Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181.

Chapter 2

2 Review: Language comprehension in face-to-face communications

2.1 Introduction

Imagine grabbing a coffee with an old friend and enjoying a long-awaited chat. While listening to her speech, you will always hear her prosody, see her hand gestures, and watch her mouth moving. Indeed, this information conveyed via "non-linguistic" channels are always present in our daily face-to-face communications. Speakers' prosody conveys information that is helpful to the segmentation of words, organisation of syntactic components and identification of new and important information (Cole, 2015). Hand gestures can imitate properties of actions or objects, single out the reference from the physical context, and draw attention to the ongoing speech (McNeill, 1992). Mouth movements carry sensory information that can help the identification of speech (Peelle and Sommers, 2015). These cues constantly co-occur during naturalistic language use in face-to-face contexts, conveying information simultaneously and interactively through different channels. For example, the production of prosody is synchronised with both hand gestures (e.g. McNeill, 1992; Esteve-Gibert & Prieto, 2013) and mouth movements (e.g. Ménard, Leclerc & Tiede, 2014). Gestures and mouth movements are generated by the motor system in an interactive style (e.g. Vainio, 2019) and are also simultaneously perceived in the visual channel (e.g. Gullberg & Holmqvist, 2006; Beattie, Webster & Ross, 2010).

The information conveyed via traditionally “non-linguistic” signals is referred here as “multimodal cues”, representing different modes or channels to convey messages (although individual cues may occupy one single modality only, e.g. gestures, which are visual only). This is a “left-over” category defined negatively, because despite the constant presence and co-occurrence of multimodal cues in real-world communication, language is primarily studied without considering these cues and their interactions. One reason is that traditional linguistic theories view language as a population-level system, where the “core” linguistic system includes the structural, categorical components of speech (e.g. phonemes, words, sentences), which are commonly shared by everyone in a population at any given time (Murgiano, Motamedi & Vigliocco, 2021). At the same time, multimodal cues that accompany single utterances are deemed “non-linguistic”. Therefore, previous studies primarily focus on studies of text/speech (at least partially) stripped away from its physical context (e.g. in texts, where all multimodal cues are removed; or in speech, where all visual cues are removed). Moreover, partially due to the constraints posed by technical challenges, the dominant experimental paradigms assume a reductionist approach, in which experimental conditions are created by manipulating a minimal difference. Thus, even in the limited number of studies investigating the impact of multimodal cues, co-occurrence between cues is removed artificially (e.g. it is a common practice to hide the speaker’s mouth when studying the impact of gestures, e.g. Holler & Gunter, 2007).

However, recent theoretical and experimental work challenges this approach, arguing for incorporating communicative and physical contexts in language studies (Holler & Levinson, 2018; Hasson, Egidi, Marelli & Willems, 2018; Murgiano, Motamedi & Vigliocco, 2021). In light of these views, language processing involves

not just linguistic input but also multimodal cues such as prosodic, gestural, mouth information, along with their co-occurrence. This view is supported by accumulating evidence from the evolution as well as the development of language. It has been proposed that language may have originated from iconic vocalisation (e.g. Perlman & Cain, 2014) or gestures (e.g. Vigliocco, Perniss & Vinson, 2014). Young infants are sensitive to caregivers' prosody (e.g. Fernald et al., 1989; Spinelli, Fasolo & Mesman, 2017), gestures (e.g. Iverson & Goldin-Meadow, 2005; Özçalışkan & Goldin-Meadow, 2005) and mouth movements (e.g. Lewkowicz & Hansen-Tift, 2012; Tenenbaum, Sobel, Sheinkopf, Malle & Morgan, 2015).

Here in this review, we focus on language comprehension and assess whether multimodal cues and their interaction modulate the processing of speech. We review behavioural, EEG and imaging studies of whether/how prosody, gestures, and mouth movements modulate listeners' comprehension. These multimodal cues are selected as they are better researched and are always present in face-to-face communication (but note that there are other multimodal cues that affect comprehension, such as speakers' gaze or torso movements, see e.g. Holler & Levinson, 2019). We aim to address two core questions: 1) are multimodal cues central to language comprehension? This question is addressed by evaluating studies focusing on the impact of each multimodal cue. Evidence that each multimodal cue modulates comprehension at different stages would support that language cannot be reduced to linguistic information alone. 2) Do multimodal cues interact to affect language comprehension? We approach this question by examining studies investigating the joint impact of more than one cue. Evidence that listeners process multimodal cues differently based on the presence of other cues would support that it is necessary to introduce their co-occurrence in future experiments. In

the last section, we will discuss the implication of these findings for future theoretical and empirical work, including the studies in this thesis.

2.2 Are multimodal cues central to language comprehension?

In this section, we review studies focusing on how each multimodal cue modulates language comprehension. Due to the volume of studies, we only aim to provide a comprehensive review of how each cue affects language processing instead of complete coverage of the research topics in each field. We will first go through behavioural, EEG and imaging studies on prosody, gestures and mouth movements one by one before discussing whether the findings support the argument that multimodal cues are central to language comprehension.

2.2.1 Prosody

Prosody, the acoustic properties of utterances that vary independently of the lexical items (Wagner & Watson, 2010), underlies every single speech sound. By modifications of acoustic features such as pitch, intensity, duration or their combinations, speakers produce various prosodic phenomena, such as prosodic stress (i.e. stressing a certain part of the speech, usually indicated by higher pitch, louder intensity and longer duration), prosodic breaks (i.e. short breaks between different parts of the speech), or prosodic changes (i.e. change of the continuous intonation of the speech, such as a rising pitch towards the end of a question). Listeners take these prosodic properties into account during all stages of language comprehension, such as the initial access of lexical information, the construction of a syntactic structure, and the processing of discourse-level information. Due to the multi-faceted nature of prosody, we don't aim to provide an in-depth review of every function of prosody (but see previous reviews in Cutler, Dahan & van Donselaar, 1997; Wagner & Watson, 2010; Cole, 2014). Rather, this section will provide a brief

overview of the behavioural, EEG and imaging studies investigating whether and how prosody modulate language comprehension at different stages.

Prosody modulates comprehension in the early word-level processing in the form of prosodic stress. Earlier behavioural studies suggested that prosodic stress can enhance the saliency of syllables (Grosjean & Gee, 1987), therefore making stressed words more recognisable (Lieberman, 1963; Cutler & Foss, 1977; McAllister, 1991), and any distortions more identifiable (Cole, Jakimik, & Cooper, 1978; Cole and Jakimik, 1980). EEG studies also support that prosodic stress affects the semantic processing of individual words, indexed by modulation of N400, an event-related-potential (ERP) peaking negatively around 400ms post-stimulus, widely associated with semantic processing difficulty (due to prediction and/or integration process, which is out of the scope of the current thesis; Kutas & Federmeier, 2011). Some studies found that words bearing prosodic stress/accent elicit more negative N400 overall, which is interpreted as deeper semantic processing, as stressed words are more salient and thus attract more cognitive resources (Li, Hagoort & Yang, 2008; Li, Lu & Zhao, 2014). However, other studies reported the opposite effect (Wang & Chu, 2012), interpreted as prosodic stress making semantic processing easier, thus the less negative N400. While the reason for the divergence of patterns remains to be explored, it may be related with the different effect of prosodic stress for words with different semantic properties: for example, studies found that prosodic stress can enlarge the classic N400 effect (i.e. the N400 difference between congruent and incongruent words, as the N400 amplitude is more negative for incongruent words; Wang, Bastiaansen, Yang & Hagoort, 2011; Li & Ren 2012). Therefore, when highlighted by prosodic stress, a incongruent word (but not necessarily a congruent one) elicits larger N400 as its

poor fit to the semantic context becomes more prominent. The effect of prosodic stress is also dependent on the predictability/information status of a word, which will be reviewed in detail in later sections.

Concerning the localisation of prosodic processing in the brain, one central debate is whether the left hemisphere is also involved in addition to the right hemisphere. A number of early imaging (and dichotic listening studies, e.g. Blumstein & Cooper, 1974) studies found that prosodic features were processed predominantly in the right hemisphere (lesion: Weintraub et al., 1981, Baum and Pell, 1999; fMRI: Gandour et al., 2003, Meyer et al., 2003). However, when prosody is used in meaning processing, other studies found that prosody may also activate the left hemisphere, such as the processing of tonal languages that rely on prosodic features to differentiate meanings (e.g. Gandour and colleagues (2003, 2004), or when the task manipulations promote semantic processing (e.g. Meyer et al., 2004). While these studies indicated that prosody contributes to meaning comprehension in general, a common paradigm employed was to compare speech with prosody against flattened speech, therefore merging different prosodic phenomena. When looking specifically at the processing of prosodic stress, Kristensen, Wang, Petersson and Hagoort (2012) presented participants with sentences in which a target congruent/incongruent word is with or without prosodic stress. They found that the prosodic manipulation showed bilateral involvement of brain areas, including bilateral superior/inferior parietal lobe (SPL and IPL), superior and middle temporal gyrus (STG and MTG) and inferior and middle parts of the frontal gyrus (IFG and MFG). The bilateral activations indicate that prosodic stress is taken into account to construct meaning from the speech. Further, these areas overlapped with the domain general attention network localized in an auditory spatial attention task (in

which participants indicated whether a beeping sound was presented in right/left ear), both activating bilateral SPL, IPL, STG and left precentral cortex. They, therefore, argued that prosodic stress engages the domain-general attention function to modulate language processing. Further, they observed an interaction between prosody and congruence in bilateral IPL, overlapping with the defined attention network, interpreted as prosodic accentuation enhancing attention especially to semantically incongruent words, This is also in line with the EEG studies reporting a larger N400 difference between congruent and incongruent words when accented (e.g. Wang, Bastiaansen, Yang & Hagoort, 2011; Li & Ren 2012). .

At the sentence level, the prosody structure of spoken languages (such as prosodic boundaries) can modulate the segmentation of longer auditory input into syntactic elements (see review in Speer, Warren & Shafer, 2003). Appropriate prosodic marking of boundaries facilitates speech processing, compared with inappropriate pairs (e.g. Warren, Grabe & Nolan, 1995; Speer, Kjelgaard & Dobroth, 1996; Kjelgaard & Speer, 1999; Watson & Gibson, 2005). Moreover, listeners make use of the mapping between prosodic and syntactic patterns to resolve ambiguity (Lehiste, Olive, & Streeter, 1976; Price et al., 1991; Snedeker & Trueswell, 2003; Snedeker & Casserly, 2010). Indeed, prosody may affect syntactic processing universally, as similar facilitatory effects are found across languages with very different syntactic and prosodic structures (Bader, 1998; Kang & Speer, 2002; Kang, Speer, & Nakayama, 2004; Nakamura, Arai & Mazuka, 2012). Interestingly, despite the strong evidence that listeners benefit from prosodic cues in syntactic processing, the exact phonological properties used to convey the same syntactic structure vary across and within speakers in the real world (Schafer, Speer, Warren & White, 2005). Therefore, the mapping between prosodic and syntactic structure in the real

world is not one-to-one but many-to-many, and listeners may be relying on prosodic structures based on the combination of pitch, intensity and duration, instead of one single parameter. Electrophysiological studies provided further support that prosodic information and syntactic information are integrated during comprehension. In an EEG study, Steinhower, Alter & Friederici (1999) found that prosodic phrase boundaries that mismatch with the syntactic structure trigger larger N400, indicating greater difficulty in processing. Further, the mismatch also induced larger P600, an ERP associated with syntactic processing that peaks ~600ms post-stimulus, indicating greater difficulty in syntax processing. Finally, the authors also found that the presence of prosodic phrase boundaries themselves triggered a slow positive shift of the ERP, which they termed Closure Positive Shifts (CPS) and considered to reflect processing of prosodic patterns. This pattern of N400 (Eckstein, & Friederici, 2005; Mietz et al., 2008; Bögels, Schriefers, Vonk, Chwilla, & Kerkhofs, 2010; Steinhauer, Abada, Pauker, Itzhak, & Baum, 2010; Pauker et al., 2011; Nickels, & Steinhauer, K, 2018), P600 (Eckstein, & Friederici, 2005; Mietz et al., 2008; Steinhauer, Abada, Pauker, Itzhak, & Baum, 2010; Hwang, & Steinhauer, 2011; Pauker et al., 2011; Roncaglia-Denissen, Schmidt-Kassow, & Kotz, 2013; Nickels, & Steinhauer, K, 2018) and CPS (Pannekamp et al., 2005; Bögels, Schriefers, Vonk, Chwilla, & Kerkhofs, 2010; Steinhauer, Abada, Pauker, Itzhak, & Baum, 2010; Hwang, & Steinhauer, 2011; Pauker et al., 2011; Nickels, & Steinhauer, K, 2018, reviewed by Bögels, Schriefers, Vonk, & Chwilla, 2011) was repeatedly replicated across different groups and languages, providing strong evidence that prosodic information is united with syntax information when comprehending languages. Bilateral prefrontal and inferior frontal areas and superior temporal areas have been associated with the processing of prosodic/syntax interface (Strelnikov et al., 2006;

Ischebeck, Friederici, & Alter, 2008; van der Burght et al., 2019). Contributing to the debate about the lateralisation of prosody processing, van der Burght and colleagues (2019) found left-lateralised activity in IFG when prosody was the only cue from which participants could establish the sentence structure. However, the inferior frontal activity was right lateralised when the prosody cue was redundant for language comprehension, even though the prosody was identical for the two conditions. This result supports both the argument that listeners make use of prosodic cues in the processing of syntactic information, as well as that the left hemisphere is activated in the processing of prosody when it contributes to the interpretation of sentence meanings.

At the discourse level, prosody affects information segmentation and marks whether any information is new or given (see review in Venditti & Hirschberg, 2003). Early behavioural works found that prosodic pattern is associated with the discourse structure: a higher pitch is associated with words in paragraph initial position (Lehiste, 1975), at the beginning of new topics (Yule, 1980; Swerts & Geluykens, 1994), and new information in the context (Aylett & Turk, 2004; Cruttenden, 2006). Listeners may potentially exploit these correlations to segment and process discourse information. Among these correlations, the association between prosodic accentuation (i.e. prosodic stress that makes a part of speech more acoustically salient, usually indicated by higher pitch, longer duration and louder sound) and information structures (i.e. whether a piece of information is new or given in the context) received most attention. New information is more likely to receive pitch accent while given information is more likely to be de-accentuated (Aylett & Turk, 2004), and this pattern is found across different languages (Cruttenden, 2006). This association between prosodic accentuation and the new (versus given) information

status not only applies to words occurring for the first time (e.g. Terken and Nooteboom, 1987; Birch and Clifton, 1995), but also words that have occurred before but now receive a new interpretation (e.g. Terken and Hirschberg, 1994), and words that are less predictable from the constraints provided by the task or the linguistic context (Gahl & Garnsey, 2004; Watson, Arnold & Tanenhaus, 2008). Thus, prosodic accentuation may be interpreted as marking information with lower predictability in the context. Indeed, listeners make use of the link between prosody and information structure when comprehending speech, as violations of this pattern hinder language comprehension while the appropriate pairing of prosody and information status facilitates comprehension, indexed by response time in different comprehension tasks (e.g. Bock & Mazzela, 1983, measuring recognition time for each sentence; Terken and Nooteboom, 1987; measuring response time for validating whether the sentence matched with a picture). Eye-tracking studies measured participants' fixiations on new/old objects upon hearing accented/unaccented nouns, and found a preference to look at the new object when the nouns were accented and/or old object when unaccentuated for adults (Dahan, Tanenhous & Chambers, 2002; Arnold, 2008), children (Arnold, 2008) as well as children as young as 24-month-old (Grassmann & Tommasello, 2010) To note, while Dahan and coleagues (2002) found both accent-new and unaccent-old associations, Arnold (2008) only found a preference for old objects when target words are not accentuated, therefore arguing that while lack of prosodic accentuation specifically marks more predictable/accessible old information, presence of accentuation serves a more general function. EEG literature harbours similar debate of whether the presence of (redundant) accentuation or the missing of accentuation (or both) drives comprehension difficulty associated with the mismatch between prosody and

information status. Compared with the appropriate pairing of accentuation and information status, inappropriate pairing elicits larger N400, indicating increased difficulty in semantic processing when processing prosodic information with linguistic information that has mismatching newness (Magne et al., 2005; Magne Heim & Alter 2006; Li, Hagoort, & Yang, 2008; Baumann & Schumacher, 2012; Dimitrova, Stowe, Redeker & Hoeks, 2012; Li & Yang, 2013). However, when separating the mismatch further into two different scenarios, namely missing accent (lack of accentuation for new information) and superfluous accent (accentuation of old information), while some studies reported that both types of mismatches trigger larger N400 (e.g. Magne et al., 2005), others associated larger N400 exclusively for missing accent (e.g. Hruska & Alter, 2004; Baumann & Schumacher, 2011; Bögels, Schriefers, Vonk, Chwilla, 2011) or superfluous accent (e.g. Dimitrova, Stowe, Redeker, Hoeks, 2012; Wang, Bastiaansen, Yang, Hagoort, 2011). Although the reason for the divergence remains unknown, It is possible that these two types of mismatches represent different underlying processing, as missing accentuation represents a new information without prosodic marker, thus might not be processed fully due to the insufficient attention; while superfluous accentuation represents a given information being prosodically highlighted, thus the listeners may pay extra attention to it or even attempt to extract additional new information from it. The failure of such process (as there is no new information to be extracted) may induce enhanced cognitive load. In an fMRI study investigating the localization of the accentuation-information status association (van Leeuwen et al., 2014), participants were presented with sentences in which the prosodic accentuation matches or mismatches the information status (both missing and superfluous accentuation included). The authors found that the mismatch condition additionally activated two distinct parts of the left IFG (namely

posterior and anterior ventral LIFG), which the authors interpreted as respectively relevant for the extraction of information structures based on prosodic features and for the unification of information status and prosodic features (van Leeuwen et al., 2014). More generally, the findings that prosodic accentuation (in association with information status) activated left hemisphere supported that prosodic accentuation affects the meaning comprehension of speech.

To sum up, behavioural, electrophysiological and imaging studies suggest that prosodic information are taken into account during language comprehension in the initial access of lexical information, the subsequent syntactic analysis and the identification of new/old information in the discourse. Prosodic stress enhances the saliency of certain words, making them more recognisable (e.g. Cutler & Foss, 1977) and their poor fit to semantic context even more difficult to process (indexed by even larger N400, e.g. Li & Ren, 2012), possibly via the activation of domain general attention networks (Kristensen, Wang, Petersson & Hagoort, 2012). On the sentence level, prosodic boundaries can mark syntactic structures, with appropriate prosodic boundaries facilitating both syntactic segmentation (e.g. Watson & Gibson, 2005) and the resolving of ambiguity (e.g. Snedeker & Casserly, 2010). Whereas, mismatch between prosodic and syntactic boundaries triggers larger N400 and P600 (e.g. Nickels, & Steinhauer, 2018), representing enhanced semantic and syntactic processing difficulty. In the discourse level, prosodic accentuation marks newness of information (e.g. Cruttenden, 2006). Inappropriate pairing of prosodic patterns and information status induces larger N400 (e.g. Magne et al., 2005) and activates inferior frontal cortices (van Leeuwen et al., 2014), indicating enhanced difficulty when integrating prosodic and semantic information. These findings showed that prosodic cues are involved in different stages of language comprehension. As

prosodic changes always accompany speech, such modulations continuously shape how listeners process every single utterance in face-to-face conversations. Therefore, prosody should be regarded as central to naturalistic language comprehension in the real world.

However, to note, these previous studies primarily employed mismatch design. Therefore, while it is clear that the disruption of prosodic patterns increase processing difficulty, a lot less is known about whether and how naturally occurring prosody (which are primarily congruent with the context) aids the processing of language on different levels in the real world.

2.2.2 Gestures

There has been no report of a culture lacking co-speech gestures (Kita, 2009). Gestures complement spoken language, as they can convey pictures and thoughts that are difficult to express by language itself (McNeill, 1992). For example, iconic gestures imitate the actions or objects referred in the speech (e.g. “drawing” - hand shaped as if holding a pencil and moved around) , thus complementing speech by conveying direct sensorimotor properties of the referent. Deictic gestures point directly to the referent (e.g. the pointing towards their own hair while saying “hair”), therefore leading listeners’ attention to the referent in the physical context. Beat gestures are not directly meaningful, but instead follow the rhythm of the speech and can highlight specific parts of the speech (e.g. vertical hand movements with flat palm when stressing something). Although other types of gestures also contribute to communication (such as emblematic gestures, which represent meaning by convention, e.g. “OK” - form a circle with thumb and index fingers while stretching out the remaining fingers; metaphoric gestures, which represent semantic properties of abstract ideas metaphorically, e.g. expressing the importance of an idea – holding

and weighing a large object; pragmatic gestures, which serve interactive functions, e.g. inviting another person to take the turn – moving a hand to the person with palm facing upwards), iconic, deictic and beat gestures received more attention in previous studies. Therefore, below I will review separately these different types of gestures.

2.2.2.1 Iconic gestures

Iconic gestures imitate concrete actions or objects and are typically related to the semantic content of the speech. Moreover, corpus studies found that the majority of gestures precedes the occurrence of its lexical affiliate (Donnellan et al., in prep; Ter Bekke, Drijvers & Holler, 2020), thus gestures have the potential to facilitate predictive language processing. Therefore, it comes as no surprise that researchers found iconic gestures to facilitate language comprehension and improve behavioural responses (see meta-analysis of behavioural studies in Hostetter, 2011; Dargue, Sweller, & Jones, 2019). Listeners are sensitive to iconic gestures and extract semantic information from them in addition to speech.

In behavioural studies, gestures have been shown to affect comprehenders' interpretation and memory of speech, measured by the content recalled, as well as accuracy and response times. For example, in a series of studies, adults were asked to watch children's conversations and identify the messages of their conversations and guess their reasonings. They found that adults incorporated information conveyed ONLY through children's gestures in their responses, indicating that comprehenders draw on resources from interlocutors' gestures when identifying the message of conversations (Goldin-Meadow, Wein, & Chang, 1992; Goldin-Meadow & Sandhofer, 1999). Conversely, when speech is accompanied by mismatching iconic gestures, comprehension is impaired as listeners' recall accuracy of semantic

information from speech has been shown to drop (Kelly & Church, 1998; McNeill, Cassell, & McCullough, 1994). Interviewers' misleading gestures can even induce children to report incorrect details about events they witnessed (Broaders & Goldin-Meadow, 2010). These findings together provided strong evidence that iconic gestures are integrated with speech in comprehension without the need of explicit instruction, and even when the gestures is not actually helpful (such as in the case of mismatching or misleading gestures). Apart from being incorporated with speech, iconic gestures also enhance the memory of information conveyed in speech, as words and sentences presented with iconic gestures are more likely to be recalled in later memory tests than those without (e.g. Cohen and Otterbein, 1992; Church, Garber & Rogalski, 2007; but see, Dargue & Sweller, 2018, which reported that only typical but not atypical iconic gestures facilitated recall). Teaching sessions where the teacher produced iconic gestures also yield better learning outcomes (symmetry: Valenzeno, Alibali, & Klatzky, 2003; foreign language: Tellier, 2008; Kelly, McDevitt & Esch, 2009; Macedonia & Klimesch, 2014; Andrä et al., 2020; Sweller, Shinooka-Phelan, & Austin, 2020). The improvement of learning associated with iconic gestures can last long after the training, as was shown by Macedonia and Klimesch (2014) that students who learnt second language words with live gestures from the teacher showed better memory of these words even 14 weeks and 14 months after training, indicating that iconic gestures enhance the memory trace of new words, making it more resilient against decay.

In electrophysiological studies, matching/mismatching iconic gestures have been shown to affect the N400 component, a biomarker of cognitive load in semantic processing (Kutas & Federmeier, 2011). Words accompanied by unrelated or incongruent gestures have been found to elicit more negative N400 than

related/congruent gestures (see review in Özyürek, 2014). This pattern has been reliably observed both with linguistic context (in a sentence or story, e.g. Özyürek et al., 2007) and without (in a single word, e.g. Wu & Coulson, 2005; Kelly, Kravitz & Hopkins, 2004; Kelly, Ward, Creigh, & Bartolotti, 2007; Bernardis, Salillas, & Caramelli, 2008), and across different groups (adults: e.g. Kelly, Kravitz & Hopkins, 2004; children: e.g. Sekine et al., 2020; non-native adults: Drijvers & Özyürek, 2018), thus providing strong evidence that comprehenders combine linguistic and gestural information in processing. Iconic gestures can also be used to disambiguate the current word and restrict predictions of the upcoming words: across a series of studies, participants were presented with sentences containing an ambiguous word, which was only disambiguated later with the appearance of the target word (e.g. “she controlled the ball … during the dance/game”, the ambiguous word “ball” is only disambiguated when target word “dance” or ”game” occur, confirming one of the meanings). The presence of iconic gestures supporting the subordinate meaning of the ambiguous word (e.g. “ball” – gesture as if holding someone while dancing) was found to reduce the N400 for the later target word confirming the subordinate meaning (e.g. “dance”, instead of “game”), indexing easier semantic comprehension (Holle & Gunter, 2007; Obermeier, Holle & Gunter, 2011; Obermeier, Dolk & Gunter, 2012; Obermeier, Kelly & Gunter, 2015). The findings suggest that listeners combine gestures and speech to construct interpretations of the speech and restrict their expectation of upcoming information based on both gestural and linguistic information.

Neuroimaging studies have addressed the questions of whether speech and iconic gesture processing overlaps and if so where (e.g. Holle et al., 2008; Green et al., 2009; Holle, Obleser, Rueschemeyer & Gunter, 2010; Straube et al., 2010;

Straube, Green, Weis & Kircher, 2012; Andric et al., 2013; Dick et al., 2014; He et al., 2018; Straube, Wroblewski, Jansen & He, 2018; see meta-analysis in Marstaller & Burianová, 2014; Yang, Andric and Mathew, 2015). Across these studies, the S/MTG and IFG are typically found to be activated by iconic gestures, and indeed disruption of these areas impairs gesture-speech integration (Zhao, Riggs, Schindler, Holle, 2018). As both areas are widely associated with semantic processing, this pattern across fMRI studies supports that iconic gestures contribute to speech comprehension. Further, Skipper and colleagues found that when participants watch videos of speech but also with iconic gestures (compared with grooming gestures or without gestures), their Broca's area (pars triangularis and pars opercularis of the IFG) exerts the least influence on other areas. Therefore, they argued that iconic gestures reduced the cognitive load of linguistic processing, leading to a decreased need for Broca's area to select and retrieve semantic information (Skipper, Goldin Meadow, Nusbaum, & Small, 2007).

2.2.2.2 Deictic gestures

Deictic gestures are used in every culture around the world to establish joint attention between interlocutors and link the speech with the referent (Kita, 2003). Apart from its importance in the early development of language in the initial grounding of lexical items (e.g. Thompson & Massaro, 1994; Morissette, Ricard & Décarie, 1995; Behne, Liszkowski, Carpenter, & Tomasello, 2012), deictic gestures have also been found to facilitate language comprehension. Concrete deictic gestures, or pointing gestures that refer to the physical existence of a referent, have been shown to improve the recall of information (Cameron & Xu, 2011; Macoun & Sweller, 2016; Austen & Sweller, 2017) and the learning outcome in general (Symmetry: Valenzeno, Alibali, & Klatzky, 2003; weather phenomena and historical

events: Beege et al., 2020) just like iconic gestures. Similarly, the incongruent use of pointing gestures also induces a larger N400 (Steven & Zhang, 2013, 2014), suggesting that listeners incorporate the referential information from concrete deictic gestures with speech during language comprehension. Unlike concrete deictic gestures, abstract deictic gestures assign a particular spatial area to a referent that is not physically present (e.g. “do you prefer chocolate or strawberry flavour?” – pointing left for “chocolate” and right for “strawberry” but without the actual presence of these objects). Indeed, around 35% of the gestures are produced in the same spatial location previous assigned to a referent (So, Kita & Goldin-Meadow, 2009), and the cohesive use of abstract pointing gestures has been shown to facilitate the interpretation of narratives (Sekine & Kita, 2015), whereas inconsistent use of abstract deictic gestures (e.g. a speaker who previously assigned left for referent A “chocolate” uttered A with pointing to the right) has been shown to impair comprehension as indexed by larger N400 (Gunter, Weinbrenner & Holle, 2015; Gunter & Weinbrenner, 2017). An fMRI study found that a mismatch between speech and deictic gestures activates the left IFG and posterior MTG, interpreted as semantic integration of gestural and speech signal (Peeters, Snijders, Hagoort & Ozyurek, 2017).

2.2.2.3 Beat gestures

Compared with iconic or deictic gestures, beat gestures are more rhythmic and do not convey meaning by themselves. Some researchers argued that beat gestures serve a similar function as prosodic accentuation (e.g. Krahmer & Swerts, 2007; Hubbard, Wilson, Callan & Dapretto, 2009). Indeed, words accompanied by beat gestures are also perceived as being more salient (similar to prosodic stress, e.g. Krahmer & Swerts, 2007). Further, the presence of beat gestures can induce

illusions of prosodic stress: in pseudowords consisting of two syllables (e.g. wasol), the syllable accompanied with beat gestures is more likely to be perceived as stressed (e.g. if “sol” is accompanied with beat gestures, the word is more likely to be perceived as waSOL, Bosker & Peeters, 2021). However, while prosodic stress has been found to robustly modulate comprehension, namely the initial recognition of words (e.g. Cutler & Foss, 1977), the processing of the congruency of a word (e.g. Li & Ren 2012) as well as the information status of a word given previous linguistic context (e.g. Cruttenden, 2006, Magne et al., 2005), studies on beat gestures provided little evidence regarding whether beat gestures also serves these functions.

Some behavioural studies reported that beat gestures improve information recall, although the results has been mixed. Different studies found that beat gestures enhances memory exclusively in adults but not children (words: So, Sim Chen-Hui & Low Wei-Shan, 2012), exclusively in children but not adults (spatial information: Austin & Sweller, 2014), no facilitation at all (in sentences/narratives, adults: Feyereisen, 2006, Beege et al., 2020; children: Macoun & Sweller, 2016), or even negative effect on selected groups (namely non-native adult speakers when presented with narratives, Rohrer, Delais-Roussarie & Prieto, 2020). These differences suggested that the effect of beat gestures may be modulated by many factors. Igualada and colleagues (2017) argued that beat gestures promote word learning by making words more salient. However, the function of beat gestures is strictly local: beat gestures only facilitate the recall of the single words that co-occur with them and thus are highlighted. In storytelling, they found that children recalled the words in the story directly accompanied by a beat gesture better than without beat gestures. In contrast, the memory of the adjacent non-target words did not differ. In support of this argument, previous studies reporting a positive effect for beat

gestures usually measure recall of single words (So, Sim Chen-Hui & Low Wei-Shan, 2012; Kushch & Prieto, 2016). Whereas studies reporting null effects typically test for sentence or discourse level information (Feyereisen, 2006; Macoun & Sweller, 2016; Beege et al., 2020). However, to note, Austin and Sweller (2014) found that 3 to 4-years-old children remember spatial route better when beat gestures were presented in the teaching sessions, which challenges the theory that the effect of beat gestures is strictly local. Further, as Igualada and colleagues (2017) only tested children 3-5 years old, the pattern may only represent a phase during language development and therefore may not be generalisable towards all groups of comprehenders. Rohrer, Delais-Roussarie & Prieto (2020) further found that naturally produced beat gestures (more frequent and more continuous, instead of experimentally manipulated beat gestures that involve one clear stroke only, used in e.g. Kushch & Prieto, 2016) did not impact recall of longer narratives in native comprehenders, and even induced worse performance in non-native comprehenders. The authors reasoned that more natural beat gestures may have a different effect with the artificial single stroke beat gestures, as the more natural beats are more continuous and less visually salient. This may further contribute to the divergent results of previous studies, as some studies presented participants with manipulated beat gestures (e.g. So, Sim Chen-Hui & Low Wei-Shan, 2012) while some presented participants with educational materials, where the beat gestures may be more continuous (e.g. Beege et al., 2020).

EEG studies of the neural time course of beat gestures also provide mixed results. Some studies found that beat gestures modulated early EEG signals, indicating a modulation of early sensory processing, although the exact EEG components differ across studies (Biau & Soto-Faraco, 2013: P200; Biau, Fromont &

Soto-Faraco, 2018: N100 & P200; Dimitrova et al., 2016: P300). Two studies reported that words produced with beat gestures induced less negative N400, similarly to meaningful gestures (Wang & Chu, 2013; Morett, Landi, Irwin & McPartland, 2020), which was interpreted as beat gestures highlighting semantic information, making it easier to process. Finally, some studies found a modulation of beat gestures at approximately 600-900ms post-stimulus window: Holle and colleagues (2012) found that presence of beat gestures reduced the P600 associated with less preferred syntactic structure (Object – Subject – Verb in German, compared with Subject – Object – Verb, which is more common in German), while Biau, Fromont & Soto-Faraco (2018) found that in sentences with a relative clause (e.g. "Someone shot the servant $_{NP1}$ of the actress $_{NP2}$ who was on the balcony $_{RC}$", in this sentence both nouns, NP1 "the servant" and NP2 "the actress", may be attached to the relative clause, or RC), beat gestures can reduce P600 for both NP1 and NP2. As P600 is regarded as a marker of syntactic processing (Coulson, King, Kutas, 1998), these findings indicate that beat gestures can modulate processing of syntactic information. Dimitrova and colleagues (2016) also reported a modulation of beats on the 600-900ms window post word onset. They presented participants with short sentences in which the target word is either in the focus position or not (induced by prior question, e.g. for sentence "She received an **email** from the teacher" with target word "**email**", a question of "Did she receive an email or a letter from the teacher" would put "**email**" in the focused position, while a question of "Did she receive an email from the teacher or the rector" would put "**email**" in the non-focused position), either with or without beat gestures. They found that when beat gestures are present, non-focused target words (but not the focused target words) elicited more positive electrophysiological signal 600-900ms after their

onset. As no syntactic manipulation is included in the experiment, they interpret this signal as late positivity complex (LPC), reflecting the enhanced meta-cognitive processing load when integrating the speech and gestures with the general context. The authors argued that beat gestures are regarded as non-verbal cues for focus, and the mismatch resulted from the presence of beats and lack of focus resulted in enhanced meta-cognitive processing difficulty. Only two imaging studies so far have investigated beat gestures (Hubbard, Wilson, Callan & Dapretto, 2009; Biau et al., 2016). Hubbard and colleagues (2009) reported that speech accompanied by beat gestures showed increased activation in auditory cortex bilaterally and the right planum temporale, which is interpreted as beat being processed in a similar way as prosody. Biau and colleagues (2016) further found that speech synchronised with beats showed increased activation in left MTG and IFG compared with the asynchronous pairing of beat gestures. The authors argued that beat gestures might convey communicative intent, which generates linguistic value and activates semantic areas.

To sum up, behavioural, electrophysiological and imaging work suggests that different types of gestures contribute to language comprehension. Iconic gestures are incorporated with speech to make sense of it without the need of explicit instruction (e.g. Goldin-Meadow, Wein, & Chang, 1992), and additional iconic gestures improves memory of the speech immediately after exposure as well as long term learning outcome of the topics (e.g. Cohen & Otterbein, 1992; Macedonia & Klimesch, 2014). In contrast, iconic gestures that mismatch the speech induce larger difficulty in semantic comprehension, indexed by larger N400 (e.g. Özyürek et al., 2007). Iconic gestures were found to activate traditionally language related areas including medial temporal gyrus and inferior frontal gyrus (e.g. Marstaller &

Burianová, 2014; Yang, Andric and Mathew, 2015), which further supports that iconic gestures contributes to semantic comprehension in general. Deictic gestures, similar to iconic gestures, is linked with improved recall of specific information (e.g. Cameron & Xu, 2011) as well as better learning outcome in general (e.g. Valenzeno, Alibali, & Klatzky, 2003). When mismatching with the context (concrete deictic gesture pointing at the wrong object, or abstract deictic gesture used inconsistently), deictic gestures impair comprehension, indexed by larger N400. Deictic gestures were also found to activate inferior frontal and medial temporal areas (Peeters, Snijders, Hagoort & Ozyurek, 2017), indicating that deictic gestures are integrated with the speech signal in comprehension. Studies on beat gestures, on the other hand, provided mixed results. While some studies suggested that beat gestures also improve memory (e.g. So, Sim Chen-Hui & Low Wei-Shan, 2012; Austin & Sweller, 2014), other studies reported no such effects (e.g. Feyereisen, 2006; Macoun & Sweller, 2016). Beat gestures were also found to activate different EEG signals, including early N100, P200 or P300 associated with earlier sensory processing of words (e.g. Biau, Fromont & Soto-Faraco, 2018), N400 associated with semantic processing (e.g. Wang & Chu, 2013) as well as P600/LPC associated with syntactic or meta-cognitive processing (e.g. Holle et al., 2012; Dimitrova et al., 2016). One possible explanation for the diverging results is that beat gesture serves a more general function - increasing word prominence - which may affect different processing stages based on the exact context, in contrast with iconic or deictic gestures, which have a clear function (i.e. providing semantic or referential information) and almost exclusively affect semantic processing of language. Another possibility is that beat gestures may contain different sub-types (e.g. more sharp and clear stroke v.s. more continuous rhythmic movements, see Prieto, Cravotta,

Kushch, Rohere & Vilà-Giménez, 2018), which may serve different functions (e.g. Rohrer, Delais-Roussarie & Prieto, 2020).

Although there is strong evidence across the board suggesting that each type of gesture modulates language comprehension (although some gestures, such as iconic gestures, may have a larger effect), listeners may be able to dynamically adjust the weight placed on gestures based on how useful they are in general. Listeners make more use of gestures when there a lot to gain, such as when linguistic information is challenging (for second language listeners with lower proficiency, Sueyoshi & Hardison, 2005; or for children comprehending difficult message, McNeil, Alibali, & Evans, 2000). Conversely, when gestures provide little additional information, listeners make less use of them, such as when the gestures are believed to be unintentional (produced by different individuals, Kelly et al., 2007; not directly facing the listener, He et al., 2020) or unreliable (involve many grooming gestures along side iconic gestures, Holle & Gunter, 2007; Obermeier, Kelly & Gunter, 2015; using abstract deictic gestures in an inconsistent style, e.g. when expressing two objects A and B, sometimes point left for A but sometimes point right, Gunter & Weinbrenner, 2017). Apart from the usefulness of the gesture, listeners' use of gestural information is also affected by the accessibility of such information: typical but not atypical iconic gestures facilitate narrative comprehension (Dargue & Sweller, 2020), as typical gestures are more easily interpretable; similarily, synchronous but not asynchronous gestures are incorporated with speech, as asynchronous gestures need to be stored in working memory in order to be incorporated (Habets et al., 2011; Obermeier et al., 2011). Indeed, to what extent listeners rely on gestures may be decided by actively balancing the gains (how useful gestures are) and costs (how difficult the gestures are to interpret) of using

gestural information. For example, although asynchronous gestures are not integrated with speech automatically; they can nonetheless be integrated under explicit task instruction (Obermeier et al., 2011) or when the vocal signal is less clear, making gestural information more important (Obermeier et al., 2012). To note, in naturalistic face-to-face conversations, listeners are likely to see a mixture of different types of gestures in dynamically changing context, which can modify the gains and costs of incorporating gestural information. In contrast, previous studies typically investigate each type of gesture individually, therefore artificially minimizing the inherent variance of informativeness of gestures, potentially enlarging the effect of gestures. Moreover, previous studies usually restrict the linguistic and multimodal context of the speech (e.g. presenting single words or short sentences with limited linguistic context, or presenting gestures while hiding other visual cues), which potentially enhances the attention to gestural information. Therefore, it remains to be asked to what extent are listeners relying on gestural information in real-world conversations.

2.2.3 Mouth movements

Mouth movements always go hand-in-hand with speech. The mouth movement patterns change along with the production of every syllable, providing dynamic perceptual information that can help listeners identify the produced sounds. Indeed, the famous McGurk effect (e.g. sound /ba/ accompanied by mouth movement /ga/ can induce the perception of /da/, McGurk & MacDonald, 1976) suggests that visual information is powerful enough to mislead sound perception. This section will focus exclusively on language comprehension, evaluating whether mouth movements facilitate comprehension based on behavioural, electrophysiological and imaging studies.

As mouth movement carries visual information that can disambiguate phonological information from speech, it comes as no surprise that many studies reported a facilitatory effect of mouth information when perceiving speech. Audiovisual speech, which contains mouth/facial information apart from phonological information, can be recognised more accurately than auditory-only speech (e.g. Sumby and Pollack, 1954; Munhall et al., 2004; Tye-Murray, Sommers, & Spehar, 2007). This is true even at larger noise levels (e.g. Macleod & Summerfield, 1987; Grant & Seitz, 2000; Bidelman, Brown, Mankel & Price, 2020). Additional mouth information facilitates word recognition because it can convey information that is not available from auditory signals, such as the place of articulation. This visual information can narrow down competitor phonemes and exclude confusable neighbours, resulting in more accurate perceptions (e.g. Mattys, Bernstein & Auer, 2002; Cappelletta and Harte, 2012). Apart from new information unavailable from speech, mouth movements can also provide information that is redundant with the auditory signal. A recent study indicated mouth and facial areas contain enough information to reconstruct the articulation of speech (i.e. vocal tract movements, Scholes, Skipper & Johnson, 2020). This cross-modal redundancy enhances the reliability of the language signal, which can be especially helpful in noise (see review in Massaro & Jesse, 2007). Finally, as mouth movements are typically 150-300 ms earlier than the auditory signal in natural speech (Chandrasekaran et al., 2009), it can provide a visual head start, which can be used to predict upcoming sounds (see review in Peelle and Sommers, 2015). For example, participants benefit more from mouth movements in those words where the mouth starts first (e.g. “drive”, where the mouth shape for the phoneme /d/ start before the actual sound), compared with

words where the sound starts earlier or simultaneous with the mouth movements (e.g. “known”, Karas et al., 2019).

In line with behavioural work studies, electrophysiological studies show that mouth movements are integrated with speech sounds as quick as 100-200ms post word onset. Audiovisual stimuli (containing mouth movements) elicit earlier or smaller N100 and P200 than auditory-only stimuli, termed N1/P2 effect (see review in Pilling, 2009). The N1/P2 effect has been shown to be robust whenever auditory and visual information is integrated, both when the mouth correctly predicts the upcoming sounds (e.g. Besle, Fort, Delpuech & Giard, 2004; van Wassenhove et al., 2005; see review in Pilling, 2009), and in the McGurk paradigm where mouth information created auditory illusion (e.g. Alsius et al., 2014; Baart et al., 2014; Knowland et al., 2014). Therefore, electrophysiological evidence suggests that mouth movements are immediately integrated with speech, as the N1/P2 pattern indexes domain-general audiovisual integration (activated both for language and non-speech events, e.g. cutting a tree motion + sound, Stekelenburg & Vroomen, 2007; 2010; 2012)

In the brain imaging literature, superior temporal sulcus has been associated with domain-general audiovisual integration, similar with the N1/P2, as this area has also been reliably associated with multisensory events (both language and non-speech events, e.g. face + voice and tools + sound, Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Bernstein et al., 2011). In the context of language comprehension, studies found mouth movements activate superior temporal sulcus (STS) and STG (e.g. Calvert, Campbell, & Brammer, 2000; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003; Bernstein, Auer, Wagner & Ponton, 2007; Nath and Beauchamp, 2011), suggesting its reliable integration with auditory speech. Further,

mouth movements can activate the auditory cortices without sound (e.g. Calvert et al., 1997; MacSweeney et al., 2000; Calvert and Campbell, 2003; Karas et al., 2019). For example, in an intracranial EEG study, Besle and colleagues (2008) found that mouth movements activated secondary auditory areas shortly after the activation of the visual motion area, suggesting that mouth movements are used in the initial phonological processing. They also found that the auditory cortex shows less activation when participants simultaneously hear the syllables and see the corresponding mouth movements, indicating that converging information from the auditory and visual channel reduces the cognitive load of sound processing. Audio-visual speech also activate brain regions associated with language production (e.g. motor areas and IFG, Calvert, Campbell, & Brammer, 2000; Hall, Fussell, & Summerfield, 2005; Skipper et al., 2005, 2007; Fridriksson et al., 2008). Therefore, some researchers argue that mouth movements facilitate language perception through the simulation of production. For example, Skipper, Nusbaum and Small (2005) found that audiovisual speech containing mouth movement activated a distributed network associated with language production (i.e. STS/G, the pars opercularis, primary motor and premotor cortex, and the cerebellum). They argued that the interpretation of mouth movement as phonetic information is based on the motor command that can produce the corresponding sound.

Apart from the large number of studies reporting a facilitatory effect of mouth movements in language perception, some studies also found that the effects of mouth movements interact with the semantic properties of words, therefore speculate that mouth movements might affect semantic processing. Some behavioural studies indicate that the facilitatory effect of mouth movements may interact with lexical level variables. For example, it is well established that an

auditory target words primed with related mouth movements are processed faster/more accurately (Dodd, Oerlemans, & Robinson, 1988; Kim, Davis & Krins , 2004; Buchwald, Winters, & Pisoni, 2009). However, this cross-modal priming effect is larger for low-frequency words (Fort et al., 2013). Similarly, mouth movements have also been shown to provide a larger facilitatory effect when recognizing meaningful sentences, compared with the anomalous ones (e.g. "The hot sun warmed the ground" v.s. "The green week did the page", Van Engen et al., 2014), indicating that linguistic context and additional mouth movements jointly contributed to the recognition of words. Finally, additional mouth movements were found to provide larger facilitatory effect when the speech is syntactically or semantically more challenging (Arnold & Hill, 2001). Taken together, these findings suggest that the impact of mouth movements on the sensory processing of a word may depend its linguistic properties, being larger for words that are meaningful in the context (Van Engen et al., 2014) but difficult to process (due to lower frequency, e.g. Fort et al., 2013; or more complex syntactic/semantic properties, Arnold & Hill, 2001). Some EEG studies, on the other hand, reported mouth modulating later event-related potentials (e.g. N400 or LPC), suggesting that mouth movements contribute to comprehension beyond perception level. However, the results remain inconsistent. One study compared auditory with audiovisual continuous speech, where the target word has different mouth informativeness (e.g. words starting with /p/ is more informative than /k/), and found that audiovisual target words elicited more negative N400 when their mouth informativeness is high (Brunellière et al., 2013). Whereas another study did not find any N400 differences when comparing the static picture of a face with one that contains dynamic mouth movements. Instead, they found that the mouth movement condition elicited a longer-lasting late posterior positivity,

associated with meta-cognitive processing such as the unification of information into the context (Hernández-Gutiérrez et al., 2018). Such impact of mouth movements on semantic processing may be mediated by improved acoustic processing: an fMRI study by McGettigan and colleagues (2012) presented participants with videos of spoken sentences differing in their semantic predictabilities (showing only face area), but further manipulated the clarities of the auditory and visual signals. They reported a significant interaction between auditory and visual clarity (in left supramarginal gyrus) and between auditory clarity and linguistic predictability (in left supplementary motor area and cuneus), such that enhanced visual clarity and linguistic predictability showed largest effect on the brain activations when acoustic signal is moderately degraded, potentially because here visual and contextual linguistic information is useful (unlike in clearer speech) but also interpretable (unlike in further degraded speech). However, they did not find any area sensitive to the interaction between visual clarity and linguistic predictability, so that mouth and facial information may not directly modulate semantic processing. This is also in line with the hypothesis that mouth movements improves language comprehension by simulating the production process, which is most directly acoustic (e.g. Skipper et al., 2005).

To sum up, there is clear evidence that mouth movements support phonological processing by providing visual information that is helpful for the identification of sound before when the sound is available (e.g. Karas et al., 2019). It is clear that mouth movements are integrated with phonological processing very early as indicated by the robust finding of N100/P200 associated to audiovisual presentation (e.g. van Wassenhove et al., 2005). fMRI studies further suggest that visible mouth movement activate cortical areas linked to both auditory processing and speech production, suggesting that mouth movements and speech are

intertwined (e.g. Skipper et al., 2005). However, it is unclear whether mouth movements contribute to processing beyond the phonological level meaning and modulate later event-related potentials. Some studies suggested that the linguistic context affect the facilitatory effect of mouth movements in recognition (e.g. Fort et al., 2014), indicating that the processing of mouth draws on resources from the contextual information. Some EEG studies further found that the presence and informativeness of mouth movements modulates N400 or LPC (e.g. Brunellière et al., 2013; Hernández-Gutiérrez et al., 2018), indicating that mouth movements may contribute to the semantic and meta-cognitive processing of a word, although such effect may be mediated by acoustic processing (e.g. McGettigan et al., 2012). Despite the small number of studies and mixed results, it is not inconceivable that mouth movements contribute more than the recognition of words, as easier perceptual processing may potentially develop into quicker semantic access and integration with the context.

2.2.4 Summary and conclusions: Are multimodal cues central to language comprehension?

The previous sections reviewed behavioural, electrophysiological and imaging studies of how each multimodal cue modulates language comprehension. The results support the claim that each of these cues contributes to how listeners process speech input. Mouth information and prosodic stress modify the intelligibility of speech signals, facilitating the initial perception of language. Semantic and referential information carried by speakers' gestures can then support semantic and pragmatic processing. With the help of prosodic structures, the word information is segmented and combined to build meaningful sentences. Throughout the comprehension of the entire discourse, listeners assign importance to each piece of

linguistic information based on the guidance by speakers' gestures and prosody, modulating their interpretation and the cognitive resources assigned. Regions identified central parts of the language network (Fedorenko & Thompson-Schill, 2014; Chai, Mattar, Blank, Fedorenko & Bassett, 2016), such as left inferior frontal areas and left superior temporal areas, are activated by multimodal cues, indicating that language processing carried out in these areas draws on multimodal resources. Based on the evidence provided above, we argue that multimodal cues are central to language comprehension. As language is most often used in the face-to-face setting rich with multimodal cues, comprehending the speech while drawing resources from prosody, gestures, and mouth movements should be considered the "default" mode of language processing. In contrast, the processing of text, divorced from the rich multimodal context should be considered as representing only a case, perhaps a special case, of language processing.

If multimodal cues are central to language comprehension, one should not neglect the fact that they constantly co-occur in the real world. Despite their fruitful outcomes, previous studies of multimodal cues primarily employed a reductionist approach, in which one cue was manipulated while the others were elimintated. Therefore, studies of prosody typically do not include any visual cue (e.g. Heim & Alter 2006), studies of gestures usually paint the face black or block the mouth areas (e.g. Holler & Gunter, 2007), and studies of mouth movements usually only show the face of the speaker (e.g. van Wassenhove et al., 2005). However, this (e.g. see someone's hands but not mouth or gaze) is not how we comprehend language in the real world. In naturalistic language processing in which the different cues co-occur, the processing of one cue can be affected by the availability and reliability of other cues present (Bidelman, Brown, Mankel & Price, 2020; Fourtassi & Frank, 2020).

Therefore, these experimental manipulations may not reflect how multimodal cues are processed in the real world when all cues co-occur. The next section will review the existing studies in which more than one cue co-occur and evaluate whether their co-occurrence affects language comprehension.

2.3 Are comprehenders sensitive to the co-occurence of multimodal cues?

Multimodal cues generally co-occur with each other in daily conversations and they are correlated in a number of ways. For example, it has been shown that motor movements, including both gestures and mouth openings, are correlated with prosodic features (e.g. McNeill, 1992; Esteve-Gibert & Prieto, 2013; Ménard, Leclerc & Tiede, 2014). Visual cues, such as gestures and mouth movements, are perceived via the same visual channel and share the limited attentional resources (e.g. Gullberg & Holmqvist, 2006; Beattie, Webster & Ross, 2010). Does the presence of an additional cue introduce an even larger facilitatory effect than one cue alone (usually termed double-enhancement)? Do multimodal cues interact with each other, so that the additional cue not only contributes to comprehension but also modulates the effect of the other cue, enlarging or reducing it? A rather limited number of studies investigated these questions, mainly centring on the combination of prosody + beat gestures, or mouth movements + iconic gestures. This section will review these research areas before summarising and discussing whether comprehenders are sensitive to the co-occurrence of multimodal cues, which necessarily occur in real life.

2.3.1 Prosody and beat gestures

It has been long known that production of gestures and prosody is aligned (e.g. Birdwhistell, 1952). It was reported that infants as young as 11-19 months old align the stroke of their gestures with the onset of the prominent syllable in their

speech (Esteve-Gibert & Prieto, 2014). However, how comprehenders make use of such correlations is much less studied in comparison.

Beat gestures have been argued to represent “visual prosody” (e.g. Krahmer & Swerts, 2007; Wang & Chu, 2013; Hubbard et al., 2009), as beat gestures enhance the saliency of perceived speech similarly to prosodic accentuation. Some studies suggested that the presence of beat and prosody induce larger processing benefits than the presence of one cue alone, which is usually called “double enhancement”. In a word recall study, Llanes-Coromina and colleagues (2018) presented children with short stories. The target words are either with prosodic accentuation, beat gesture or both. They found that words with both cues are remembered better. Similarly, Kushch, Igualada, & Prieto (2018) found that adults learn second language vocabulary better when both cues are present, while the effect of beat gestures alone is smaller than prosodic accentuation alone. Further, studies found that participants’ pupil size was largest when presented with target words accompanied by prosodic accentuation and beat gestures, indicating an increase in cognitive load. However, beats alone lead to larger pupil size than prosody (Morett, Roche, Fraundorf and McPartland, 2020). The presence of double enhancement indicates that the redundant audiovisual information from prosody + beat gestures further benefits language comprehension, possibly by further highlighting the speech.

However, it remains unclear whether the effects of these cues are additive or interactive. Some behavioural and EEG studies reported a lack of interaction between the effect of beat and prosody despite the double enhancement (Eye-tracking: Morett, Fraundorf and McPartland, 2019; EEG: Wang & Chu, 2013). However, other evidence suggested that beat gestures and prosodic accentuation

are integrated. Asynchronization between the two cues elicits a larger N400, indicating enhanced processing difficulty (Morett, Landi, Irwin & McPartland, 2020). Further, some studies found that the presence of gestures modify how prosody is processed, indicating a potential interaction. Morett and Fraundorf (2019) found that when the speaker emphasised the target word with beat gestures only on some words, prosodic accentuation facilitated memory only when beat gestures were also present. Conversely, when the speaker never uses beat gestures, prosodic accentuation always enhances memory. They inferred that listeners consider words without beat gestures unimportant when the speaker uses beat gestures to emphasise some words. Therefore, these words were processed less deeply even when being accentuated prosodically.

Thus, overall these studies indicate that the co-occurrence of beat gestures and prosodic modulation affect how listeners process words. It leads to double enhancement, indexed by a larger facilitatory effect than one single cue. However, whether these two cues are processed independently or interactively remains under debate. While some studies suggest that their impacts are additive, others indicate that prosody and beat gestures are integrated, and the presence of one cue affects the processing of the other. Therefore, further studies are required to elucidate whether they are processed in parallel or interactively.

2.3.2 Mouth movements and gestures

Both mouth movements and gestures occur in the visual modality alongside speech, and therefore are constantly perceived simultaneously. However, we know little about their joint impact. One reason is that previous studies only show participants one cue while blocking the others (e.g. paint the head black when studying gestures, or cut the torso out when studying mouth movements). But even

when both cues co-occur in a study, the untargeted cue is usually kept constant, without being explicitly measured or manipulated (e.g. Kelly, Kravitz & Hopkins, 2004; Drijver, van der Plasc, Özyürek & Jensen, 2019; Dick, Goldin-Meadow, Hasson, Skipper & Small, 2009). This section will review the limited number of studies investigating the potential interaction between mouth movements and gestures (primarily iconic gestures).

When both mouth movements and gestures are present, do listeners show a greater benefit than when only one is present? Hirata and Kelly (2010) investigated whether mouth movements and beat gestures representing short/long features (e.g. quick stroke v.s. prolonged movement extending horizontally) jointly facilitate English speakers learning Japanese long/short vowels. They found that the presence of mouth movements help the identification of vowels, but the additional presence of gestures do not. Conversely, other studies found a double enhancement of mouth information and iconic gesture when identifying words in noise. Drijvers and Özyürek (2017) presented participants with short videos of an actress producing single words with mouth movements, iconic gestures or both, embedded in different noise levels. They found that participants' recall accuracies are highest when both cues are present, especially when the noise level is higher, Similar pattern was further found across different populations (non-native: Drijvers & Özyürek, 2020, Drijvers, Vaitonytė & Özyürek, 2019; older adults: Schubotz, Holler, Drijvers & Özyürek, 2019), indicating a robust double enhancement effect of mouth and iconic gestures when identifying words in noise. However, as both cues occupy the visual domain and therefore share listeners' visual attention, it is possible that a potential trade-off exists, leading listeners to pay greater attention to the one most relevant to the task. A recent study further quantified the informativeness of mouth movements and iconic

gestures per word, and found that higher informativeness of both mouth movements and gestures lead to better performance in picture-matching task (Krason, Fenton, Varley & Vigliocco, 2021). This study also reported a marginally significant interaction between iconic gestures and mouth movements, showing that higher mouth informative words are recognised faster but the effect is larger when there are no iconic gestures. This suggests a potential trade-off between visual cues and iconic gestures may be weighted more heavily by the listeners, as they provide direct semantic information, which is especially helpful for the picture-matching task. This hypothesis is supported by an fMRI study, which found that iconic gesture reduced the connectivity between pars opercularis and other motor and language areas (i.e. premotor and primary motor cortices, supramarginal gyrus and STS/G, Skipper, Goldin‐Meadow, Nusbaum and Small, 2007). This connectivity was assumed to represent the mouth-language link; therefore, the authors argued that this pattern indicates that iconic gestures reduced the reliance on mouth movements when processing speech.

2.3.3 Summary: Are listeners sensitive to the co-occurrence of multimodal cues?

Compared with studies on the impact of individual cues on language comprehension, only a few studies focus on their joint impact. However, these studies suggested that listeners benefit from the co-occurrence between these cues, such as prosody + beat gestures and mouth movements + iconic gestures. Thus, these findings indicate that the co-occurrence of multimodal cues also affects language comprehension.

However, apart from the studies above, whether and how listeners respond to the co-occurrence of other multimodal cues remains largely unknown. For example, apart from beat gestures, the production of prosody also closely correlates with

iconic gestures (e.g. Brentari, Marotta, Margherita, & Ott, 2013) and pointing gestures (e.g. Esteve-Gibert & Prieto, 2013). However, whether participants are sensitive to the joint impact between these cues and whether they are processed independently or interactively remains largely unexplored (but see Prieto, Borràs-Comes, Tubau and Espinal, 2013; Esteve-Gibert, Prieto, & Liszkowski, 2017, reporting that prosody is integrated with iconic gesture and pointing gestures respectively). For another instance, mouth movements are also perceived together with deictic and beat gestures. However, whether these gestures have the same effect as iconic gestures when co-occurring with mouth movements remains unclear (but see Hirata & Kelly, 2010). Finally, daily face-to-face conversations usually includes more than one or two cues, instead, prosody, gestures and mouth movements can all occur, which invites future study.

2.4 Language is multimodal: Implications and way forward

In the world that we live in, multimodal cues, such as prosody, gestures or mouth movements, always accompany linguistic signals in face-to-face communication (this is true even when we are socially distanced, as video meetings embed these cues nonetheless). Behavioural, electrophysiological and imaging works strongly suggested that these multimodal cues are central to language comprehension. Mouth movements and prosodic stress both facilitate the initial perceptual identification of speech signals. Gestures contribute to the semantic processing. In contrast, prosody guides successful syntactic segmentation and constructions of sentence structures. Finally, gestures and prosody both contribute to directing attentional resources to the new or important information in the speech. Moreover, these multimodal cues typically co-occur, and evidence pointed out that listeners make use of their co-occurrence, showing double enhancement or different

processing of each cue depending on the other cues present. For example, beat gestures and prosodic accentuation further enhance word saliency (e.g. Kushch, Igualada, & Prieto, 2018; Llanes-Coromina et al., 2018), while the neurological connectivity representing the processing of mouth movements is decreased with additional hand gestures (Skipper et al., 2007)

These different lines of inquiry call for a rethinking of the traditional view, deeming language as purely linguistic while all other multimodal information is "paralinguistic" or "non-linguistic" (see further discussion in Murgiano, Motamedi & Vigliocco, 2021). Language processing constantly consults the information from other cues. This is true even when experimenters manipulated the multimodal information to be misleading, such as prosodic patterns mismatching the actual information status (e.g. Dahan, Tanenhous & Chambers, 2002; Li, Hagoort, & Yang, 2008; Baumann & Schumacher, 2012), or iconic gestures introducing false information (Kelly & Church, 1998; Broaders & Goldin-Meadow, 2010). Moreover, listeners dynamically adjust how much weight to put on each cue based on other cues, both linguistic and multimodal, in the context. Gestures and mouth movements both show larger facilitatory effects for challenging linguistic information (McNeil, Alibali, & Evans, 2000; Sueyoshi & Hardison, 2005; Arnold & Hill, 2001; Fort et al., 2013), while the presence of beat gestures changes how prosody is processed (Morett and Fraundorf, 2019). The processing of linguistic and multimodal information are intertwined, allowing smooth communication in daily face-to-face conversations.

Moreover, recognising that language is multimodal would also imply a change of experimental design. The traditional reductionist approach has no doubt been fruitful. However, as multimodal cues and their co-occurrence are always present in

real-world comprehension, their removal may result in lower ecological validity. For example, studying comprehension based on written text without any multimodal cues may result in a different processing mechanism. Similarly, studying one individual multimodal cue while blocking other cues will hide all potential interactions and might introduce task-specific effects.

Therefore, this PhD thesis aims to investigate how multimodal cues individually and jointly modulate language comprehension in more naturalistic settings, where more than one cue co-occur. In order to maintain scientific rigour and avoid potential confound, co-occurring cues are either kept constant (but present), manipulated, or measured and accounted for in the statistical analysis. Chapter 3 introduces a corpus of mouth informativeness for English words that we built with a novel method. This corpus provides insight into the recognisability of mouth patterns in more naturalistic settings, but also serves as a building block for the following studies using this corpus to track mouth informativeness. Chapter 4 describes an EEG experiment and a subsequent replication where we investigated how multimodal cues dynamically affect language comprehension. As N400 has been associated with processing difficulty, arising due to semantic prediction and/or integration process (Kutas & Federmeier, 2011), we took N400 amplitude per word as a marker of cognitive load in comprehension, with more negative N400 indicating more difficult processing. We presented participants with videos with naturally occurring prosody, gestures and mouth, and measured their joint impact in language processing, namely N400 amplitude per word. Chapter 5 presents another EEG study using the same paradigm but testing non-native speakers, investigating whether and how multimodal cues modulate second language processing. Chapter 6 introduces an online study investigating whether and how gestures and prosody

jointly modulate the learning of new concepts. We manipulated the interaction between the two variables while keeping the naturalistic setting. Finally, Chapter 7 discusses the general findings and implications of all studies presented.

Reference

1. Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., & Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Frontiers in psychology*, *5*, 727.

2. Andrä, C., Mathias, B., Schwager, A., Macedonia, M., & von Kriegstein, K. (2020). Learning foreign language vocabulary with gestures and pictures enhances vocabulary memory for several months post-learning in eight-year-old school children. *Educational Psychology Review*, *32*(3), 815-850.

3. Andric, M., Solodkin, A., Buccino, G., Goldin-Meadow, S., Rizzolatti, G., & Small, S. L. (2013). Brain function overlaps when people observe emblems, speech, and grasping. *Neuropsychologia*, *51*(8), 1619-1629.

4. Arnold, J. E. (2008). THE BACON not the bacon: How children and adults understand accented and unaccented noun phrases. *Cognition*, *108*(1), 69-99.

5. Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339-355.

6. Austin, E. E., & Sweller, N. (2014). Presentation and production: The role of gesture in spatial communication. *Journal of experimental child psychology*, *122*, 92-103.

7. Austin, E. E., & Sweller, N. (2017). Getting to the elephants: Gesture and preschoolers' comprehension of route direction information. *Journal of experimental child psychology*, *163*, 1-14.

8. Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, *47*(1), 31-56.

9. Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*, 115-121.

10. Bader, M. (1998). Prosodic influences on reading syntactically ambiguous sentences. In *Reanalysis in sentence processing* (pp. 1-46). Springer, Dordrecht.

11. Baum, S. R., & Pell, M. D. (1999). The neural bases of prosody: Insights from lesion studies and neuroimaging. *Aphasiology*, *13*(8), 581-608.

12. Baumann, Stefan, and Petra B. Schumacher. "(De-) accentuation and the processing of information status: evidence from event-related brain potentials." *Language and speech* 55.3 (2012): 361-381.

13. Beattie, G., Webster, K., & Ross, J. (2010). The fixation and processing of the iconic gestures that accompany talk. *Journal of Language and Social Psychology*, *29*(2), 194-213.

14. Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature neuroscience*, *7*(11), 1190-1192.

15. Beege, M., Ninaus, M., Schneider, S., Nebel, S., Schlemmel, J., Weidenmüller, J., ... & Rey, G. D. (2020). Investigating the effects of beat and deictic gestures of a lecturer in educational videos. *Computers & Education*, *156*, 103955.

16. Behne, T., Liszkowski, U., Carpenter, M., & Tomasello, M. (2012). Twelve-month-olds' comprehension and production of pointing. *British Journal of Developmental Psychology*, *30*(3), 359-375.

17. Bernardis, P., Salillas, E., & Caramelli, N. (2008). Behavioural and neurophysiological evidence of semantic interaction between iconic gestures and words. *Cognitive Neuropsychology*, *25*(7-8), 1114-1128.

18. Bernstein, L. E., Auer Jr, E. T., Wagner, M., & Ponton, C. W. (2008). Spatiotemporal dynamics of audiovisual speech processing. *Neuroimage*, *39*(1), 423-435.

19. Bernstein, L. E., Jiang, J., Pantazis, D., Lu, Z. L., & Joshi, A. (2011). Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. *Human brain mapping*, *32*(10), 1660-1676.

20. Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *European journal of Neuroscience*, *20*(8), 2225-2234.

21. Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and language*, *124*(2), 143-152.

22. Biau, E., Fernández, L. M., Holle, H., Avila, C., & Soto-Faraco, S. (2016). Hand gestures as visual prosody: BOLD responses to audio–visual alignment are modulated by the communicative nature of the stimuli. *Neuroimage*, *132*, 129-137.

23. Biau, E., Fromont, L. A., & Soto-Faraco, S. (2018). Beat gestures and syntactic parsing: an ERP study. *Language Learning*, *68*, 102-126.

24. Bidelman, G. M., Brown, B., Mankel, K., & Price, C. N. (2020). Psychobiological responses reveal audiovisual noise differentially challenges speech recognition. *Ear and hearing*, *41*(2), 268.

25. Birch, S., & Clifton Jr, C. (1995). Focus, accent, and argument structure: Effects on language comprehension. *Language and speech*, *38*(4), 365-391.

26. Birdwhistell, R. L. (1952). *Introduction to kinesics: An annotation system for analysis of body motion and gesture*. Department of State, Foreign Service Institute.

27. Blumstein, S., & Cooper, W. E. (1974). Hemispheric processing of intonation contours. *Cortex*, *10*(2), 146-158.

28. Bock, J. K., & Mazzella, J. R. (1983). Intonational marking of given and new information: Some consequences for comprehension. *Memory & Cognition*, *11*(1), 64-76.

29. Bögels, S., Schriefers, H., Vonk, W., Chwilla, D. J., & Kerkhofs, R. (2010). The interplay between prosody and syntax in sentence processing: The case of subject-and object-control verbs. *Journal of Cognitive Neuroscience*, *22*(5), 1036-1053.

30. Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B*, *288*(1943), 20202419.

31. Brentari, D., Marotta, G., Margherita, I., & Ott, A. (2013). The interaction of pitch accent and gesture production in Italian and English. *Studi e saggi linguistici*, *51*(1), 83-101.

32. Broaders, S. C., & Goldin-Meadow, S. (2010). Truth is at hand: How gesture adds information during investigative interviews. *Psychological Science*, *21*(5), 623-628.

33. 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.

34. Buchwald, A. B., Winters, S. J., & Pisoni, D. B. (2009). Visual speech primes open-set recognition of spoken words. *Language and cognitive processes*, *24*(4), 580-610.

35. Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *Journal of cognitive neuroscience*, *15*(1), 57-70.

36. Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., ... & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *science*, *276*(5312), 593-596.

37. Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current biology*, *10*(11), 649-657.

38. Cameron, H., & Xu, X. (2011). Representational gesture, pointing gesture, and memory recall of preschool children. *Journal of Nonverbal Behavior*, *35*(2), 155-171.

39. Cappelletta, L., & Harte, N. (2012, February). Phoneme-to-viseme mapping for visual speech

40. Chai, L. R., Mattar, M. G., Blank, I. A., Fedorenko, E., & Bassett, D. S. (2016). Functional network dynamics of the language system. *Cerebral Cortex*, *26*(11), 4148-4159.

41. Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS computational biology*, *5*(7), e1000436.

42. Church, R. B., Garber, P., & Rogalski, K. (2007). The role of gesture in memory and social communication. *Gesture*, *7*(2), 137-158.

43. Cohen, R. L., & Otterbein, N. (1992). The mnemonic effect of speech gestures: Pantomimic and non-pantomimic gestures compared. *European Journal of Cognitive Psychology*, *4*(2), 113-139.

44. Cole, J. (2015). Prosody in context: a review. *Language, Cognition and Neuroscience*, *30*(1-2), 1-31.

45. Cole, R. A., & Jakimik, J. (1980). A model of speech perception. *Perception and production of fluent speech*, *133*(64), 133-42.

46. Cole, R. A., Jakimik, J., & Cooper, W. E. (1978). Perceptibility of phonetic features in fluent speech. *The Journal of the Acoustical Society of America*, *64*(1), 44-56.

47. Cruttenden, A. (2011). The de-accenting of given information: A cognitive universal?. In *Pragmatic organization of discourse in the languages of Europe* (pp. 311-356). De Gruyter Mouton.

48. Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, *20*(1), 1-10.

49. Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*(2), 292-314.

50. Dargue, N., & Sweller, N. (2018). Not all gestures are created equal: The effects of typical and atypical iconic gestures on narrative comprehension. *Journal of Nonverbal Behavior*, *42*(3), 327-345.

51. Dargue, N., & Sweller, N. (2020). Two hands and a tale: When gestures benefit adult narrative comprehension. *Learning and Instruction*, *68*, 101331.

52. Dargue, N., Sweller, N., & Jones, M. P. (2019). When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. *Psychological Bulletin, 145*(8), 765–784.

53. Dick, A. S., Goldin-Meadow, S., Hasson, U., Skipper, J. I., & Small, S. L. (2009). Co-speech gestures influence neural activity in brain regions associated with processing semantic information. *Human brain mapping*, *30*(11), 3509-3526.

54. Dick, A. S., Mok, E. H., Beharelle, A. R., Goldin-Meadow, S., & Small, S. L. (2014). Frontal and temporal contributions to understanding the iconic co-speech gestures that accompany speech. *Human brain mapping*, *35*(3), 900-917.

55. Dimitrova, D. V., Stowe, L. A., Redeker, G., & Hoeks, J. C. (2012). Less is not more: Neural responses to missing and superfluous accents in context. *Journal of cognitive neuroscience*, *24*(12), 2400-2418.

56. Dimitrova, D., Chu, M., Wang, L., Özyürek, A., & Hagoort, P. (2016). Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience*, *28*(9), 1255-1269.

57. Dodd, B., Oerlemans, M., & Robinson, R. (1988). Cross-Modal Effects in Repetition Priming: A Comparison of Lipread, Graphic, and Heard Stimuli. *Visible Language*, *22*(1), 58.

58. Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212-222.

59. 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*, 7-17.

60. Drijvers, L., & Özyürek, A. (2020). Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language and speech*, *63*(2), 209-220.

61. 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.

62. 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.

63. Eckstein, K., & Friederici, A. D. (2005). Late interaction of syntactic and prosodic processes in sentence comprehension as revealed by ERPs. *Cognitive Brain Research*, *25*(1), 130-143.

64. Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements.

65. Esteve-Gibert, N., & Prieto, P. (2014). Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Communication*, *57*, 301-316.

66. Esteve-Gibert, N., Prieto, P., & Liszkowski, U. (2017). Twelve-month-olds understand social intentions based on prosody and gesture shape. *Infancy*, *22*(1), 108-129.

67. Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in cognitive sciences*, *18*(3), 120-126.

68. Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of child language*, *16*(3), 477-501.

69. Feyereisen, P. (2006). How could gesture facilitate lexical access?. *Advances in Speech Language Pathology*, *8*(2), 128-133.

70. Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes*, *28*(8), 1207-1223.

71. Fourtassi, A., & Frank, M. C. (2020). How optimal is word recognition under multimodal uncertainty?. *Cognition*, *199*, 104092.

72. Fridriksson, J., Moss, J., Davis, B., Baylis, G. C., Bonilha, L., & Rorden, C. (2008). Motor speech perception modulates the cortical language areas. *Neuroimage*, *41*(2), 605-613.

73. Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 748-775.

74. Gandour, J., Dzemidzic, M., Wong, D., Lowe, M., Tong, Y., Hsieh, L., ... & Lurito, J. (2003). Temporal integration of speech prosody is shaped by language experience: An fMRI study. *Brain and language*, *84*(3), 318-336.

75. Gandour, J., Tong, Y., Wong, D., Talavage, T., Dzemidzic, M., Xu, Y., ... & Lowe, M. (2004). Hemispheric roles in the perception of speech prosody. *Neuroimage*, *23*(1), 344-357.

76. Goldin-Meadow, S., Wein, D., & Chang, C. (1992). Assessing knowledge through gesture: Using children's hands to read their minds. *Cognition and Instruction*, *9*(3), 201-219.

77. Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science*, *2*(1), 67-74.

78. Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197-1208.

79. Grassmann, S., & Tomasello, M. (2010). Prosodic stress on a word directs 24-month-olds' attention to a contextually new referent. *Journal of Pragmatics*, *42*(11), 3098-3105.

80. Green, A., Straube, B., Weis, S., Jansen, A., Willmes, K., Konrad, K., & Kircher, T. (2009). Neural integration of iconic and unrelated coverbal gestures: A functional MRI study. *Human brain mapping*, *30*(10), 3309-3324.

81. Grosjean, F., & Gee, J. P. (1987). Prosodic structure and spoken word recognition. *Cognition*, *25*(1-2), 135-155.

82. Gullberg, M., & Holmqvist, K. (2006). What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition*, *14*(1), 53-82.

83. Gunter, T. C., & Weinbrenner, J. D. (2017). When to take a gesture seriously: On how we use and prioritize communicative cues. *Journal of Cognitive Neuroscience*, *29*(8), 1355-1367.

84. Gunter, T. C., Weinbrenner, J. E., & Holle, H. (2015). Inconsistent use of gesture space during abstract pointing impairs language comprehension. *Frontiers in psychology*, *6*, 80.

85. Habets, B., Kita, S., Shao, Z., Özyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of cognitive neuroscience*, *23*(8), 1845-1854.

86. Hall, D. A., Fussell, C., & Summerfield, A. Q. (2005). Reading fluent speech from talking faces: typical brain networks and individual differences. *Journal of cognitive neuroscience*, *17*(6), 939-953.

87. Hasson, U., Egidi, G., Marelli, M., & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, *180*, 135-157.

88. He, Y., Luell, S., Muralikrishnan, R., Straube, B., & Nagels, A. (2020). Gesture's body orientation modulates the N400 for visual sentences primed by gestures. *Human brain mapping*, *41*(17), 4901-4911.

89. He, Y., Steines, M., Sommer, J., Gebhardt, H., Nagels, A., Sammer, G., ... & Straube, B. (2018). Spatial–temporal dynamics of gesture–speech integration: a simultaneous EEG-fMRI study. *Brain Structure and Function*, *223*(7), 3073-3089.

90. Heim, S., & Alter, K. (2006). Prosodic pitch accents in language comprehension and production: ERP data and acoustic analyses. *Acta Neurobiologiae Experimentalis*, *66*(1), 55.

91. Hernández-Gutiérrez, D., Rahman, R. A., 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.

92. Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds.

93. Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of cognitive neuroscience*, *19*(7), 1175-1192.

94. Holle, H., Gunter, T. C., Rüschemeyer, S. A., Hennenlotter, A., & Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *Neuroimage*, *39*(4), 2010-2024.

95. Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A. D., Ward, J., & Gunter, T. C. (2012). Gesture facilitates the syntactic analysis of speech. *Frontiers in psychology*, *3*, 74.

96. Holle, H., Obleser, J., Rueschemeyer, S. A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *Neuroimage*, *49*(1), 875-884.

97. Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639-652.

98. Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological bulletin*, *137*(2), 297.

99. 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.

100. Hwang, H., & Steinhauer, K. (2011). Phrase length matters: The interplay between implicit prosody and syntax in Korean “garden path” sentences. *Journal of cognitive neuroscience*, *23*(11), 3555-3575.

101. Igualada, A., Esteve-Gibert, N., & Prieto, P. (2017). Beat gestures improve word recall in 3-to 5-year-old children. *Journal of Experimental Child Psychology*, *156*, 99-112.

102. Ischebeck, A. K., Friederici, A. D., & Alter, K. (2008). Processing prosodic boundaries in natural and hummed speech: An fMRI study. *Cerebral Cortex*, *18*(3), 541-552.

103. Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological science*, *16*(5), 367-371.

104. Kang, S., & Speer, S. R. (2003). Prosodic disambiguation of syntactic clause boundaries in Korean. In *Proceedings of the 22nd West Coast Conference on Formal Linguistics* (pp. 259-272).

105. Kang, S., Speer, S. R., & Nakayama, M. (2004, October). Effects of prosodic boundaries on ambiguous syntactic clause boundaries in Japanese. In *Proceedings of Interspeech* (Vol. 2004, p. 8t).

106. Karas, P. J., Magnotti, J. F., Metzger, B. A., Zhu, L. L., Smith, K. B., Yoshor, D., & Beauchamp, M. S. (2019). The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. *Elife*, *8*, e48116.

107. Karas, P. J., Magnotti, J. F., Wang, Z., Metzger, B. A., Yoshor, D., & Beauchamp, M. S. (2019). Audiovisual Speech Enhancement via Cross-Modal Suppression of Auditory Association Cortex by Visual Speech. *Neurosurgery*, *66*(Supplement_1), nyz310_695.

108. Kelly, S. D., & Church, R. B. (1998). A comparison between children's and adults' ability to detect conceptual information conveyed through representational gestures. *Child development*, *69*(1), 85-93.

109. Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and language*, *89*(1), 253-260.

110. Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and cognitive processes*, *24*(2), 313-334.

111. Kelly, S. D., Ward, S., Creigh, P., & Bartolotti, J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. *Brain and language*, *101*(3), 222-233.

112. Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition*, *93*(1), B39-B47.

113. Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes*, *24*(2), 145-167.

114. Kita, S. (Ed.). (2003). *Pointing: Where language, culture, and cognition meet*. Psychology Press.

115. Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language*, *40*(2), 153-194.

116. Knowland, V. C., Mercure, E., Karmiloff-Smith, A., Dick, F., & Thomas, M. S. (2014). Audio-visual speech perception: A developmental ERP investigation. *Developmental Science*, *17*(1), 110-124.

117. 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.

118. Krason, A., Fenton, R., Varley, R., & Vigliocco, G. (2021). The role of iconic gestures and mouth movements in face-to-face communication. *Psychonomic Bulletin & Review*, 1-13.

119. Krason*, Zhang* and Vigliocco (2020)

120. 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.

121. Kushch, O., & Prieto Vives, P. (2016). The effects of pitch accentuation and beat gestures on information recall in contrastive discourse. *Barnes J, Brugos A, Shattuck-Hufnagel S, Veilleux N, editors. Speech Prosody 2016; 2016 May 31-June 3; Boston, United States of America.[place unknown]: International Speech Communication Association; 2016. p. 922-5. DOI: 10.21437/SpeechProsody. 2016-189*.

122. 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.

123. Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, *62*, 621-647.

124. Lehiste, I. (1975). The phonetic structure of paragraphs. In *Structure and process in speech perception* (pp. 195-206). Springer, Berlin, Heidelberg.

125. Lehiste, I., Olive, J. P., & Streeter, L. A. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *The Journal of the Acoustical Society of America*, *60*(5), 1199-1202.

126. Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, *109*(5), 1431-1436.

127. Li, X. Q., & Ren, G. Q. (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.

128. Li, X., & Yang, Y. (2013). How long-term memory and accentuation interact during spoken language comprehension. *Neuropsychologia*, *51*(5), 967-978.

129. Li, X., Hagoort, P., & Yang, Y. (2008). Event-related potential evidence on the influence of accentuation in spoken discourse comprehension in Chinese. *Journal of Cognitive Neuroscience*, *20*(5), 906-915.

130. Li, X., Lu, Y., & Zhao, H. (2014). How and when predictability interacts with accentuation in temporally selective attention during speech comprehension. *Neuropsychologia*, *64*, 71-84.

131. Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and speech*, *6*(3), 172-187.

132. Llanes-Coromina, J., Vilà-Giménez, I., Kushch, O., Borràs-Comes, J., & Prieto, P. (2018). Beat gestures help preschoolers recall and comprehend discourse information. *Journal of Experimental Child Psychology*, *172*, 168-188.

133. MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British journal of audiology*, *21*(2), 131-141.

134. Macoun, A., & Sweller, N. (2016). Listening and watching: The effects of observing gesture on preschoolers' narrative comprehension. *Cognitive Development*, *40*, 68-81.

135. MacSweeney, M., Amaro, E., Calvert, G. A., Campbell, R., David, A. S., McGuire, P., ... & Brammer, M. J. (2000). Silent speechreading in the absence of scanner noise: an event-related fMRI study. *Neuroreport*, *11*(8), 1729-1733.

136. Marstaller, L., & Burianová, H. (2014). The multisensory perception of co-speech gestures–A review and meta-analysis of neuroimaging studies. *Journal of Neurolinguistics*, *30*, 69-77.

137. Massaro, D. W., & Jesse, A. (2007). Audiovisual speech perception and word. *The Oxford handbook of psycholinguistics*, 19-36.

138. Mattys, S. L., Bernstein, L. E., & Auer, E. T. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception & Psychophysics*, *64*(4), 667-679.

139. McAllister, J. (1991). The processing of lexically stressed syllables in read and spontaneous speech. *Language and Speech*, *34*(1), 1-26.

140. McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: behavioural and neural interactions. Neuropsychologia, 50(5), 762-776.

141. McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748.

142. McNeil, N. M., Alibali, M. W., & Evans, J. L. (2000). The role of gesture in children's comprehension of spoken language: Now they need it, now they don't. *Journal of Nonverbal Behavior*, *24*(2), 131-150.

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

144. McNeill, D., Cassell, J., & McCullough, K. E. (1994). Communicative effects of speech-mismatched gestures. *Research on language and social interaction*, *27*(3), 223-237.

145. Ménard, L., Leclerc, A., & Tiede, M. (2014). Articulatory and acoustic correlates of contrastive focus in congenitally blind adults and sighted adults. *Journal of Speech, Language, and Hearing Research*, *57*(3), 793-804.

146. Meyer, M., Alter, K., & Friederici, A. (2003). Functional MR imaging exposes differential brain responses to syntax and prosody during auditory sentence comprehension. *Journal of Neurolinguistics*, *16*(4-5), 277-300.

147. Meyer, M., Steinhauer, K., Alter, K., Friederici, A. D., & von Cramon, D. Y. (2004). Brain activity varies with modulation of dynamic pitch variance in sentence melody. *Brain and language*, *89*(2), 277-289.

148. Mietz, A., Toepel, U., Ischebeck, A., & Alter, K. (2008). Inadequate and infrequent are not alike: ERPs to deviant prosodic patterns in spoken sentence comprehension. *Brain and Language*, *104*(2), 159-169.

149. Morett, L. M., & Fraundorf, S. H. (2019). Listeners consider alternative speaker productions in discourse comprehension and memory: Evidence from beat gesture and pitch accenting. *Memory & Cognition*, *47*(8), 1515-1530.

150. Morett, L. M., Landi, N., Irwin, J., & McPartland, J. C. (2020). N400 amplitude, latency, and variability reflect temporal integration of beat gesture and pitch accent during language processing. *Brain Research*, *1747*, 147059.

151. Morett, L. M., Roche, J. M., Fraundorf, S. H., & McPartland, J. C. (2020). Contrast is in the eye of the beholder: Infelicitous beat gesture increases cognitive load during online spoken discourse comprehension. *Cognitive Science*, *44*(10), e12912.

152. Morett, L., Fraundorf, S. H., & McPartland, J. C. (2019). Eye See What You're Saying: Beat Gesture Facilitates Online Resolution of Contrastive Referring Expressions in Spoken Discourse. In *CogSci* (pp. 843-848).

153. Morissette, P., Ricard, M., & Décarie, T. G. (1995). Joint visual attention and pointing in infancy: A longitudinal study of comprehension. *British Journal of Developmental Psychology*, *13*(2), 163-175.

154. Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological science*, *15*(2), 133-137.

155. Murgiano, M., Motamedi, Y., & Vigliocco, G. (2021). Situating language in the real-world: the role of multimodal iconicity and indexicality. *Journal of Cognition*, *4*(1).

156. Nakamura, C., Arai, M., & Mazuka, R. (2012). Immediate use of prosody and context in predicting a syntactic structure. *Cognition*, *125*(2), 317-323.

157. Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience*, *31*(5), 1704-1714.

158. Nickels, S., & Steinhauer, K. (2018). Prosody–syntax integration in a second language: Contrasting event-related potentials from German and Chinese learners of English using linear mixed effect models. *Second Language Research*, *34*(1), 9-37.

159. Obermeier, C., Dolk, T., & Gunter, T. C. (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex*, *48*(7), 857-870.

160. Obermeier, C., Holle, H., & Gunter, T. C. (2011). What iconic gesture fragments reveal about gesture–speech integration: When synchrony is lost, memory can help. *Journal of Cognitive Neuroscience*, *23*(7), 1648-1663.

161. Obermeier, C., Kelly, S. D., & Gunter, T. C. (2015). A speaker's gesture style can affect language comprehension: ERP evidence from gesture-speech integration. *Social cognitive and affective neuroscience*, *10*(9), 1236-1243.

162. Özçalışkan, Ş., & Goldin-Meadow, S. (2005). Gesture is at the cutting edge of early language development. *Cognition*, *96*(3), B101-B113.

163. Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130296.

164. Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of cognitive neuroscience*, *19*(4), 605-616.

165. Pannekamp, A., Toepel, U., Alter, K., Hahne, A., & Friederici, A. D. (2005). Prosody-driven sentence processing: an event-related brain potential study. *Journal of cognitive neuroscience*, *17*(3), 407-421.

166. Pauker, E., Itzhak, I., Baum, S. R., & Steinhauer, K. (2011). Effects of cooperating and conflicting prosody in spoken English garden path sentences: ERP evidence for the boundary deletion hypothesis. *Journal of Cognitive Neuroscience*, *23*(10), 2731-2751.

167. Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181.

168. Peeters, D., Snijders, T. M., Hagoort, P., & Özyürek, A. (2017). Linking language to the visual world: Neural correlates of comprehending verbal reference to objects through pointing and visual cues. *Neuropsychologia*, *95*, 21-29.

169. Perlman, M., & Cain, A. A. (2014). Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language. *Gesture*, *14*(3), 320-350.

170. Perrone, M., Dohen, M., Loevenbruck, H., Sato, M., Pichat, C., Yvert, G., & Baciu, M. (2010). An fMRI study of the perception of contrastive prosodic focus in French. In *Speech Prosody 2010-Fifth International Conference*.

171. Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception.

172. Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *the Journal of the Acoustical Society of America*, *90*(6), 2956-2970.

173. recognition. In *ICPRAM (2)* (pp. 322-329).

174. Rohrer, P. L., Delais-Roussarie, E., & Prieto, P. (2020). Beat gestures for comprehension and recall: Differential effects of language learners and native listeners. *Frontiers in Psychology*, *11*.

175. Roncaglia-Denissen, M. P., Schmidt-Kassow, M., & Kotz, S. A. (2013). Speech rhythm facilitates syntactic ambiguity resolution: ERP evidence. *PloS one*, *8*(2), e56000.

176. Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2005). Prosodic influences on the production and comprehension of syntactic ambiguity in a game-based conversation task. *Approaches to studying world-situated language use*, 209-225.

177. Schubotz, L., Holler, J., Drijvers, L., & Özyürek, A. (2021). Aging and working memory modulate the ability to benefit from visible speech and iconic gestures during speech-in-noise comprehension. *Psychological research*, *85*(5), 1997-2011.

178. Sekine, K., Schoechl, C., Mulder, K., Holler, J., Kelly, S., Furman, R., & Özyürek, A. (2020). Evidence for children's online integration of simultaneous information from speech and iconic gestures: an ERP study. *Language, Cognition and Neuroscience*, *35*(10), 1283-1294.

179. Sekine, K., Sowden, H., & Kita, S. (2015). The development of the ability to semantically integrate information in speech and iconic gesture in comprehension. *Cognitive Science*, *39*(8), 1855-1880.

180. Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2007). Speech-associated gestures, Broca's area, and the human mirror system. *Brain and language*, *101*(3), 260-277.

181. Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage*, *25*(1), 76-89.

182. Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, *17*(10), 2387-2399.

183. Snedeker, J., & Casserly, E. (2010). Is it all relative? Effects of prosodic boundaries on the comprehension and production of attachment ambiguities. *Language and Cognitive Processes*, *25*(7-9), 1234-1264.

184. Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and language*, *48*(1), 103-130.

185. So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive science*, *33*(1), 115-125.

186. So, W. C., Sim Chen-Hui, C., & Low Wei-Shan, J. (2012). Mnemonic effect of iconic gesture and beat gesture in adults and children: Is meaning in gesture important for memory recall?. *Language and Cognitive Processes*, *27*(5), 665-681.

187. Speer, S. R., Kjelgaard, M. M., & Dobroth, K. M. (1996). The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities. *Journal of psycholinguistic research*, *25*(2), 249-271.

188. Speer, S. R., Warren, P., & Schafer, A. (2003). Intonation and sentence processing. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 95-105).

189. Spinelli, M., Fasolo, M., & Mesman, J. (2017). Does prosody make the difference? A meta-analysis on relations between prosodic aspects of infant-directed speech and infant outcomes. *Developmental Review*, *44*, 1-18.

190. Steinhauer, K., Abada, S. H., Pauker, E., Itzhak, I., & Baum, S. R. (2010). Prosody–syntax interactions in aging: Event-related potentials reveal dissociations between on-line and off-line measures. *Neuroscience Letters*, *472*(2), 133-138.

191. Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature neuroscience*, *2*(2), 191-196.

192. Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of cognitive neuroscience*, *19*(12), 1964-1973.

193. Stekelenburg, J., & Vroomen, J. (2012). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. *Frontiers in Integrative Neuroscience*, *6*, 26.

194. Stevens, J., & Zhang, Y. (2013). Relative distance and gaze in the use of entity-referring spatial demonstratives: An event-related potential study. *Journal of Neurolinguistics*, *26*(1), 31-45.

195. Stevens, J., & Zhang, Y. (2014). Brain mechanisms for processing co-speech gesture: a cross-language study of spatial demonstratives. *Journal of Neurolinguistics*, *30*, 27-47.

196. Straube, B., Green, A., Jansen, A., Chatterjee, A., & Kircher, T. (2010). Social cues, mentalizing and the neural processing of speech accompanied by gestures. *Neuropsychologia*, *48*(2), 382-393.

197. Straube, B., Green, A., Weis, S., & Kircher, T. (2012). A supramodal neural network for speech and gesture semantics: an fMRI study. *PloS one*, *7*(11), e51207.

198. Straube, B., Wroblewski, A., Jansen, A., & He, Y. (2018). The connectivity signature of co-speech gesture integration: The superior temporal sulcus modulates connectivity between areas related to visual gesture and auditory speech processing. *NeuroImage*, *181*, 539-549.

199. Strelnikov, K. N., Vorobyev, V. A., Chernigovskaya, T. V., & Medvedev, S. V. (2006). Prosodic clues to syntactic processing—a PET and ERP study. *Neuroimage*, *29*(4), 1127-1134.

200. Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661-699.

201. 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.

202. Sweller, N., Shinooka-Phelan, A., & Austin, E. (2020). The effects of observing and producing gestures on Japanese word learning. *Acta Psychologica*, *207*, 103079.

203. Swerts, M., & Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and speech*, *37*(1), 21-43.

204. Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, *8*(2), 219-235.

205. Tenenbaum, E. J., Sobel, D. M., Sheinkopf, S. J., Malle, B. F., & Morgan, J. L. (2015). Attention to the mouth and gaze following in infancy predict language development. *Journal of Child Language*, *42*(6), 1173-1190.

206. Ter Bekke, M., Drijvers, L., & Holler, J. (2020). The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech.

207. Terken, J., & Hirschberg, J. (1994). Deaccentuation of words representing 'given'information: Effects of persistence of grammatical function and surface position. *Language and Speech*, *37*(2), 125-145.

208. Terken, J., & Nooteboom, S. G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and cognitive processes*, *2*(3-4), 145-163.

209. Thompson, L. A., & Massaro, D. W. (1994). Children′ s integration of speech and pointing gestures in comprehension. *Journal of Experimental Child Psychology*, *57*(3), 327-354.

210. Tong, Y., Gandour, J., Talavage, T., Wong, D., Dzemidzic, M., Xu, Y., ... & Lowe, M. (2005). Neural circuitry underlying sentence-level linguistic prosody. *NeuroImage*, *28*(2), 417-428.

211. Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification*, *11*(4), 233-241.

212. Vainio, L. (2019). Connection between movements of mouth and hand: Perspectives on development and evolution of speech. *Neuroscience & Biobehavioral Reviews*, *100*, 211-223.

213. Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, *28*(2), 187-204.

214. van der Burght, C. L., Goucha, T., Friederici, A. D., Kreitewolf, J., & Hartwigsen, G. (2019). Intonation guides sentence processing in the left inferior frontal gyrus. *Cortex*, *117*, 122-134.

215. Van Engen, K. J., Phelps, J. E., Smiljanic, R., & Chandrasekaran, B. (2014). Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1908-1918.

216. Van Leeuwen, T. M., Lamers, M. J., Petersson, K. M., Gussenhoven, C., Rietveld, T., Poser, B., & Hagoort, P. (2014). Phonological markers of information structure: An fMRI study. *Neuropsychologia*, *58*, 64-74.

217. Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181-1186.

218. Venditti, J. J., & Hirschberg, J. (2003). Intonation and discourse processing. In *Proceedings of the international congress of phonetic sciences* (pp. 315-318). Saarbrücken,, Germany: University of Saarbrücken.

219. Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: implications for language learning, processing and evolution.

220. Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of cognitive neuroscience*, *22*(7), 1583-1596.

221. Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, *25*(7-9), 905-945.

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

223. Wang, L., Bastiaansen, M., Yang, Y., & Hagoort, P. (2011). The influence of information structure on the depth of semantic processing: How focus and pitch accent determine the size of the N400 effect. *Neuropsychologia*, *49*(5), 813-820.

224. Warren, P., Grabe, E., & Nolan, F. (1995). Prosody, phonology and parsing in closure ambiguities. *Language and cognitive processes*, *10*(5), 457-486.

225. Watson, D. G., Arnold, J. E., & Tanenhaus, M. K. (2008). Tic Tac TOE: Effects of predictability and importance on acoustic prominence in language production. *Cognition*, *106*(3), 1548-1557.

226. Watson, D., & Gibson, E. (2005). Intonational phrasing and constituency in language production and comprehension. *Studia linguistica*, *59*(2-3), 279-300.

227. Weintraub, S., Mesulam, M. M., & Kramer, L. (1981). Disturbances in prosody: A right-hemisphere contribution to language. *Archives of Neurology*, *38*(12), 742-744.

228. Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral cortex*, *13*(10), 1034-1043.

229. Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology*, *42*(6), 654-667.

230. Yang, J., Andric, M., & Mathew, M. M. (2015). The neural basis of hand gesture comprehension: a meta-analysis of functional magnetic resonance imaging studies. *Neuroscience & Biobehavioral Reviews*, *57*, 88-104.

231. Yule, G. (1980). Speakers' topics and major paratones. *Lingua*, *52*(1-2), 33-47.

232. Zhao, W., Riggs, K., Schindler, I., & Holle, H. (2018). Transcranial magnetic stimulation over left inferior frontal and posterior temporal cortex disrupts gesture-speech integration. *Journal of Neuroscience*, *38*(8), 1891-1900.

Chapter 3

3 Informativeness of mouth movements for 1,743 English words

3.1 Introduction

Mouth movements are always present in daily face-to-face communication and facilitate listeners to identify auditory signals (see review in Chapter 2). However, previous studies investigating the impact of mouth movements on language processing predominately manipulated the visibility of mouth movements, either contrasting audiovisual with audio stimuli (e.g. Arnold & Hill, 2010) or contrasting videos with mouth present versus blurred (to different degrees, e.g. Drijvers & Özyürek, 2017; Scott et al., 2002; McGettigan et al., 2012). These studies clearly showed that visual information accompanying speech, including mouth movements, is incorporated in language processing. However, it is possible that different words, each with distinct mouth movements, carry different amount of information in the visual modality. As a fine-grained quantification of any word's mouth informativeness (i.e. how easy it is to identify a word based on its mouth movement patterns) remains unknown, here in this chapter, we proposed a novel measure of mouth informativeness, assessed by lipreading tasks (Krason*, Zhang*, Vigliocco, submitted).

Although it is widely acknowledged that mouth movements contribute to identifying speech signals (see review in Pelle & Sommers, 2015), some mouth movements are more visually salient while others are less informative. For example, consonants produced at labial (e.g., /b/, /p/, /m/) or labial-dental (e.g., /f/, /v/)

positions (e.g., Binnie, Montgomery, & Jackson, 1974; Benguerel & Pichora-Fuller, 1982), as well as vowels with a rounding feature (e.g., /u/, /o/, /ɔ/; Robert-Ribes et al., 1998; Traunmüller & Öhrström, 2007) have long been recognized as visually more identifiable. However, phoneme categories are used to describe acoustically distinct speech segments; whereas, different phonemes may look very similar visually (e.g. /b/ and /p/, which are not visually distinguishable based on mouth movements only). Therefore, phoneme categories may not be suitable to describe different mouth movements in the visual domain. Some researchers further classified phonemes into visemes (Fisher, 1968). Each viseme consists of a group of sounds with the same mouth movements (e.g. /b/ and /p/ are visually the same and therefore belong to the same viseme class; whereas /b/ and /d/ have different mouth movements, and therefore are separated into two viseme classes). Jesse and Massaro (2010) investigated visual features that make words visually more salient. They presented participants with a set of one-syllable words with CVC structures in a gating task in which participants needed to identify the word based on the partial information available. The author associated improved word identification performance with viseme features including lower lip tuck (i.e. tucking the lower lip under the upper teeth, e.g., /v/); protrusion (i.e. sticking the lips out, e.g., /ʃ/), labial closure (i.e. drawing the upper and lower lips closer, e.g. /p/), mouth narrowing (i.e. horizontally bringing the lips closer, e.g. /w/), and finally rounding (i.e. creating a rounded shape with the lips, e.g. /r/).

Apart from identifying the informativeness of mouth movement features, Jesse and Massaro (2010) further reported that the impact of informative visual features was largest before the end of the first phoneme in a word, indicating that mouth information might be beneficial only in the early perception of words. Indeed, visual

information from the mouth occurs 100-300ms prior to the auditory signal (Chandrasekaran et al., 2009; van Wassenhove et al., 2005), and thus, might constrain phoneme identification and predict specific phonemes. For example, Karas et al. (2019) found that words with a "visual head start" (in which the mouth movements begin significantly earlier than the auditory information, e.g. 'drive' compared with 'known') showed a larger audiovisual benefit over auditory-only speech. Whereas mouth movements later in a word may be less influential.

However, these studies investigating the informativeness of mouth movements predominately looked at very short words (e.g. single-syllable words, Jesse & Massaro, 2010) or a single phoneme in a word (e.g. initial phoneme, Karas et al., 2019), which may not be sufficient to capture the dynamic nature of mouth movements and quantify mouth informativeness in longer words (with two or more syllables). It is difficult, if possible at all, to calculate the informativeness of a certain word based on our knowledge of phoneme/visemes informativeness. For example, is the word “base (/beɪs/)” more or less informative based on mouth movement than the word “subscription (/səbˈskrɪpʃn/)”? The former contains one visible labial phoneme/viseme (“/b/”), but the latter contains two labial phoneme/viseme (“/b/” and “/p/”), one rounding (“/r/”) and one protrusion (“/ʃ/”). However, the latter is also three times as long as the former, thus containing a long string of other mouth movements, which may either increase its mouth informativeness (as these movements nonetheless convey visual information that comprehenders may use) or reduce it (as these mouth movements may not be easily identifiable and therefore may simply be a distraction). Another important factor contributing to the difficulty of such comparison is the position effect established above: a phoneme/viseme’s contribution to the informativeness of a word varies based on its location within a

word. Therefore, although both words contain the sound /b/, it may not be equally informative as it appears in the initial position for “base” but non-initial position in “subscription”. Finally, the phonetic context can modify the identifiability of mouth movements. For example, Benguerel and Pichora-Fuller (1982) found that while lipreading performance of VCV syllables with visually more salient mouth movements (including articulation of /p/, /f/, /u/) was high regardless of the subsequent phonemes, the identification performance of mouth movements with lower visual saliency (e.g. /t/ or /k/) is modified by the phonetic context. The vastly different phonetic contexts in word “base” and “subscription” make their comparison even more difficult based on their phonemes.

To identify the informativeness of words based on mouth movements, we adopted a new approach and constructed a corpus of mouth informativeness for 1,743 English words, differing in their visual saliency, length, frequency, concreteness, and age of acquisition (AoA). We presented participants with muted videos of single words being pronounced and invited them to guess the identity of the word. We then calculated the phonological distance between lipreading guesses and target words (referred to as “distance” below). Shorter distance represents more accurate guesses, indicating higher mouth informativeness of a word. The norms can be accessed via Open Science Framework (OSF, https://osf.io/mna8j/), and we aim to continue the collection of norms for more words.

This corpus would be helpful for the researchers investigating the informativeness of each mouth features. Researchers may potentially investigate how the dynamic mouth movements affect the identifiability of words, how phonological context modulates such effect, and how other lexical level variables modulate the impact of mouth movements (e.g. frequency). This corpus may also

help computer scientists to validate computation models attempting to capture mouth movements informativeness. Finally, our corpus may offer another tool for language and psychology researchers investigating the impact of other multimodal cues (e.g. gestures), allowing them to move towards a more naturalistic design. As reviewed in Chapter 2, the constant presence and co-occurrence are key properties of multimodal communication in the real world. However, researchers were forced to hide some cues, such as mouth movements, to achieve experimental control (e.g. Holle & Gunter, 2007; Drijvers & Özyürek, 2017). Our corpus allows researchers to track and account for the co-occurring mouth information, therefore increasing the naturalness while maintaining scientific rigour.

In this chapter, we first introduce the collection of and quantification of mouth informativeness, involving three separate studies using native British speakers and native American speakers. In order to validate our corpus, we then conducted a confirmatory analysis, assessing whether visually salient mouth features also predict higher mouth informativeness, indexed by shorter distance between participants' guess and the actual word. Finally, we carried out two exploratory analyses, assessing whether word informativeness is also associated with salient visual features in non-initial positions or their total number of occurrences. Each analysis was performed on: (i) the combined corpus of British and American words (1743 words in total without duplicates); (ii) the British corpus only (1097 words in total); (iii) the American corpus only (745 words in total) to identify cross-accent differences. We discuss the implications of our findings in the last section.

3.2 Methods

3.2.1 Participants

Native English speakers were recruited from Prolific (http://www.prolific.co/) to take part in three separate studies. Studies 1 recruited 150 (111 females, 37 males, and 2 non-binary; *Mean age* = 28, *SD* = 6.45) British English-speaking participants. Study 2 recruited 59 (40 females, 18 males and 1 non-binary, *Mean age* = 26, *SD* =7.13) British participants. Study 3 recruited 145 American participants, out of which eight were excluded due to experiencing technical issues or incorrectly answering the catch trials (see below), leaving data from 137 participants (71 females, 64 males, and 2 non-binary; *Mean age*=29, *SD*=6.24). All participants provided consent to take part in the experiment and were paid £6/hour for their time. The ethical approval was obtained from University College London (UCL; Research Ethics Committee 0143/003).

3.2.2 Materials

A total of 1842 words were video-recorded (study 1: 315 words, study: 782 words, study 3: 745 words). 1097 words were included the British slice (Study 1 and 2) and 745 words in the American slice of the corpus (Study 3), with 99 words included in both slices. A native British English actress produced the words in Study 1 and 2 while a native American English actress produced the words in Study 3. The videos were recorded with a professional camera (Panasonic HC-V180) either at UCL in a sound-proof recording booth (study 2 and study 3) or at an actress' home due to Covid-19 restrictions that were present in the UK at the time of stimuli preparation (study 1). Each video depicted the face of the actress uttering an English word with a neutral accent and facial expression. The videos were then muted for the purpose of the experiment. The mean length of each video was ~1000ms. The

uttered words varied in the number of phonemes (range: 1-12), syllables (range: 1-5), log-frequency (Balota et al., 2007, range: 3.315-15.897), concreteness (Brysbaert et al., 2014; range: 1.19-5 out of 5; 84 words missing concreteness norms), AoA (Kuperman et al. 2012, range: 2.37-14.75; 128 words missing AoA norms).

3.2.3 Procedure

Participants took part in an online experiment created on Gorilla (https://gorilla.sc/). Participants were asked to watch the muted videos and guess the word produced by the speaker by typing it in the answer box provided. Individuals were randomly assigned to respond to ~50 trials in study 1, ~60 trials in study 2, and ~100 trials in study 3, and each word was guessed by at least 10 different participants. Participants initiated the videos by clicking on them and each video was automatically presented twice in a row (to minimize the difficulty of the task, and to make sure participants did not miss the beginning of each trial). A typing box appeared simultaneously with the second presentation of a video. There was a 250ms interval between the trials. Before the experiment, participants were exposed to seven practice trials, where participants first guessed a word and then saw the correct word on the screen. Participants were encouraged to make their best guess if unsure of the correct answer. Additionally, we included several control trials, consisting of a lexical decision task where we showed participants pictures of everyday objects followed by a question (e.g., '*Was this a candle?'*). The control trials were randomly distributed within the experiment to identify participants who did not pay attention to the task. The entire experiment lasted between 20-40 minutes. Figure 1 depicts an example of trial types used in the studies.

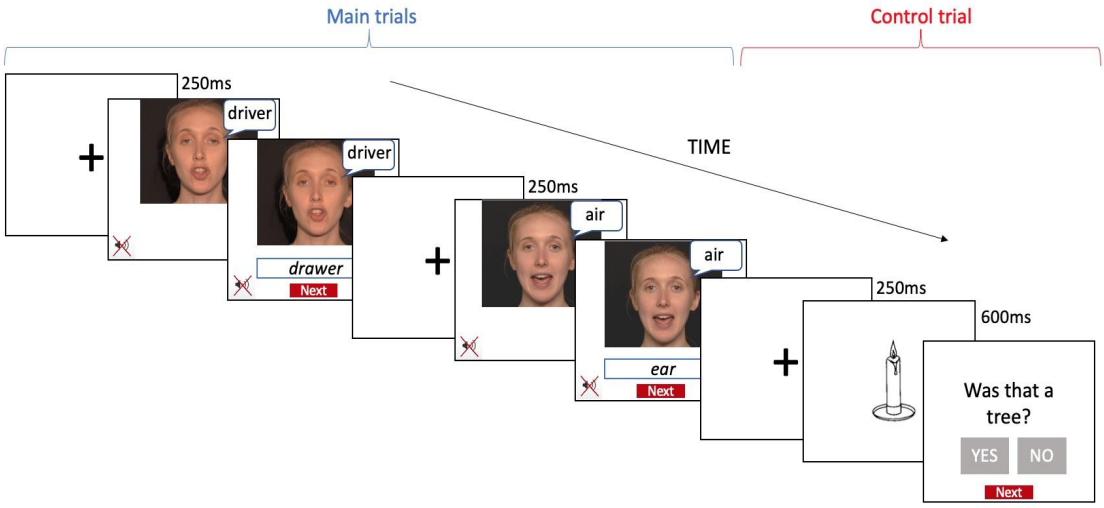


Figure 1.

Example of two experimental trials and one control trial.

3.2.4 Quantifying mouth informativeness

To assess informativeness of the mouth and facial movements, we decided to calculate how similar (or distinct) are participants' lipreading guesses to the target words. We used a lipreading task instead of a matching task (e.g., watching silent video-clips and guessing what was uttered by the speaker by choosing the correct answer among a number of foils) to (1) establish as accurate measure of mouth informativeness as possible, (2) avoid any effect of foil selection, and (3) ensure response variability between participants. After collecting the data, we first corrected accidental spaces and obvious typing errors (e.g., "barbeque" was corrected into "barbecue"). We then phonetically transcribed the target words and participants' responses, according to either British (studies 1 and 2) or American (study 3) International Phonetic Alphabet (IPA). Next, we calculated the string distance between the two words by taking into account words' length and their phonological features (equally weighted), using the *PanPhon* package (Mortensen et al., 2016). This package took the IPA transcriptions of the two words, conveyed them to two

separate sequences of phonological feature vectors, and computed the Levinshtein distance between these two vectors. Any missing responses were removed from the analysis (~0.6% in study 1, ~0.5% in study 2; <0.5% in study 3). Finally, for each target word, we averaged the distance values obtained from different participants, which hereafter we call 'distances.' Thus, distances have any values from 0 (correct guesses, highly informative words) onwards, and the larger the score, the more difficult it is to guess the words based on lipreading only. We additionally calculated the accuracy (i.e. whether participants guessed the word correctly), number of phonemes correctly identified, and the percentage of phonemes correctly identified. Calculations were carried out in PyCharm (2018.2.4) and the summary of the results is presented in Table 1.

Table 1.

Mean distances between participants' guesses and target words collected from three studies. Smaller scores represent more informative mouth movements.

Study	Word Number	Accent	Mean Score	SD	Range	Mean Accuracy	Correct Phoneme Number	Correct Phoneme %
Study 1	315	British	0.85	0.27	0.00-1.51	0.221	4.213	65.9%
Study 2	782	British	0.86	0.30	0.00-1.55	0.224	3.037	59.5%
Study 3	745	American	1.06	0.27	0.06-1.69	0.162	2.715	50.6%

3.3 Data analysis

The analyses were carried out in RStudio (V. 4.0.4) and the R code is available on OSF (https://osf.io/mna8j/). We performed confirmatory and exploratory analyses (see below) separately for (1) a combined corpus of British and American words; (2) a British corpus only; (3) an American corpus only. For all the regression

models reported here, the categorical variables were dummy coded and the continuous variables were centred and scaled using the scale() function embedded in R. Summary of the models with their main predictors and outcome is depicted in Table 2. Additionally, we included number of phonemes, AoA (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), log-frequency (Balota et al., 2007), phonological neighborhood density (Luce & Pisoni, 1998) as control variables in all the models. We initially included orthographical neighborhood density as a control variable but removed it later due to correlation with phonological neighborhood density.

3.3.1 Confirmatory analyses

Confirmatory analyses were run to validate our datasets. Based on previous literature, we predicted 1) Word initial phonemes with a front place of articulation, including bilabials (/b/, /p/, /m/) and labial-dentals (/f/, /v/; e.g., Binnie et al., 1974; Benguerel & Pichora-Fuller, 1982; Jesse & Massaro, 2010), as well as phonemes with a rounding feature (/r/, /w/, /u/, /o/, /ɔ/; Robert-Ribes et al., 1998; Traunmüller & Öhrström, 2007; Jesse & Massaro, 2010) should convey more visual information than other phonemes. Thus, the words including these features should be visually more informative, as indicated by lower distance (i.e., small, averaged distance). This effect should be the strongest in a word initial position (Jesse & Massaro, 2010). 2) Word initial visemes with lower lip tuck (viseme {f}, including phoneme /f/, /v/), protrusion (viseme {ch}, including phoneme /ʃ/, /tʃ/, /dʒ/, viseme {w}, including phoneme /w/), labial closure (viseme {p}, including phoneme /b/, /p/, /m/), mouth narrowing (viseme {w}, including phoneme /w/), and lip rounding (viseme {j}, including phoneme /j/, viseme {r}, including phoneme /r/, and viseme {w}, including

phoneme /w/) are visually more salient (Jesse & Massaro, 2010) and thus they should be present in words with distances closer to 0 (i.e., small averaged).

To test our predictions, we carried out two separate multiple linear regression analyses with distances as our dependent variable and the following predictors: rounding and frontness of the initial phoneme (Model 1.1); lower lip tuck, protrusion, lip closure and rounding in the initial position (Model 1.2). We removed the mouth narrowing feature from all the models due to multiple collinearities. Note that mouth narrowing contains only a single phoneme /w/, which is also present in the protrusion and rounding categories.

3.3.2 Exploratory analyses

Exploratory analyses were carried out to investigate: 1) whether the informative features found in confirmatory analyses further increase informativeness of the mouth when they are present in a non-initial position (i.e., after the first phoneme); and 2) whether the number of informative features within a word (so called “informativeness load”) predicts mouth informativeness.

The first exploratory analysis investigates the effect of salient phonemes in non-initial positions. We carried out another set of multiple linear regression analyses with mouth informativeness as our outcome variable and the following predictors: frontness (/b/, /p/, /m/, /f/, /v/) and rounding (/r/, /w/, /u/, /o/, /ɔ/) of initial and non-initial phonemes (Model 2.1) and informative viseme features (lower lip tuck {f}, protrusion {ch}, {w}, labial closure {p}, and lip rounding {j}, {r}, {w}) in initial and non-initial positions (Model 2.2) We then carried out model comparisons (comparing Model 1.1 vs. Model 2.1, and Model 1.2 vs. Model 2.2) using likelihood ratio test to examine whether features in non-initial positions significantly improved the model fit.

In the second exploratory analysis, we asked whether the number of informative features in a given word quantitatively contributes to the overall informativeness of mouth movements (with more features leading to more informative movements). We coded a new variable - informativeness load - by counting the occurrence of informative features in a word, and dividing it by the word length. We built two multiple regression models (Model 2.3 & 2.4) based on phoneme and viseme features separately with distance as dependent variable and index of informativeness as the predictor.

Table 2.

Summary of the models and their main predictors tested in this study. Each of these models was run for a combined corpus of British-American words, British corpus only, and finally American corpus only.

Confirmatory analyses			
	Predictors		Outcome
	Feature in initial position	Feature in non-initial position	
Model 1.1	rounding, frontness	n/a	Distance
Model 1.2[a]	lower lip tuck, protrusion, lip closure, rounding	n/a	
Exploratory analyses			
	Predictors		Outcome
	Feature in initial position	Feature in non-initial position	
Model 2.1	rounding, frontness	rounding, frontness	Distance
Model 2.2[a]	lower lip tuck, protrusion, lip closure, rounding	lower lip tuck, protrusion, lip closure, rounding	
Model 2.3	Informativeness load based on rounding and frontness features		
Model 2.4[a]	Informativeness load based on lower lip tuck, protrusion, lip closure, and rounding features		

Note: We also included the number of phonemes within a word, word frequency and age of acquisition, as well as phonological neighborhood density as control variables in every model.

[a]Mouth narrowing was removed due to multicollinearity.

3.4 Results

We only present significant results of target variables below for simplicity. The statistics of the full models is presented in Table 3-6.

3.4.1 Confirmatory analyses

As predicted, words starting with phonemes with rounding and frontness features had overall a lower distance, indicating words with more informative mouth movements. The pattern was the same across the corpus ($B_{rounding} = -0.72$, $t_{rounding} = -6.78$, $p_{rounding} < .001$; $B_{frontness} = -0.33$, $t_{frontness} = -6.12$, $p_{frontness} < .001$), for the British slice ($B_{rounding} = -0.67$, $t_{rounding} = -5.16$, $p_{rounding} < .001$; $B_{frontness} = -0.28$, $t_{frontness} = -4.01$, $p_{frontness} < .001$) and the American slice ($B_{rounding} = -0.88$, $t_{rounding} = -5.80$, $p_{rounding} < .001$; $B_{frontness} = -0.51$, $t_{frontness} = -7.01$, $p_{frontness} < .001$).

We also found that the viseme features of lower lip tuck, labial closure and lip rounding led to lower distances, showing that words starting with these features have more informative mouth movements. The effect of protrusion in the initial position was not significant. The pattern was similar across the corpus ($B_{lowerLipTuck} = -0.51$, $t_{lowerLipTuck} = -5.24$, $p_{lowerLipTuck} < .001$; $B_{labialClosure} = -0.25$, $t_{labialClosure} = -4.17$, $p_{labialClosure} < .001$; $B_{lipRounding} = -0.42$, $t_{lipRounding} = -2.88$, $p_{lipRounding} = 0.004$); for the British slice ($B_{lowerLipTuck} = -0.42$, $t_{lowerLipTuck} = -3.42$, $p_{lowerLipTuck} < .001$; $B_{labialClosure} = -0.20$, $t_{labialClosure} = -2.53$, $p_{labialClosure} = 0.012$; $B_{lipRounding} = -0.38$, $t_{lipRounding} = -2.06$, $p_{lipRounding} = 0.039$) and the American slice ($B_{lowerLipTuck} = -0.74$, $t_{lowerLipTuck} = -5.42$, $p_{lowerLipTuck} < .001$; $B_{labialClosure} = -0.44$, $t_{labialClosure} = -5.56$, $p_{labialClosure} < .001$; $B_{lipRounding}$

= -0.57, $t_{lipRounding}$ = -2.87, $p_{lipRounding}$ = 0.004). Table 3 shows the full results for the confirmatory analysis. Figure 3 shows the mean informativeness score for informative features in the initial position for the combined British-American corpus as an example, as the pattern is very similar for the British and American slice.

Table 3.

Full results from confirmatory analysis.

	Across Corpus				**British Slice**				**American Slice**			
Model 1.1: Phoneme feature in initial position												
	β	SE	t	p	β	SE	t	p	β	SE	t	p
Initial												
Rounding	-0.72	0.11	-6.78	<.001	-0.67	0.13	-5.16	<.001	-0.88	0.15	-5.80	<.001
Frontness	-0.33	0.05	-6.12	<.001	-0.28	0.07	-4.01	<.001	-0.51	0.07	-7.01	<.001
Control												
Freq	-0.20	0.03	-7.29	<.001	-0.03	0.04	-0.84	0.402	-0.05	0.05	-0.94	0.348
PhonNBH	0.01	0.03	0.28	0.781	0.03	0.04	0.83	0.410	-0.04	0.05	-0.80	0.423
AoA	0.05	0.03	2.02	0.043	0.17	0.04	4.72	<.001	0.20	0.05	4.37	<.001
PhonNum	0.03	0.04	0.91	0.363	0.03	0.05	0.59	0.553	-0.01	0.05	-0.16	0.885
Adjusted R^2	0.10				0.09				0.14			
Model 1.2: Viseme feature in initial position												
	β	SE	t	p	β	SE	t	p	β	SE	t	p
Initial												
LowerLipTuck	-0.51	0.10	-5.24	<.001	-0.42	0.12	-3.42	0.001	-0.74	0.14	-5.42	<.001
Protrusion	-0.10	0.11	-0.87	0.387	-0.05	0.15	-0.37	0.714	-0.18	0.15	-1.25	0.212
LabialClosure	-0.25	0.06	-4.17	<.001	-0.20	0.08	-2.53	0.012	-0.44	0.08	-5.56	<.001
LipRounding	-0.42	0.15	-2.88	0.004	-0.38	0.19	-2.06	0.039	-0.57	0.20	-2.87	0.004
Control												
Freq	-0.21	0.03	-7.50	<.001	-0.05	0.04	-1.16	0.249	-0.05	0.05	-0.99	0.324
PhonNBH	0.02	0.03	0.59	0.554	0.05	0.04	1.22	0.223	-0.04	0.05	-0.91	0.361
AoA	0.06	0.03	2.08	0.038	0.16	0.04	4.58	<.001	0.21	0.05	4.44	0.000
PhonNum	0.04	0.04	1.15	0.252	0.05	0.05	1.02	0.309	-0.02	0.05	-0.37	0.715
Adjusted R^2	0.09				0.07				0.13			

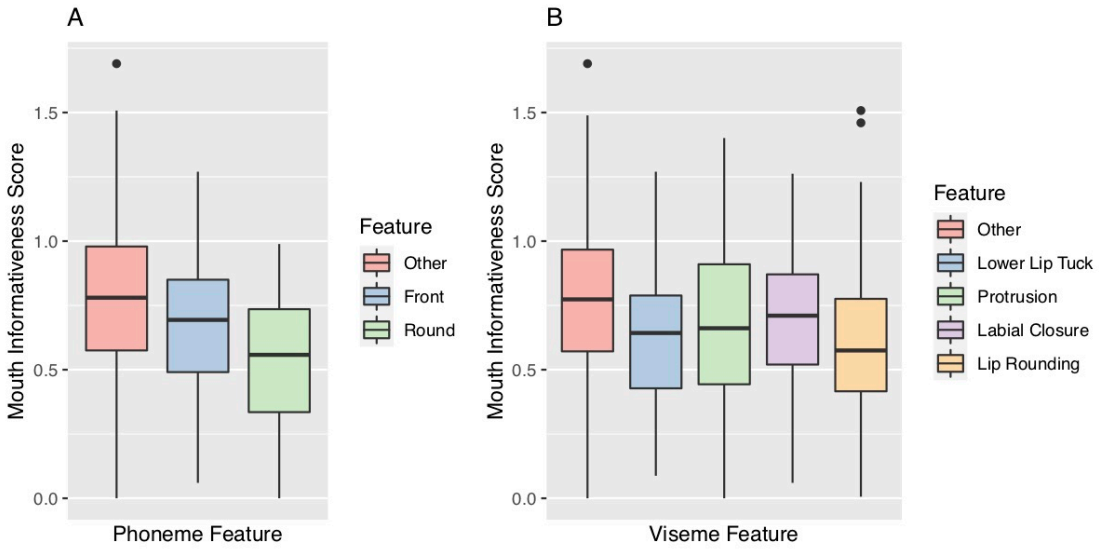


Figure 2.

Mean distances with informative phoneme (A) and viseme (B) features in the initial position for the combined British-American corpus. Lower distance indicates more informative features.

3.4.2 Exploratory analyses

We tested whether the distance is modulated by informative mouth features in non-initial position. We found that in addition to the initial position, words with phonemes with rounding or frontness features in a non-initial position led to a lower distance, suggesting more informative mouth movements. Model comparisons (Model 1.1 vs. Model 2.1) showed that the model including informative phonemes in both initial and non-initial positions was significantly better ($F_{courpus} = 79.98$, $p<.001$; $F_{British} = 43.96$, $p<.001$; $F_{American} = 53.87$, $p<.001$). This pattern is similar across the corpus ($B_{rounding} = -0.27$, $t_{rounding} = -5.10$, $p_{rounding} <.001$; $B_{frontness} = -0.62$, $t_{frontness} = -11.72$, $p_{frontness} <.001$), for the British slice ($B_{rounding} = -0.18$, $t_{rounding} = -2.62$, $p_{rounding} = 0.009$; $B_{frontness} = -0.62$, $t_{frontness} = -9.11$, $p_{frontness} <.001$) and the American slice (B

$_{rounding}$ = -0.45, t $_{rounding}$ = -6.58, p $_{rounding}$ <.001; B $_{frontness}$ = -0.58, t $_{frontness}$ = -8.13, p $_{frontness}$ <.001).

Similarly, when looking at visemes, we found that lower lip tuck, labial closure and lip rounding in the non-initial position predict mouth informativeness suggesting that words with these features have more informative mouth movements. The pattern was overall similar across the corpus (B $_{lowerLipTuck}$ = -0.96, t $_{lowerLipTuck}$ = -11.72, p $_{lowerLipTuck}$ <.001; B $_{labialClosure}$ = -0.48, t $_{labialClosure}$ = -8.24, p $_{labialClosure}$ <.001); for the British slice (B $_{lowerLipTuck}$ = -0.84, t $_{lowerLipTuck}$ = -8.51, p $_{lowerLipTuck}$ <.001; B $_{labialClosure}$ = -0.49, t $_{labialClosure}$ = -6.34, p $_{labialClosure}$ <.001) and the American slice (B $_{lowerLipTuck}$ = -1.00, t $_{lowerLipTuck}$ = -7.76, p $_{lowerLipTuck}$ <.001; B $_{labialClosure}$ = -0.46, t $_{labialClosure}$ = -5.98, p $_{labialClosure}$ <.001). To note, the effect of lip rounding in the American slice is only marginally significant (B $_{lipRounding}$ = -0.13, t $_{lipRounding}$ = -1.83, p $_{lipRounding}$ = 0.068), compared with the British slice (B $_{lipRounding}$ = -0.17, t $_{lipRounding}$ = -2.63, p $_{lipRounding}$ = 0.009) or entire corpus (B $_{lipRounding}$ = -0.16, t $_{lipRounding}$ = -3.07, p $_{lipRounding}$ = 0.004). We refrain from further interpretation as this may be associated with the smaller number of words in the American slice.

Interestingly, although we did not find evidence that protrusion feature in the initial position significantly affects mouth informativeness, the presence of protrusion in the non-initial position makes word recognition easier, which is true across the corpus (B $_{protrusion}$ = -0.25, t $_{protrusion}$ = -4.11, p $_{protrusion}$ <.001), for the British slice (B $_{protrusion}$ = -0.26, t $_{protrusion}$ = -3.21, p $_{protrusion}$ = 0.001) and the American slice (B $_{protrusion}$ = -0.28, t $_{protrusion}$ = -3.46, p $_{protrusion}$ = 0.001). Model comparisons (Model 1.2 vs. Model 2.2) indicated that the model with informative visemes in both initial and non-initial positions was significantly better (F $_{corpus}$ = 79.98, p<.001; F $_{British}$ = 43.96, p<.001; F $_{American}$ = 53.87, p<.001). See full results in table 4.

Table 4.

Full results for analysis in non-initial positions.

	Across Corpus				British Slice				American Slice			
Model 2.1: Phoneme feature in non-initial position												
	β	SE	t	p	*β*	SE	t	p	*β*	SE	t	p
Initial												
Rounding	-0.76	0.10	-7.50	<.001	-0.76	0.13	-6.05	<.001	-0.88	0.14	-6.25	<.001
Frontness	-0.43	0.05	-8.24	<.001	-0.38	0.07	-5.56	<.001	-0.60	0.07	-8.81	<.001
Non-initial												
Rounding	-0.27	0.05	-5.10	<.001	-0.18	0.07	-2.62	0.009	-0.45	0.07	-6.58	<.001
Frontness	-0.62	0.05	-11.72	<.001	-0.62	0.07	-9.11	<.001	-0.58	0.07	-8.13	<.001
Control												
Freq	-0.22	0.03	-8.41	<.001	-0.04	0.04	-1.14	0.255	-0.09	0.05	-1.89	0.060
PhonNBH	-0.05	0.03	-1.47	0.141	-0.04	0.04	-0.97	0.334	-0.05	0.04	-1.23	0.220
AoA	0.04	0.03	1.41	0.158	0.16	0.03	4.66	<.001	0.19	0.04	4.35	<.001
PhonNum	0.09	0.03	2.63	0.009	0.06	0.04	1.32	0.186	0.10	0.05	1.98	0.049
Comparison with M 1.1	F(2)=79.98, p<.001				F(2)=43.96, p<.001				F(2)=53.87, p<.001			
Adjusted R2	0.18				0.16				0.25			
Model 2.2: Viseme feature in non-initial position												
	β	SE	t	p	*β*	SE	t	p	*β*	SE	t	p
Initial												
LowerLipTuck	-0.64	0.09	-6.92	<.001	-0.56	0.12	-4.81	<.001	-0.80	0.13	-6.30	<.001
Protrusion	-0.13	0.10	-1.28	0.200	-0.11	0.14	-0.77	0.442	-0.21	0.14	-1.56	0.119
LabialClosure	-0.37	0.06	-6.37	<.001	-0.32	0.08	-4.16	<.001	-0.55	0.08	-7.25	<.001
LipRounding	-0.49	0.14	-3.54	<.001	-0.50	0.18	-2.82	0.005	-0.53	0.19	-2.86	0.004
Non-initial												
LowerLipTuck	-0.96	0.08	-11.72	<.001	-0.84	0.10	-8.51	<.001	-1.00	0.13	-7.76	<.001
Protrusion	-0.25	0.06	-4.11	<.001	-0.26	0.08	-3.21	0.001	-0.28	0.08	-3.46	0.001
LabialClosure	-0.48	0.06	-8.24	<.001	-0.49	0.08	-6.34	<.001	-0.46	0.08	-5.98	<.001
LipRounding	-0.16	0.05	-3.07	<.001	-0.17	0.07	-2.63	0.009	-0.13	0.07	-1.83	0.068
Control												
Freq	-0.21	0.03	-8.25	<.001	-0.06	0.04	-1.46	0.144	-0.07	0.05	-1.62	0.106
PhonNBH	-0.05	0.03	-1.54	0.125	-0.04	0.04	-0.95	0.343	-0.07	0.04	-1.48	0.138
AoA	0.06	0.03	2.39	0.017	0.17	0.03	5.03	<.001	0.20	0.04	4.63	<.001
PhonNum	0.09	0.03	2.63	0.009	0.07	0.04	1.54	0.123	0.08	0.05	1.59	0.113
Comparison with M 1.2	F(4)=52.88, p<.001				F(4)=29.18, p<.001				F(4)=27.19, p<.001			
Adjusted R2	0.20				0.17				0.25			

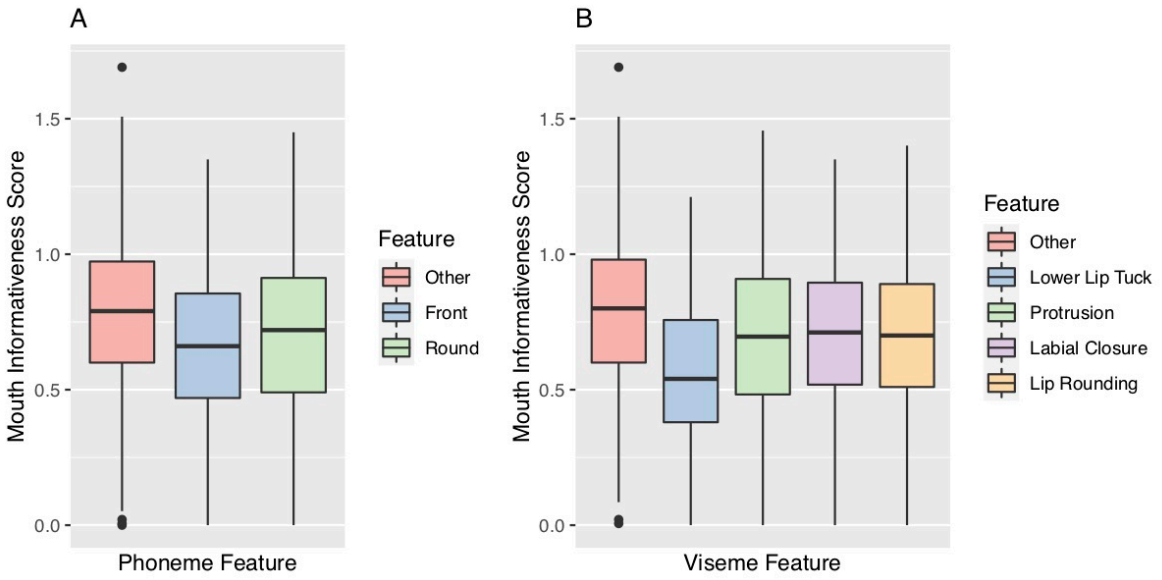


Figure 3.

Mean distances with informative phoneme (A) and viseme (B) features in non-initial position for the combined British-American corpus. Lower distance indicates more informative features.

Finally, we analysed whether informativeness load, or the number of informative features in a word, modulated mouth informativeness. We found that larger informativeness load calculated from both phonemes and visemes led to lower distance. The pattern is similar across the corpus ($B_{phoneme}$ = -0.32, $t_{phoneme}$ = -13.85, $p_{phoneme}$ <.001; B_{viseme} = -0.29, t_{viseme} = -12.45, p_{viseme} <.001), for the British slice ($B_{phoneme}$ = -0.29, $t_{phoneme}$ = -10.02, $p_{phoneme}$ <.001; B_{viseme} = -0.28, t_{viseme} = -9.46, p_{viseme} <.001) and the American slice ($B_{phoneme}$ = -0.39, $t_{phoneme}$ = -12.46, $p_{phoneme}$ <.001; B_{viseme} = -0.33, t_{viseme} = -10.29, p_{viseme} <.001). See table 5 for the full results. Figure 5 shows a scatterplot of informativeness load for the combined British-American corpus as an example.

Table 5

Full results from the exploratory analysis testing the impact of informativeness load on mouth informativeness.

	Across Corpus				British Slice				American Slice			
Model 2.3: Informativeness load (phoneme)												
	β	SE	t	p	β	SE	t	p	β	SE	t	p
InfoLoad_ Phoneme	-0.32	0.02	-13.85	<.001	-0.29	0.03	-10.02	<.001	-0.39	0.03	-12.46	<.001
Control												
Freq	-0.22	0.03	-8.37	<.001	-0.05	0.04	-1.21	0.228	-0.100	0.05	-2.18	0.030
PhonNBH	-0.01	0.03	-0.06	0.950	0.01	0.04	0.17	0.865	-0.011	0.04	-0.26	0.799
AoA	0.051	0.03	1.97	0.049	0.16	0.04	4.74	<.001	0.207	0.04	4.66	<.001
PhonNum	-0.01	0.04	-0.39	0.699	-0.02	0.05	-0.52	0.602	-0.04	0.05	-0.86	0.393
Adjusted R2	0.16				0.14				0.22			
Model 2.4: Informativeness load (viseme)												
	β	SE	t	p	β	SE	t	p	β	SE	t	p
InfoLoad_ Viseme	-0.29	0.02	-12.45	<.001	-0.28	0.03	-9.46	<.001	-0.33	0.03	-10.29	<.001
Control												
Freq	-0.22	0.03	-8.42	<.001	-0.06	0.04	-1.65	0.099	-0.08	0.05	-1.64	0.102
PhonNBH	-0.01	0.03	-0.26	0.793	0.01	0.04	0.16	0.872	-0.05	0.05	-1.17	0.242
AoA	0.07	0.03	2.85	0.004	0.19	0.03	5.34	<.001	0.23	0.05	4.93	<.001
PhonNum	-0.04	0.04	-1.05	0.294	-0.05	0.05	-1.03	0.305	-0.08	0.05	-1.67	0.095
Adjusted R2	0.14				0.13				0.17			

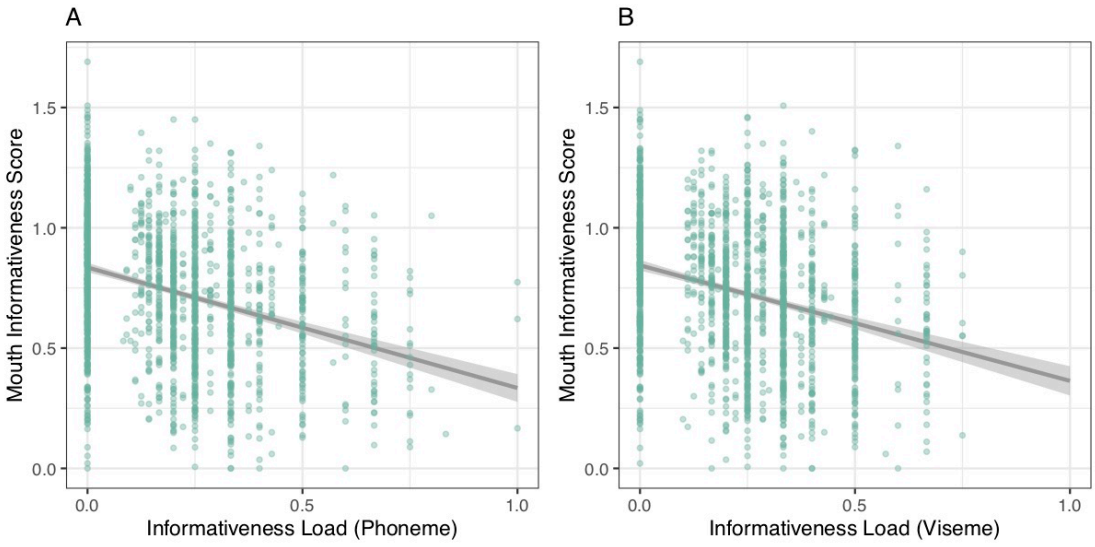


Figure 4.

Scatterplot of informativeness load, i.e., the number of informative features based on phoneme (A) and viseme (B) classifications as a predictor of mouth informativeness for the combined British-American corpus. Lower distance indicates more informative features.

3.5 Discussion

Mouth movements are always present in face-to-face communication, facilitating speech processing. Here, we present the first corpus of mouth informativeness for 1,743 English words, quantified using a novel approach to measure how identifiable words are based only on mouth information.

The confirmative analysis confirmed that our corpus captured informative visual features. Previous studies found phoneme and viseme features that are more visually identifiable (especially in initial positions), and our analysis showed that the presence of these features, namely frontness and rounding phonemes (/b/, /p/, /m/, /f/, /v/, and /r/, /w/, /u/, /o/, /ɔ/, respectively) or visemes characterized by lower lip tuck ({f}), labial closure ({p}), and lip rounding ({j}, {r}, {w}), lead to higher mouth informativeness, indexed by shorter distance between the target words and participants' guesses. Our findings replicated previous studies (e.g., Binnie et al., 1974; Benguerel & Pichora Fuller, 1982; Robert-Ribes et al., 1998; Traunmüller & Öhrström, 2007; Jesse & Massaro, 2010), indicating the reliability of our norms.

The exploratory analysis examined whether those informative features are also good predictors of mouth informativeness when presented in a non-initial position. We found that these features significantly predict mouth informativeness in both positions, and the model with these features in both positions outperformed the initial-only model. This indicates that apart from the beginning of a word (e.g., Karas et al., 2019) or in one-syllable words (e.g., Jesse & Massaro, 2010), the informative mouth features are generally helpful throughout the word. Indeed, the informativeness load per word (i.e. the number of informative features per word given its length) also predicts mouth informativeness, indicating that the more visible

features a word includes, the more informative they becomes based on corresponding mouth movements. Altogether, our results show that words such as 'woman', 'roof', 'further', 'mushroom', and 'ball' are highly informative in the visual context (all have a mouth informativeness score below 0.1) and can inform a perceiver to a larger extent than words without these features, e.g., 'gun', 'leek', 'hang', 'example', and 'neck' (all have a mouth informativeness score above 1.0).

Interestingly, in contrast to Jesse and Massaro (2010), we found that lip protrusion (including viseme {ch} and {w}) leads to higher mouth informativeness only in non-initial position, but not initial position. The reason for this difference remains unknown due to the exploratory nature of the analysis. However, note that viseme {w} in the lip protrusion category also appears in the lip rounding feature category. Therefore, the effect of protrusion features (for viseme {w} at least) may be captured by other features, thus leaving protrusion with less statistical power. Alternatively, this may be associated with the coarticulation that naturally occurs for longer words in our corpus. As is observed from the norm, the effect of protrusion in word-initial position may be largely context-dependent, e.g., 'change' is highly informative, but 'chick' is a lot less so. This may result in higher variance, deeming the effect less significant.

Since viseme directly captures mouth movements' visual features, it should theoretically better predict the mouth informativeness scores. However, models with viseme features produced similar results to the phoneme-based models in model fits. As some features overlap in terms of their phonemes (e.g. phonemes /f/ /v/ are produced frontally, but also with lower lip tuck features), it may suggest that both types of analysis capture similar mouth patterns.

Finally, provided that our mouth informativeness measure is based on silent lipreading, one can argue that the method is prone to large individual variability in lipreading skills as well as differences in pronunciation, not only across English accents, but also across speakers. Here we showed that our norms for both the British and the American corpora accurately, and in a similar way (despite differences in pronunciation between the two accents) capture the informativeness of mouth movement patterns that are specific to individual words. To note, the British and American slices of the corpus are produced by two separate actresses. Therefore, it remains unknown whether any difference between the two slices are due to accents or individual pronunciation styles. Further research is needed to investigate more thoroughly the accent and speaker-related differences, which is beyond the scope of the present study.

Apart from its potential for lipreading researchers, this corpus can also benefit language and psychology researchers in general. Words are common units in experimental designs and psycholinguistic analysis. For any studies of audiovisual speech processing using words as a basic unit, the mouth informativeness value can be incorporated into the statistical model to account for the impact of mouth movements without the need to manipulate their presence. As will be introduced in Chapters 4 and 5, the mouth informativeness norm is proved useful for the studies about language processing in naturalistic materials and can affect how the listener processes linguistic information.

Reference

1. Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology (London, England: 1953)*, *92 Part 2*, 339–355.

2. Auer Jr, E. T. (2009). Spoken word recognition by eye. *Scandinavian Journal of Psychology*, *50*(5), 419–425. https://doi.org/10.1111/j.1467-9450.2009.00751.x

3. Auer, E. T., & Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *The Journal of the Acoustical Society of America*, *102*(6), 3704–3710. https://doi.org/10.1121/1.420402

4. Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

5. Benguerel, A.-P., & Pichora-Fuller, M. K. (1982). Coarticulation effects in lipreading. *Journal of Speech & Hearing Research*, *25*(4), 600–607. https://doi.org/10.1044/jshr.2504.600

6. Binnie Carl A., Montgomery Allen A., & Jackson Pamela L. (1974). Auditory and Visual Contributions to the Perception of Consonants. *Journal of Speech and Hearing Research*, *17*(4), 619–630. https://doi.org/10.1044/jshr.1704.619

7. Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 46(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

8. Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The Natural Statistics of Audiovisual Speech. *PLoS Computational Biology*, *5*(7), e1000436. https://doi.org/10.1371/journal.pcbi.1000436

9. 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. https://doi.org/10.1016/j.bandl.2018.01.003

10. 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. https://doi.org/10.1111/cogs.12789

11. Drijvers, L., & Özyürek, A. (2017). Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

12. Fisher, C. G. (1968). Confusions Among Visually Perceived Consonants. *Journal of Speech and Hearing Research*, *11*(4), 796–804. https://doi.org/10.1044/jshr.1104.796

13. Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3 Pt 1), 1197–1208. https://doi.org/10.1121/1.1288668

14. Hirata, Y., & Kelly, S. D. (2010). Effects of Lips and Hands on Auditory Learning of Second-Language Speech Sounds. *Journal of Speech, Language, and Hearing Research*, *53*(2), 298–310. https://doi.org/10.1044/1092-4388(2009/08-0243)

15. Holle, H., & Gunter, T. C. (2007). The Role of Iconic Gestures in Speech Disambiguation: ERP Evidence. *Journal of Cognitive Neuroscience*, *19*(7), 1175–1192. https://doi.org/10.1162/jocn.2007.19.7.1175

16. Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics*, *72*(1), 209–225. https://doi.org/10.3758/APP.72.1.209

17. Karas, P. J., Magnotti, J. F., Metzger, B. A., Zhu, L. L., Smith, K. B., Yoshor, D., & Beauchamp, M. S. (2019). The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. *ELife*, *8*. https://doi.org/10.7554/eLife.48116

18. Krason, A., Fenton, R., Varley, R., & Vigliocco, G. (2020). *The Role of Iconic Gestures and Mouth Movements in Face-to-Face Communication* [submitted].

19. Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

20. Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, *19*(1), 1–36.

21. Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space. *PLoS ONE*, *4*(3), e4638. https://doi.org/10.1371/journal.pone.0004638

22. Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle* (pp. xii, 500). The MIT Press.

23. Massaro, D. W, Cohen, M. M, Tabain, M, Beskow, J, and Clark, R. "Animated Speech." Audiovisual Speech Processing. Cambridge UP, 2012. 309-45. Web.

24. Mattys, S. L., Bernstein, L. E., & Auer, E. T. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception & Psychophysics*, *64*(4), 667–679. https://doi.org/10.3758/BF03194734

25. McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: behavioural and neural interactions. Neuropsychologia, 50(5), 762-776.

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

27. Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. (2016). PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 3475–3484. https://www.aclweb.org/anthology/C16-1328

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

29. Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In *Hearing by eye: The psychology of lip-reading* (pp. 97–113). Lawrence Erlbaum Associates, Inc.

30. Rennig, J., Wegner-Clemens, K., & Beauchamp, M. S. (2020). Face viewing behavior predicts multisensory gain during speech perception. *Psychonomic Bulletin & Review*, *27*(1), 70–77. https://doi.org/10.3758/s13423-019-01665-y

31. Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise. *The Journal of the Acoustical Society of America*, *103*(6), 3677–3689. https://doi.org/10.1121/1.423069

32. Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex (New York, N.Y.: 1991)*, *17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

33. Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69–B78. https://doi.org/10.1016/j.cognition.2004.01.006

34. Schubotz, L., Holler, J., Drijvers, L., & Özyürek, A. (2020). Aging and working memory modulate the ability to benefit from visible speech and iconic gestures during speech-in-noise comprehension. *Psychological Research*. https://doi.org/10.1007/s00426-020-01363-8

35. Scott, S., Rosen, S., Spitsyna, G., Faulkner, A., Neville, L., & Wise, R. (2002). The neural basis of cross modal enhancement in speech perception: A pet study. In Society for Neuroscience.

36. Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

37. Traunmüller, H., & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*, *35*(2), 244–258. https://doi.org/10.1016/j.wocn.2006.03.002

38. Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and Visual Lexical Neighborhoods in Audiovisual Speech Perception. *Trends in Amplification*, *11*(4), 233–241. https://doi.org/10.1177/1084713807307409

39. Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception, & Psychophysics*, *79*(2), 396–403. https://doi.org/10.3758/s13414-016-1238-9

Chapter 4

4 Multimodal cues jointly modify linguistic predictions

4.1 Introduction

There is increasing evidence for predictions as a mechanistic account of brain function in general (e.g. Arnal, Wyart and Giraud, 2011; Clark, 2013; Friston and Kiebel, 2009). For language comprehension, in particular, it has been hypothesised that listeners construct predictions of upcoming words based on previous linguistic context (see review in Kuperberg & Jeager, 2016). The N400, a negative-going event-related potential (ERP) recorded from centro-parietal electrodes 200-600ms after word onset, is a biological marker of semantic processing difficulty, at least partly associated with word predictability (Kutas & Federmeier, 2011). Words that are less predictable in linguistic context elicit more negative N400, either outright incompatible with the context (e.g. Kutas & Hillyard, 1980) or less probable given the linguistic context (e.g. Kutas & Hillyard, 1984). Recent studies showed that surprisal, a computational measure of how unpredictable a word is given the linguistic context, modulates N400, with higher surprisal reliably predicting more negative N400 (with one standard deviation increase of surprisal causing around 0.2uV change in N400 amplitude, Frank et al., 2015; Frank, 2017). While it remains debatable whether this surprisal effect on N400 reflects the prediction or integration process, it is clear that less predictable words are more difficult to process. However, very little is known about whether and how multimodal cues jointly modify the effect of linguistic

predictability when comprehending more naturalistic materials where multimodal cues co-occur with one another. Therefore, in this chapter, we present the first investigation (and a direct replication) of the EEG signature of multimodal communication using naturalistic style materials. We ask two main questions: 1) do multimodal cues always **modulate** language comprehension in more naturalistic materials? 2) Are comprehenders sensitive to the **interactions** between multimodal cues? This study has been published as Zhang, Frassinelli, Tuomainen, Skipper & Vigliocco (2021).

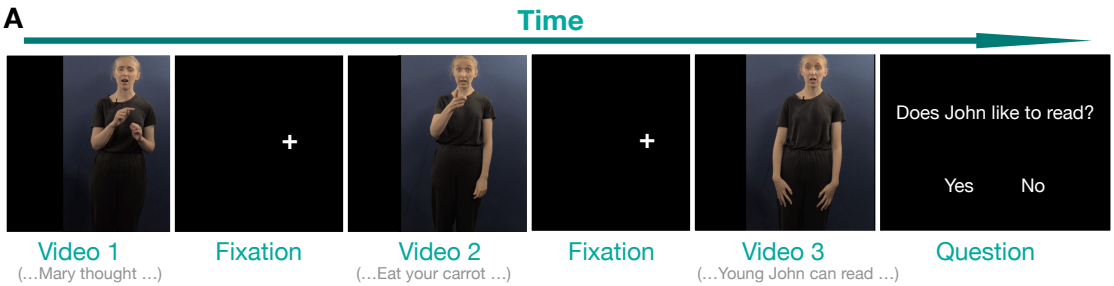
Previous studies suggest that multimodal cues may modify the effect of linguistic predictability when presented individually. As reviewed in Chapter 2, less predictable information introduced by prosodic stress elicits smaller N400 amplitude than unstressed words (e.g. Magne et al, 2005), indicating that prosodic information was taken into account and made new information more predictable. Meaningful co-speech gestures (i.e. gestures that carries direct meaning, such as iconic gesture, deictic gestures or emblematic gestures) can also increase the predictability of upcoming words by providing associated semantic information. For example, activating the less predictable meaning of the homonymous word “ball” using a “dancing” gesture, reduces the N400 response to a later mention of “dance” (Holle and Gunter, 2007; Obermeier et al., 2011; Obermeier et al., 2012). Very few studies have investigated how beat gestures modified the effect of linguistic predictability or N400 (but see Wang and Chu, 2013), and similarly, studies of how mouth movements affect word predictability are also inconclusive (Brunellière et al., 2013; Hernández-Gutiérrez et al., 2018).

However, most previous studies focused on single cues (usually linguistic cues) despite the evidence that multimodal cues modulates language

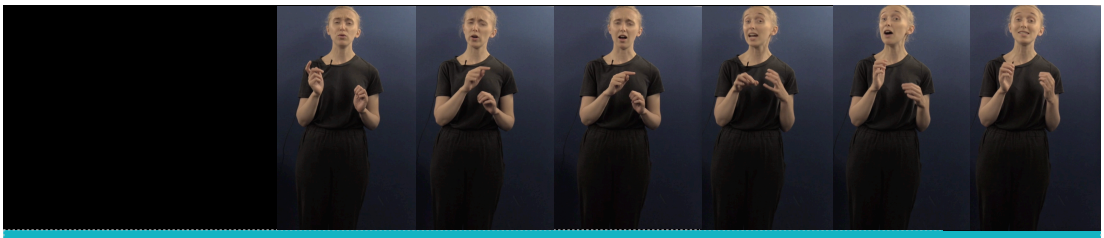
comprehension and that multimodal cues typically co-occur. One important reason for this pertains to the challenge of doing experimental research with naturalistic materials (Alday, 2019). For example, it is difficult to account for the confounding variables in naturalistic materials (e.g. when speakers produce gestures, their speech tend to also have higher pitch) without modern statistical techniques such as mixed-effect models, which only became popular in psychology in the recent years. Thus, many have used a reductionist approach to ensure experimental control (e.g., “normalising” prosody and avoiding audiovisual presentation when studying speech; showing only the mouth when studying audiovisual speech perception; hiding the face when studying gestures). This approach, however, risks breaking the natural (and predictive) correlation among cues with unknown consequences on processing (Hasson et al., 2018; Skipper, 2015).

Here we ask whether/how the multimodal cues affect the impact of linguistic predictability in more naturalistic settings. In doing so, we also address the methodological issues above. Across an initial study and a subsequent replication, we measured the electrophysiological responses to each cue and their interactions, elicited by words in naturalistic speech videos (see figure 1). In Experiment 1, we first established the effect of word predictability in naturalistic speech. We quantified predictability per content word using surprisal, measuring how unpredictable a word is given prior linguistic context, and then identified the EEG time window sensitive to surprisal. Then, we quantified prosody, gesture and mouth informativeness per word, and analysed how these cues and their interactions modulates surprisal effects in a linear-mixed-effect model. In Experiment 2, we replicated the process with different participants and materials to test the robustness of the impact of multimodal cues. We ask two questions in the current study: 1) to what extent is the processing of

multimodal cues central to comprehension of naturalistic style speech? We address this question by assessing whether individual cues modulate the word predictabilities based on linguistic context, indexed by N400 amplitude. Based on previous results, we predict that N400 amplitude will be bigger for less predictable (higher surprisal) words but smaller when informative multimodal cues are present (e.g. meaningful gestures). If more than one multimodal cue modulates N400, and such modulations are replicable, we will conclude that multimodal cues are central to language comprehension. Whereas, statistical outcomes of only one cue affecting N400 will be treated as evidence that this particular cue (but not multimodal cues in general) is central to comprehension; and statistical outcomes of unreplicable effects of multimodal cues will be deemed as evidence that multimodal cues do not reliably modulate comprehension and are thus not central to comprehension. Second, we ask what the dynamics of online multimodal comprehension are. We answer this question by analysing the interactions between multimodal cues. If the presence of other cues actively modifies the impact of a certain cue, then the listeners dynamically change the weight assigned to multimodal information depending on the context.



B



	Mary	thought	of	joining	Clara.	They	might	pick	flowers	together.
Surprisal	-	6.73	-	9.02	-	-	-	8.63	9.41	7.40
Pitch 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.

Design of the current experiments. Participants watched videos of an actress uttering short passages and answered comprehension questions. We then quantified the linguistic predictability, prosody, gestures and mouth movements per word and analysed how they jointly affect N400.

4.2 Experiment 1

4.2.1 Methods

4.2.1.1 Participants

36 Native English speakers with normal hearing and normal/corrected to normal vision were paid £7.5/hour to participate after giving written consent. 31 participants were included in the analysis (mean age = 27, 17 women) while 5 participants were excluded for technical issues. N was decided on the basis of the previous study by

Frank et al (2015). They used 24 subjects and found a significant effect of surprisal (computed as in the present paper) in word-by-word reading. We decided to increase N because of the difference in presentation modality. All methods were approved by University College London (UCL; Research Ethics Committee 0143/003).

4.2.1.2 Materials

103 naturalistic passages (unrelated to one another) were selected from the British National Corpus. Two-hundred and forty-six naturalistic passages (containing two consecutive sentences) were initially extracted from the British National Corpus (BNC, University of Oxford, 2007). In particular, we used the BNC (and not a web corpus) because it offers more standard sentences. BNC contains 100 million words of language material selected from both written (90%) and spoken language (10%). In order to obtain complete grammatically valid sentences we excluded the spoken part. In the written part there are newspapers, published articles and novels. Passages were selected in a semi-random fashion with the only constraints that 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 passages for grammaticality, meaningfulness and gesturability on a 1-5 likert scale. We selected 103 passages that had averaged gesturability > 2 (and SD < 2.5); and had no grammatical errors or semantic anomalies. 3 passages were used as practice trials while 100 are included as test trials. The averaged number of words in the included passages is 23, and the mean duration of the resulting videos is 8.50s.

A native British English-speaking actress produced the passages with natural speed, prosody and facial expressions. Thus, although not fully naturalistic, our materials preserve the natural co-occurrences among the different cues. The actress has given informed consent for publication of identifying information. The onset and offset of each word were automatically detected using a word-phoneme aligner based on a Hidden Markov Model (Rapp, 1995) and was checked manually (word duration: Mean=440ms, SD=376ms). For each content word (i.e., nouns, adjectives, verbs and adverbs) we quantified the informativeness of each cue (linguistic predictability, pitch prosody, gesture and mouth, see 4.2.1.3 for quantification methods). Function words (i.e., articles, pronouns, auxiliary verbs and prepositions) were excluded because Frank and colleagues failed to show any effect of the predictability (measured as surprisal) for such words (Frank et al., 2015).

We recorded two versions of each passage: one in which she was instructed to gesture freely and one in which she was instructed not to gesture. In the analyses we compare the same word across with/without gesture conditions. In contrast to other cues that is present for each word (e.g. prosody or mouth movements), gestures are not always produced and words likely to be accompanied by meaningful gestures (e.g., *combing*) are semantically very different from words that are not (e.g., *pleasing*) and these differences, unrelated to surprisal differences, could nonetheless be confounded. Thus, comparison of the same words produces clearer results. There are small differences between the with/without gesture videos. a) Duration: words with gestures tend to be ~15ms longer than words without (no gesture videos=437.28ms, gesture videos=455.12ms, pairewise t test not significant, p=0.14). b) Pitch prosody value (mean F0): pitch prosody tends to be slightly higher in videos with gesture than without (no gesture videos = 295.90Hz, gesture videos = 300.79Hz, pairewise t test not

significant, p=0.12). c). Mouth informativeness: we cannot assess whether the mouth informativeness differs across gesture conditions, as mouth informativeness is measured separately in a single word recognition task, which is independent from the current study. All differences above are very small and not statistically significant, so we do not believe may impact our results.

4.2.1.3 Quantification of cues

Linguistic predictability for each word was measured using surprisal (Mean=7.92, SD=2.10), defined as the negative log-transformed conditional probability of a word given its preceding context (Shannon, 1949). Surprisal provides a good measure of predictability and predicts reading times (e.g. Smith & Levy, 2013) and N400 amplitude (e.g. Frank et al., 2015). Here, surprisal was generated using a bigram language model trained on the lemmatized version of the first slice (~31-million tokens) of the ENCOW14-AX corpus (Schäfer, & Bildhauer, 2012). Previous studies found that surprisal derived by an n-gram model can predict the N400 amplitude per word in written (Frank et al., 2015) and audio stimuli (Alday et al., 2017). 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 (Frank et al., 2015). Therefore, we chose a bigram model to reduce data sparsity and, consequently, increase the robustness of our surprisal measures. Once trained, the model was used to calculate the surprisal of each word in based on previous content words as below:

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

where w_{t+1} indicates the current word, and $w_{1\ldots t}$ stands for previous content words. We also compared surprisal generated from a window size of 1, 2, 3, 4, 5 and

all previous words (see 4.2.2.3 for more details). Given the minor differences, we calculated surprisal based on all previous words to produce our final results.

Pitch prosody per word was quantified as mean F0 (Mean=298Hz, SD=84Hz) extracted using Praat (version 6.0.29, Boersma, 2001). Apart from mean F0, other acoustic properties (e.g. minimum F0, maximum F0, mean intensity and F0 change) may also represent prosodic accentuation. Therefore, we also compared the result of these different operationalisations. We ran separate linear mixed effect model with each of the above variable as a predictor measuring prosody changes, while keeping other multimodal cues constant. The results obtained from different operationalisations were very similar. Therefore, we selected mean F0, which is commonly used to represent prosody (e.g. Kakourous, Salminen & Räsänen, 2018), in all our following analysis in both experiments.

Gestures were coded as meaningful gestures or beats by two expert coders in ELAN (version 5.0.0, Sloetjes & Wittenburg, 2008). Meaningful gestures (Exp.1: N=359; Exp.2) comprised iconic gestures (e.g. drawing movements for the word “drawing”) and deictic gestures (e.g. pointing to the hair for “hair”). Beat gestures (Exp.1: N=229) comprised rhythmic movements of the hands without clear meaning (McNeill, 1992). Coders annotated the category, phases and lexical affiliate (meaningful gesture only) of gestures. To associate words with gestures, two variables, meaningful gesture and beat gesture, were then created. Words received 1 for meaningful gesture if it is the lexical affiliate of a corresponding meaningful gesture, and 1 for beat gesture if it overlapped with the stroke of a beat gesture. Another expert coder annotated 10% of the videos to check for reliability (inter-rater agreement=95.3, kappa=0.922, $p<.001$).

Mouth informativeness (Exp.1: mean= 0.65, SD=0.28) per word was extracted from the mouth informativeness corpus described in Chapter 3. An actress (who produced the Briths English words in the corpus in Chapter 3, and also all stimuli in the current study) produced individual words, and participants from the online study watched each word twice and guess the words based on the mouth shape. Every word was rated by 10 participants. The phonological distance between the guesses and the target words were calculated. We then reversed the distance so that larger value indicates more accurate guess thus higher mouth informativeness.

4.2.1.4 Procedure

Participants watched videos (N=100) presented using Presentation software (V. 18.0), counterbalanced for gestures presence, while their EEG responses were recorded. Videos were separated by a 2000ms interval. Participants were instructed to watch the videos attentively and answer comprehension questions following some of the videos. The questions are about the content of the immediately preceding passages (e.g. Passage: “Emma screamed and swore at them. She was especially angry if the girls dared to eat any of her food or drink her coffee.”, Question: “Is Emma going to share her sweets with the other girls?”). Out of 100 passages, 35 questions were presented, accounting for 35% of the total passages (14 are Yes and 21 are No). Participants were instructed answer the questions, when they were presented, as quickly and accurately as possible (prioritizing accuracy) by pressing the left (“Yes”) or right (“No”) control key. Participants sat ~1m away from the screen (resolution=1024*768) with 50Ω headphones. They were asked to avoid moving, keep their facial muscles relaxed and reduce blinking (when comfortable). The recording took ~30 mins in Exp.1.

4.2.1.5 EEG recording and preprocessing

A 32-channel BioSimi system (Ag/AgCl electrodes, 24 bit resolution, 10-10 international system layout) was used to collect EEG data. A common reference included the CMS electrode and DRL electrode. Elastic head caps were used to keep the electrodes in place. Two external electrodes were attached (left/right mastoids) for off-line reference, while two other external eye electrodes were attached (below left eye and on right canthus) to detect blinks and eye movements. Electrolyte gel was inserted to improve connectivity. 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.

Raw data were pre-processed with EEGLAB (version 14.1.1) and ERPLAB (version 7.0.0) in MATLAB (R2017b). All electrodes were included. Triggers were sent per video, and word onsets were calculated from the word boundary annotation. Any lag between trigger and stimuli presentation was also measured and corrected (Mean= 210.33ms, SD=69.92). The EEG files were re-referenced to average of the mastoids, down-sampled to 256Hz, separated into epochs (-100 to 1200ms), and filtered with a 0.05-30Hz band-pass filter. Due to the likely overlap between baseline (-100 to 0ms) and the EEG signal of the previous word, we did not perform baseline correction, but instead extracted the mean EEG amplitude in baseline interval and later included it in the regression model as control (Frank et al., 2015; Alday et al., 2017). We conducted independent component analysis to label and remove noise components (e.g. eye movement, heart beat), and artifact rejection using moving window peak-to-peak analysis (Voltage Threshold=100 μV, moving window full width=200 ms, window step=20 ms) and 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 in Exp.1, and 12.18% (SD=14.43) in Exp.2.

4.2.2 Establishing the effect of surprisal

In order to investigate whether/how multimodal cues modulate the effect of linguistic predictability, we first established that surprisal affected language comprehension in our naturalistic style materials, containing various multimodal cues. Although previous studies found that surprisal derived by an n-gram model can predict the N400 amplitude per word in written (Frank et al., 2015) and audio stimuli (Alday et al., 2017), both study included linguistic information only, therefore it remains unknown whether surprisal still predicts language comprehension in more naturalistic materials.

4.2.2.1 Behavioural effect of surprisal

We analysed whether averaged surprisal per passage affected the accuracy (Mean accuracy=82.1%, SD accuracy=0.384) and response time (Mean RT=4129.8 ms, SD RT=2881.3) for the 35 comprehension questions in order to examine the behavioural impact of surprisal. We constructed separate linear mixed effect (LMER) models for accuracy and response time respectively (binomial regression model for accuracy and linear regression model for response time, using the lme4 package, Bates et al., 2015). We included mean surprisal per passage (calculated by averaging the surprisal value of all content words) as the predictor variable, and participant and passageID as random intercepts to control for by participant and by passage variation. All continuous variables (response time and mean surprisal) was standardised using the “scale()” function embedded in R, which centres a variable and calculates the z-score, while all categorical variables (accuracy, participant and passageID) were sum coded.

We found that accuracy decreased with an increase in surprisal (β=-0.784, p<.001). Similarly, we found that sentences with higher averaged surprisal had slower reaction times (β=0.089, p=.024). The reaction time are overall longer (averaged around 4 seconds) because participants were instructed to prioritise accuracy. These findings confirm that sentences with higher surprisal were harder to process, indicating that surprisal predicts behavioural measures of language comprehension even in more naturalistic materials.

4.2.2.2 Time window of surprisal

We then establish the time-window where processing is affected by linguistic predictability, measured by surprisal per word. No previous study investigated surprisal effects in audiovisual communication. Therefore, rather than making a priori assumptions about the specific event-related response we should observe, we carried out a hierarchical LInear MOdeling (LIMO toolbox, Pernet, Chauveau, Gaspar & Rousselet, 2011) to identify the EEG component sensitive to surprisal. We selected hierarchical linear modelling, instead of more traditional Mass Univariate approach or simple visual inspection, because hierarchical linear modelling can better accommodate continuous variables (surprisal here). Hierarchical linear modelling (LIMO toolbox) carries out regression based EEG analysis (Smith & Kutas, 2015 a, b), decomposing the ERP signal into a time-series of beta coefficient waveforms elicited by continuous variables. Significant differences between the beta coefficient waveforms and zero (or a flat line, indicating that the variable does not affect EEG signal) represent the existence of an effect. Therefore, hierarchical linear modelling can identify time windows sensitive to surprisal without the need of dichotomising the variable and comparing between high and low surprisal groups (as would be required for visual inspection or Mass Univariate approach). Similar regression based approach

has been used in previous EEG studies investigating the effect of continuous variables (e.g. Rousselet et al., 2011; Broderick et al., 2018).

In this analysis, we first created a single-trial file from the EEG file for each participant, and a continuous variable containing surprisal of each word that this participant was presented with. In the first level analysis for each participant, the toolbox performed a regression analysis for each data point (sample, based on sampling rate, which is 512Hz in our case) in 0-1200ms 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 of the 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, Pernet et al., 2015). The resulting significant time window represents the interval where surprisal reliability modulates the EEG response.

As is shown in Figure 2, we found that words with higher surprisal elicited more negative EEG response in the 300-600ms time window especially in central-parietal areas. No other time window was significantly sensitive to surprisal. As a result, we focused on the 300-600ms time-window in our subsequent analyses in both experiments.

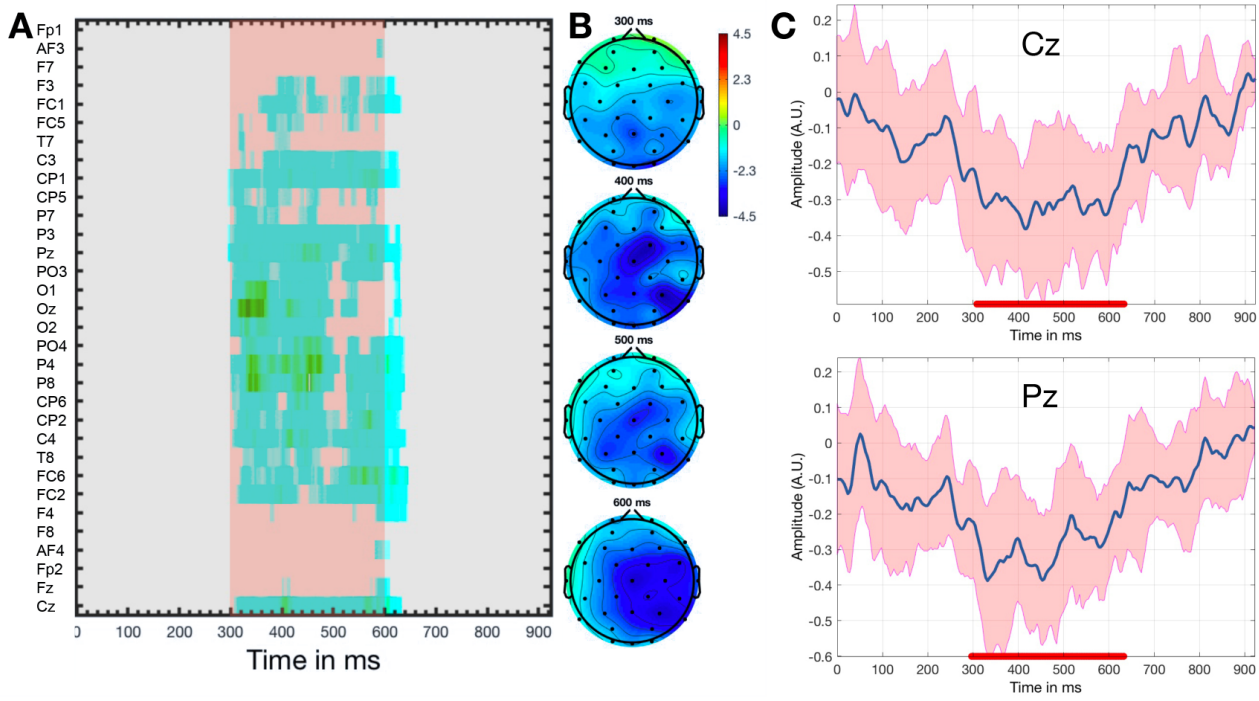


Figure 2.

Hierarchical linear modelling results showing the ERP sensitive to surprisal (one-sample t-test P<0.05, cluster-corrected). (A) Surprisal elicited a more negative ERP ~300-600ms (marked in pink) across most of the electrodes. Green areas are statistically significant while grey areas are not. (B) Surprisal effect is primarily central-parietal in the topographic maps for the 300-600 time window. Deeper blue area indicates more negative EEG response. The color bar on the right represents the scale of F values for the specific electrode. (C) Beta values for surprisal were significantly negative compared with 0 (flat waveform) in 300-600ms in Cz and Pz. 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)

4.2.2.3 Comparing surprisal calculated from different window size

Additionally, we compared the surprisal value generated based on different window sizes. For each word, we generated surprisal with varying window size n (n=1, 2, 3, 4, 5, and all), thus taking the previous n words into account when estimating the predictability of this word. In order to determine the appropriate window size, we first conducted a set of hierarchical linear modelling analysis with surprisal calculated from different window sizes as predictor variable, and then a set of multiple regression analysis with N400 (averaged ERP within 300-600ms) as the dependent variable, different surprisal and baseline ERP as independent variables (all continuous variables are standardised). As is shown in Figure 3, all different operationalisations of surprisal induced more negative ERP within around 300-600ms. As is shown in Table 1, all operationalisations were significantly negatively related with the N400 amplitude and generated similar statistics. Since the difference between the measures are minimal, we used n=all to generate surprisal in all our subsequent analysis in both experiments, thus calculating surprisal based on all previous content words.

Table 1.

Surprisal calculated from different window sizes has similar effect on N400 (300-600ms) amplitude in Experiment 1.

	β	**Std Error**	**T**	**P**	**R^2**
Surprisal 1	*-0.008*	*0.001*	*-8.479*	*<.001****	*0.613*
Surprisal 2	*-0.008*	*0.001*	*-9.150*	*<.001****	*0.613*
Surprisal 3	*-0.010*	*0.001*	*-11.080*	*<.001****	*0.613*
Surprisal 4	*-0.009*	*0.001*	*-10.150*	*<.001****	*0.613*
Surprisal 5	*-0.009*	*0.001*	*-10.050*	*<.001****	*0.613*
Surprisal All	*-0.009*	*0.001*	*-9.901*	*<.001****	*0.613*

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

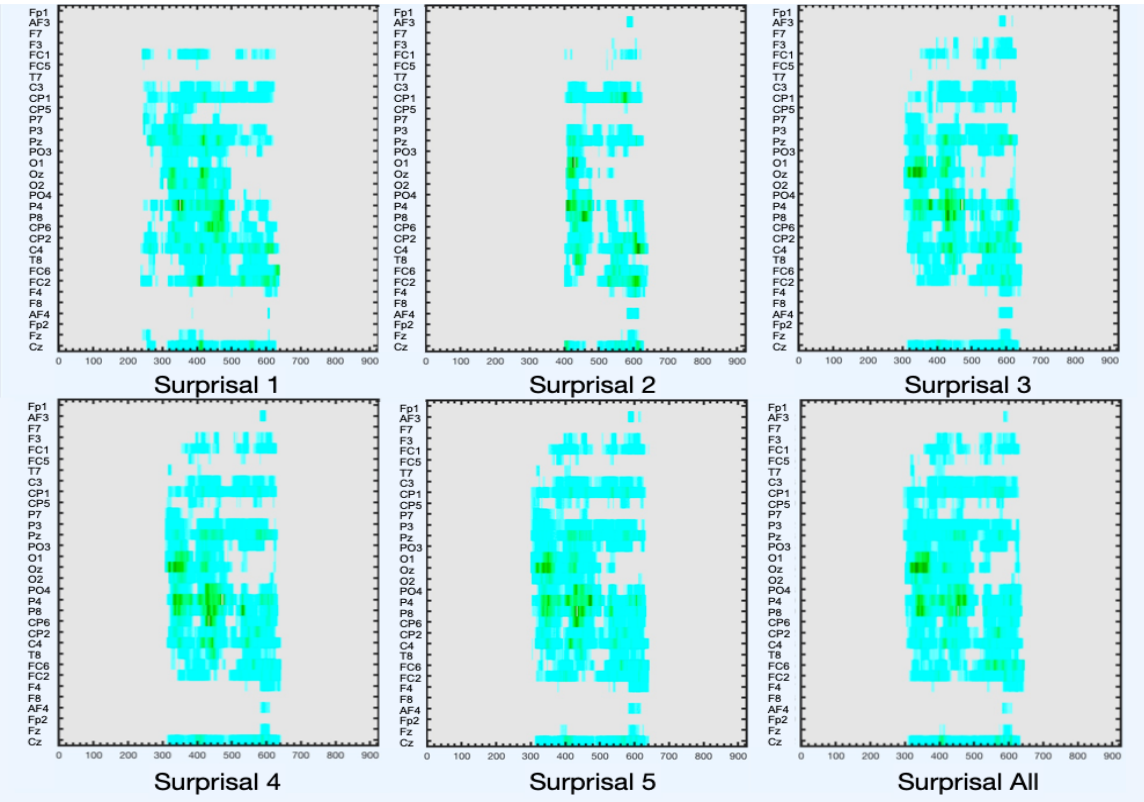


Figure 3.

Hierarchical linear modelling: surprisal generated from different window sizes all induced more negative ERP within approximately 300-600ms in Experiment 1.

4.2.3 Linear mixed effect regression analysis

After establishing that surprisal affects comprehension of naturalistic material and modulates the EEG response in 300-600ms time window, we then asked 1) do multimodal cues individually modulate the impact of surprisal and 2) what are the dynamics of multimodal cues. We addressed these questions using LMER analysis. We used LMER for its advantage in accommodating both categorical and continuous variables, thus increasing statistical power (MacCallum, Zhang, Preacher & Rucker, 2002). Mean ERP in the 300-600ms and -100-0ms time windows were extracted from 32 electrodes for each word as the dependent variable and the baseline. Due to

the likely overlap between baseline and the EEG signal of the previous word, we did not perform baseline correction during data pre-processing, but instead extracted the mean EEG amplitude in baseline interval and later included it in the regression model as control variables (Frank et al., 2015; Alday, Schlesewsky, Bornkessel-Schlesewsky, 2017). Independent variables included 1) predictors: surprisal, pitch prosody, meaningful gestures, beat gestures, mouth informativeness, and all up to three-way interactions between surprisal and cues, excluding any meaningful*beat gestures interactions (instances where the two gestures co-occur were removed), 2) control: baseline, word length, word order in the sentence, sentence order in experiment, and relative electrode positions measured by the X, Y and Z coordinates each coded as a variable (Winsler, Midgley, Grainger & Holcomb, 2018). We originally included frequency (derived from the ENCOW corpus) as control variable. However, frequency is removed from the final model due to multiple collinearity with surprisal, as both measures capture linguistic probability to different extents. Surprisal was log-transformed to normalize the data. All continuous variables were scaled so that coefficients represent the effect size. All categorical variables were sum coded so that coefficients represent the difference with the grand mean (intercept). We further included word lemma and participant as random variables. The maximal random structure failed to converge, so we included the highest interaction (three-way interactions) as random slope for participants (Barr, 2013), and surprisal as random slope for lemma. No predictors showed multicollinearity (VIF<2, kappa=4.871).

We excluded from the analyses: (a) words without a surprisal value. Very few words do not have any surprisal value (n=9), due to the lack of co- occurrence between this word and its context in the training corpus. (b) Words without a mean F0 score.

Very few words (n=4) showed pitch error when using Praat to automatically extract pitch (e.g. when the vowel is pronounced very quietly). c) Words with both meaningful and beat gesture (n=3). These instances usually represent the speaker producing a meaningful gesture but then a quick beat gesture for a word, or using one hand to produce a meaningful gesture but the other to produce a beat. Given the rarity of this phenomenon, we excluded them from the analysis thus removing any interaction between meaningful and beat gestures. d) Words occurring without any gesture in the "with gesture" condition, and the corresponding words in without gesture videos (n=406). This is to reduce the imbalance of the data, otherwise the without gesture condition would include not only the corresponding without gesture words of all with gesture words, but also all words in d), making this group ~3 times larger than the with gesture one. We compared the same item across the with/without gesture videos, instead of with or without gesture words in the with gesture videos only (different items). This is because words likely to be accompanied by meaningful gestures (e.g., *combing*) are semantically very different from words that are not (e.g., *pleasing*). Thus, comparison of the same words produces clearer results. After excluding the instances above, analysis of Exp.1 included 31 participants, 381 lemmas and 480,212 data points.

4.2.4 Results: Multimodal cues individually and jointly modulate comprehension

4.2.4.1 Are multimodal cues central to language processing?

To assess this question, we first focus on the main effects of the multimodal cues and their interaction with surprisal as predictors of N400 amplitude (see full results from LMER analysis in Table 2.).

Table 2.

Full result: linear mixed effects regression model on N400 (300-600ms) in Experiment 1.

Fixed Effects	β	SE	t	p
(Intercept)	*0.007*	*0.01*	*0.732*	*0.466*
Predictor Variables				
Surprisal	***-0.007***	***0.014***	***-0.502***	***0.616***
Mean F0	***0.011***	***0.002***	***5.262***	***< .001******
Mouth Informativeness	*0.013*	*0.008*	*1.652*	*0.1*
Meaningful Gesture (Present)	***0.006***	***0.001***	***5.232***	***< .001******
Beat Gesture (Present)	***-0.004***	***0.001***	***-2.5***	***0.012****
Surprisal:Mean F0	***0.022***	***0.003***	***7.467***	***< .001******
Surprisal:Mouth Informativeness	*0.018*	*0.015*	*1.221*	*0.224*
Surprisal:Meaningful Gesture (Present)	***0.007***	***0.001***	***4.653***	***< .001******
Surprisal:Beat Gesture (Present)	***-0.009***	***0.002***	***-5.345***	***< .001******
Mean F0:Mouth Informativeness	*-0.002*	*0.002*	*-1.553*	*0.12*
Mean F0:Meaningful Gesture (Present)	***0.005***	***0.001***	***4.148***	***< .001******
Mean F0:Beat Gesture (Present)	*-0.003*	*0.002*	*-1.924*	*0.054*
Mouth Informativeness:Meaningful Gesture (Present)	***0.004***	***0.001***	***3.072***	***0.002****
Mouth Informativeness:Beat Gesture (Present)	***0.012***	***0.002***	***8.005***	***< .001******
Surprisal:Mean F0:Mouth Informativeness	*0.008*	*0.007*	*1.164*	*0.252*
Surprisal:Mean F0:Meaningful Gesture (Present)	*0.001*	*0.006*	*0.094*	*0.926*
Surprisal:Mean F0:Beat Gesture (Present)	*0.004*	*0.006*	*0.581*	*0.565*
Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	***-0.016***	***0.006***	***-2.83***	***0.008****
Surprisal:Mouth Informativeness:Beat Gesture (Present)	*0.003*	*0.004*	*0.642*	*0.525*
Control Variables				
Word Order	***-0.011***	***0.002***	***-4.778***	***< .001******
Word Length	***-0.013***	***0.004***	***-2.847***	***0.004****
Sentence Order	***-0.007***	***0.001***	***-7.868***	***< .001******
Baseline	***0.788***	***0.001***	***862.209***	***< .001******
Electrode X	***-0.006***	***0.001***	***-6.491***	***< .001******
Electrode Y	***0.008***	***0.001***	***8.387***	***< .001******
Electrode Z	***0.001***	***0.001***	***1.271***	***0.204***

Random Effects		Variance	Std.Dev.
Lemma	*(Intercept)*	*0.012*	*0.108*
	Surprisal	*0.044*	*0.211*
Participant ID	*(Intercept)*	*0.001*	*0.034*

Surprisal:Mean F0:Mouth Informativeness	*0.001*	*0.039*
Surprisal:Mean F0:Meaningful Gesture (Present)	*0.001*	*0.03*
Surprisal:Mean F0:Beat Gesture (Present)	*0.001*	*0.033*
Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	*0.001*	*0.03*
Surprisal:Mouth Informativeness:Beat Gesture (Present)	*0*	*0.021*

Model: Experiment 1

AIC	*BIC*	*logLik*	*deviance*	*df.resid*
892159	*892746*	*-446026*	*892053*	*480159*

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

We found a main effect of pitch prosody (Panel A, Figure 4) (Exp1: β=0.010, p<.001): words with higher pitch prosody showed less negative EEG, or smaller N400 amplitude. The interaction between surprisal and pitch prosody (Exp1: β=0.017, p<.001) indicates that pitch prosody modulates the N400 response associated with surprisal: higher surprisal words showed a larger reduction of N400 amplitude when the pitch prosody was higher, in comparison to lower surprisal words.

Meaningful gestures showed similar effects (Panel B, Figure 4). Words accompanied by a meaningful gesture showed a significantly less negative N400 (Exp1: β=0.006, p<.001) and higher surprisal words elicited a larger reduction of N400 amplitude when meaningful gestures were present, in comparison to lower surprisal words (Exp1: β=0.008, p<.001; Exp2: β=0.011, p<.001).

In contrast, we found a significant negative main effect of beat gestures (Panel C, Figure 4, Exp1: β=-0.005, p=.001): words accompanied by beat gestures elicited a more negative N400. Moreover, higher surprisal words accompanied by beat gestures showed even more negative N400 compared with lower surprisal words (Exp1: β=-0.012, p<.001).

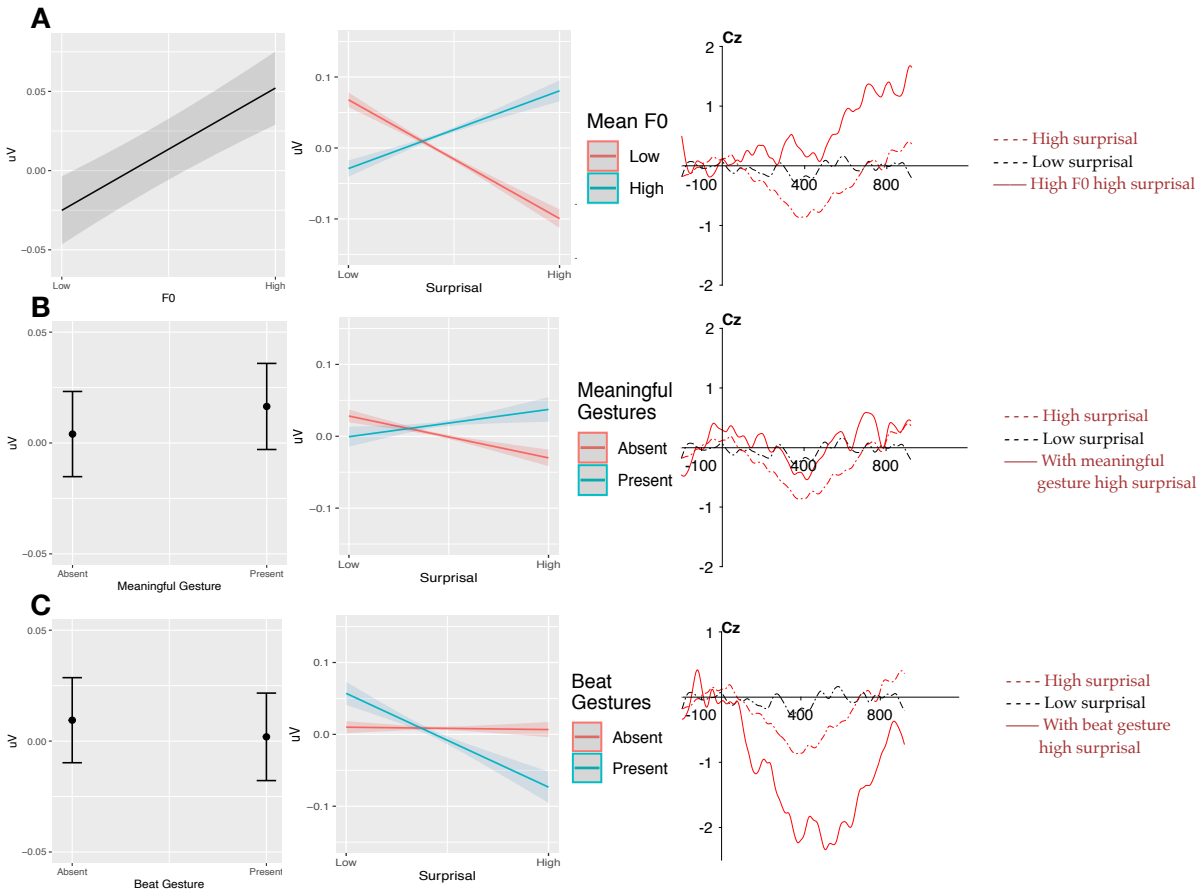


Figure 4.

Multimodal cues each modulate N400 and the impact of surprisal. Continuous variables are categorized in the EEG plots for illustration only (including F0 and surprisal, grouping the 1/3 words with the highest and lowest values into high/low categories). The same conventions apply for all plots below. A). Higher F0 induce less negative N400, especially for high surprisal words. B). Meaningful gestures induce less negative N400, especially for high surprisal words. C) Beat gestures induce more negative N400, especially for high surprisal words.

4.2.4.2 What are the dynamics of multimodal cue processing?

We found significant interactions between multimodal cues (Figure 5). First, we saw an interaction between pitch prosody (mean F0) and meaningful gesture: words

accompanied by meaningful gestures elicited even less negative N400 amplitude if their pitch prosody was higher (Exp1: β=0.004, p<.001). Second, the interactions between mouth informativeness and meaningful gesture (Exp1: β=0.004, p=0.002) and between mouth informativeness and beat gesture (Exp1: β=0.012, p<0.001) were also significant. Words with more informative mouth movement elicited less negative N400 when accompanied by either meaningful or beat gestures. The interaction between mouth informativeness and meaningful gestures is further affected by surprisal (Exp1: β=-0.016, p=0.008), indicating that the positive interaction is even stronger for words with lower surprisal.

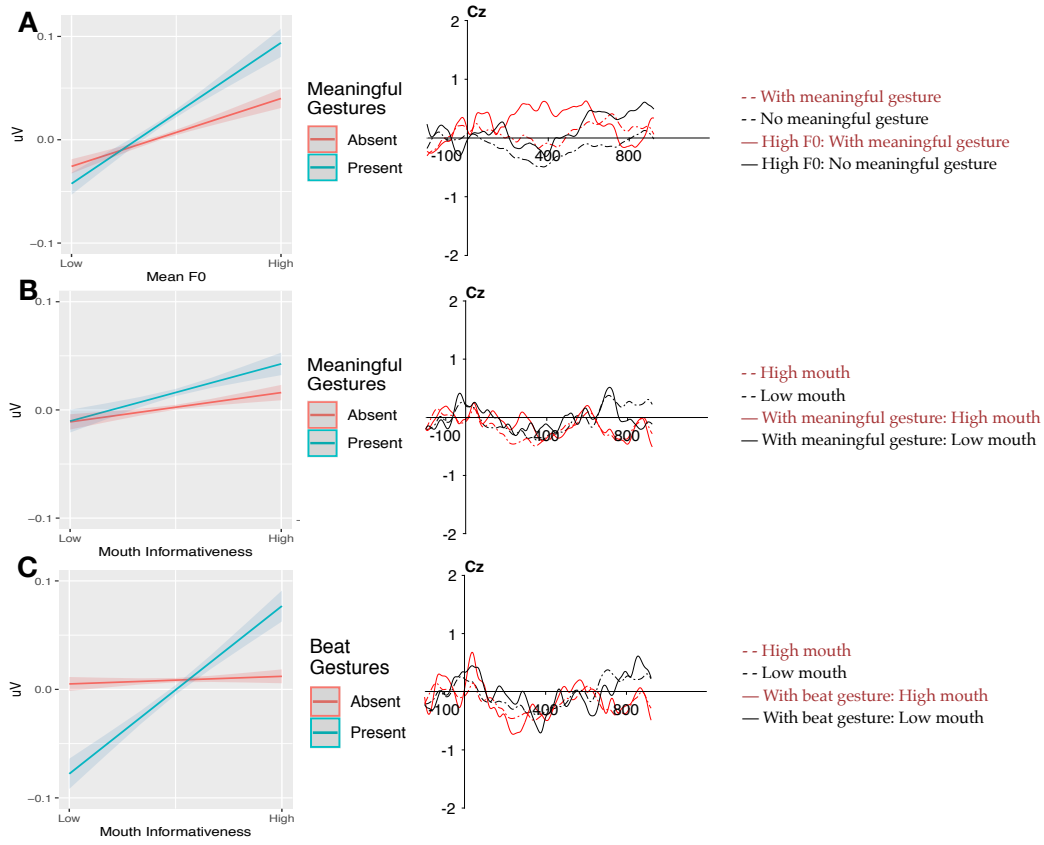


Figure 5.

Multimodal cues interact to affect N400. A) The positive effect of meaningful gestures is larger for higher F0 words; B) The positive effect of mouth movements is larger for

words with meaningful gestures; C) The positive effect of mouth movements is larger for words with beat gestures.

4.2.5 Discussion

In line with the previous studies, we confirmed that linguistic predictability, measured by surprisal, reliably predicts behavioural measures of comprehension per sentence as well as the ERP components per word within 300-600ms time window. However, more crucially, this N400 response is always modulated by each multimodal cues as well as their co-occurrence. While higher prosodic pitch and meaningful gestures induce less negative N400, especially for less predictable words, beat gestures induce the opposite effect. Moreover, the co-occurrence between higher pitch and meaningful gestures, as well as higher mouth informativeness with both gesture types, elicit even less negative N400. The interaction between mouth informativeness and meaningful gestures is further modulated by surprisal, being especially large for low surprisal words. Overall, these results show that the predictability of words based on linguistic context is *always* modulated by multimodal cues, and that the weight given to each cue depends on which other informative cues are present.

In the following Experiment 2, we tested the robustness of the findings above. We employed the same paradigm to test whether multimodal cues individually and jointly modulate language comprehension for a different set of participants and materials.

4.3 Experiment 2

4.3.1 Methods

4.3.1.1 Participants

20 Native English speakers (mean age = 25, 15 women) with normal hearing and normal/corrected to normal vision were paid £7.5/hour to participate after giving written consent. As Exp.2 has longer passages (on average 45 words per passage compared with 23 in Exp.1), we are able to obtain a similar number of observations with less participants. All methods were approved by University College London (UCL; Research Ethics Committee 0143/003).

4.3.1.2 Materials

In Exp.2, we chose 83 spoken passages from BBC TV scripts in order to further enhance the naturalness of the stimuli (as BNC corpus used in Exp.1 contains a large proportion of written materials, which may be different from the spoken language). One-hundred and ninty-six passages were initially extracted from the BBC script library (https://www.bbc.co.uk/writersroom/scripts, drama category), containing scripts of the BBC TV shows. Forty-two English speakers recruited from Prolific (https://www.prolific.co/) were paid £6 per hour to rate the chosen passages on gesturability (on a Likert scale from 0 to 5; defined in the experiment as how easily gestures could be made when uttering the sentence) as well as whether the sentence was meaningful and grammatically acceptable (with “Yes” or “No”) in an online task developed using Gorilla (https://gorilla.sc/). 83 passages were included, all had 1) the mean gesturability score above 2; 2) more than 70% of the participants indicated it was grammatical; and 3) more than 70% of the participants indicated it was meaningful. Four passages were used for practice, and 79 were used as stimuli (Mean gesturability=2.89, SD=0.47).

The same native British English-speaking actress produced the passages with similar instructions with Exp.1. Again, one with gesture version and one without gesture version of the passages were recorded. The averaged number of words in the included passages is 45, and the mean duration of the resulting videos is 15.66s.

4.3.1.3 Quantification of cues

Similar with Exp.1, the onset and offset of each word were automatically detected and was checked manually (word duration: mean=508ms, SD=306ms). For each content word (i.e., nouns, adjectives, verbs and adverbs) we quantified the informativeness of each cue. Linguistic predictability per word was measured using surprisal (Exp.2: Mean=8.17, SD=1.92), calculated using the same n-gram model based on all previous content words in a passage. Pitch prosody per word was quantified as mean F0 (Exp.2: Mean=288Hz, SD=88Hz) extracted using Praat. Gestures were coded as meaningful gestures or beats by three expert coders (inter-rater agreement A&C=95.6, kappa=0.929, p<.001, inter-rater agreement B&C=96.7, kappa=0.948, p<.001). 458 words were associated with meaningful gestures (i.e. iconic gestures and deictic gestures), while 340 words were associated with beat gestures. Mouth informativeness (Exp.2: mean=0.67, SD=0.29) was extracted per word from the corpus described in Chapter 3.

4.3.1.4 Procedure

Participants watched 79 videos, counterbalanced for the presence of gestures (i.e. every two participants shared the same passage sequence containing 39/40 videos with gestures and 40/39 without, with the order randomised. If a passage had gestures for participant A, then participant B watched this passage in the without gesture condition), while their EEG responses were recorded. Videos were separated by a 1000ms interval in Exp.2. Participants were given the same

instruction as in Exp.1. They were asked to watch videos carefully and answer the comprehension questions by pressing the left ("Yes") or right ("No") control key. 40 questions were presented after the 79 passages, accounting for 50% of the total number (20 are Yes and 20 are No). The recording took ~60 mins in Exp.2.

4.3.1.5 EEG recording and preprocessing

Exp.2 used the same EEG setup and preprocessing pipeline as in Exp.1. The artefact rejection process rejected 12.18% of trials (SD=14.43) in Exp.2.

4.3.1.6 Linear mixed effect regression analysis

In order to investigate whether the patterns in Exp.1 is replicable, we performed the same LMER analysis in Exp.2 as Exp.1. The averaged ERP in 300-600ms in Exp.2 was used as dependent variable, and the independent variables included 1) predictors: surprisal, pitch prosody, meaningful gestures, beat gestures, mouth informativeness, and all up to three-way interactions between surprisal and cues, excluding any meaningful*beat gestures interactions (instances where the two gestures co-occur were removed), 2) control: baseline, word length, word order in the sentence, sentence order in experiment, and relative electrode positions measured by X, Y and Z coordinate. We again included word lemma and participant as random variables, with the highest interaction (three-way interactions) as random slope for participants. Due to convergence issues, we did not include surprisal as the random slope (as was done in Exp.1). No predictors showed multicollinearity (Exp.2 VIF<2.5, kappa=5.76). Similar with in Exp.1, we excluded from the analyses: (a) words without a surprisal value (Exp.2: N=13); (b) words without a pitch prosody score (Exp.2: N=2); (c) words associated with both beat and meaningful gestures (Exp.2: N=6); (d) words occurring without any gesture in the "with gesture" condition, and the corresponding words in without gesture videos (Exp.2: N=685, to avoid data

unbalance). Analysis of Exp.2 included 20 participants, 510 word type lemmas and 434,944 data points.

4.3.2 Results: Multimodal cues reliably modulate language comprehension

4.3.2.1 Robustness of the impact of individual cues

In order to assess the replicability of the effect of each cue, we again first focus on the main effects of the multimodal cues and their interaction with surprisal as predictors of N400 amplitude. As we intend to identify the replicable effects only, below we only report in text the effects that are found significant in Exp.1 (See table 3 for full results). Any effect of multimodal cues only significant in Exp.2 is regarded as unreliable and will not be discussed further (this includes surprisal * mouth informativeness, pitch prosody * beat gestures, surprisal * F0 * mouth informativeness).

Table 3.

Full result: linear mixed effects regression model on N400 (300-600ms) in Experiment 2.

Fixed Effects	**β**	**SE**	**t**	**p**
(Intercept)	*0.003*	*0.009*	*0.327*	*0.745*
Predictor Variables				
Surprisal	***-0.067***	***0.004***	***-18.094***	***< .001****
Mean F0	***0.014***	***0.002***	***7.625***	***< .001****
Mouth Informativeness	*0.01*	*0.005*	*1.907*	*0.057*
Meaningful Gesture (Present)	***0.007***	***0.001***	***5.853***	***< .001****
Beat Gesture (Present)	***-0.006***	***0.001***	***-4.002***	***< .001****
Surprisal:Mean F0	***0.012***	***0.002***	***5.948***	***< .001****
Surprisal:Mouth Informativeness	***-0.013***	***0.004***	***-3.312***	***0.001****
Surprisal:Meaningful Gesture (Present)	***0.011***	***0.001***	***8.601***	***< .001****
Surprisal:Beat Gesture (Present)	***-0.01***	***0.001***	***-7.125***	***< .001****
Mean F0:Mouth Informativeness	*0*	*0.001*	*0.235*	*0.814*
Mean F0:Meaningful Gesture (Present)	***0.005***	***0.001***	***4.57***	***< .001****
Mean F0:Beat Gesture (Present)	***0.009***	***0.002***	***5.689***	***< .001****

Mouth Informativeness:Meaningful Gesture (Present)	***0.007***	***0.001***	***6.21***	***< .001****
Mouth Informativeness:Beat Gesture (Present)	***0.004***	***0.001***	***3.181***	***0.001****
Surprisal:Mean F0:Mouth Informativeness	***-0.008***	***0.004***	***-2.151***	***0.042****
Surprisal:Mean F0:Meaningful Gesture (Present)	*-0.004*	*0.003*	*-1.023*	*0.318*
Surprisal:Mean F0:Beat Gesture (Present)	*0.007*	*0.005*	*1.377*	*0.182*
Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	*-0.005*	*0.004*	*-1.19*	*0.247*
Surprisal:Mouth Informativeness:Beat Gesture (Present)	*0.003*	*0.004*	*0.726*	*0.475*
Control Variables				
Word Order	*0*	*0.002*	*-0.241*	*0.809*
Word Length	*-0.004*	*0.003*	*-1.456*	*0.145*
Sentence Order	*0.001*	*0.001*	*1.478*	*0.139*
Baseline	*0.803*	*0.001*	*884.984*	*< .001****
Electrode X	*-0.007*	*0.001*	*-7.537*	*< .001****
Electrode Y	*0.01*	*0.001*	*10.674*	*< .001****
Electrode Z	*-0.005*	*0.001*	*-5.909*	*< .001****

Random Effects		**Variance**	**Std.Dev.**
Lemma	*(Intercept)*	*0.014*	*0.118*
	Surprisal	*-*	*-*
Participant ID	*(Intercept)*	*0.001*	*0.031*
	Surprisal:Mean F0:Mouth Informativeness	*0*	*0.016*
	Surprisal:Mean F0:Meaningful Gesture (Present)	*0*	*0.015*
	Surprisal:Mean F0:Beat Gesture (Present)	*0*	*0.021*
	Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	*0*	*0.017*
	Surprisal:Mouth Informativeness:Beat Gesture (Present)	*0*	*0.018*

Model: Experiment 2				
AIC	*BIC*	*logLik*	*deviance*	*df.resid*
776081	*776630*	*-387990*	*775981*	*434894*

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

We replicated the positive main effect of pitch prosody (Panel A, Figure 6, Exp2: β=0.014, p<.001) and its interaction with surprisal (Exp2: β=0.012, p<.001), indicating that words with higher pitch prosody showed less negative N400, especially for higher surprisal words. Similarly, the main effect of meaningful gestures (Panel B, Figure 6, Exp2: β=0.007, p<.001) its interaction with surprisal (Exp2: β=0.011, p<.001) are also

replicated, indicating that words with meaningful gestures showed less negative N400 in general, and higher surprisal words elicited a larger reduction of N400 amplitude when meaningful gestures were present, in comparison to lower surprisal words. We also replicated the opposite effect of beat gestures (Panel C, Figure 6, Exp2: β=-0.006, p=.001) and the interaction between beat gestures and surprisal (Exp2: β=-0.010, p<.001): words accompanied by beat gestures elicited a more negative N400, especially for higher surprisal words.

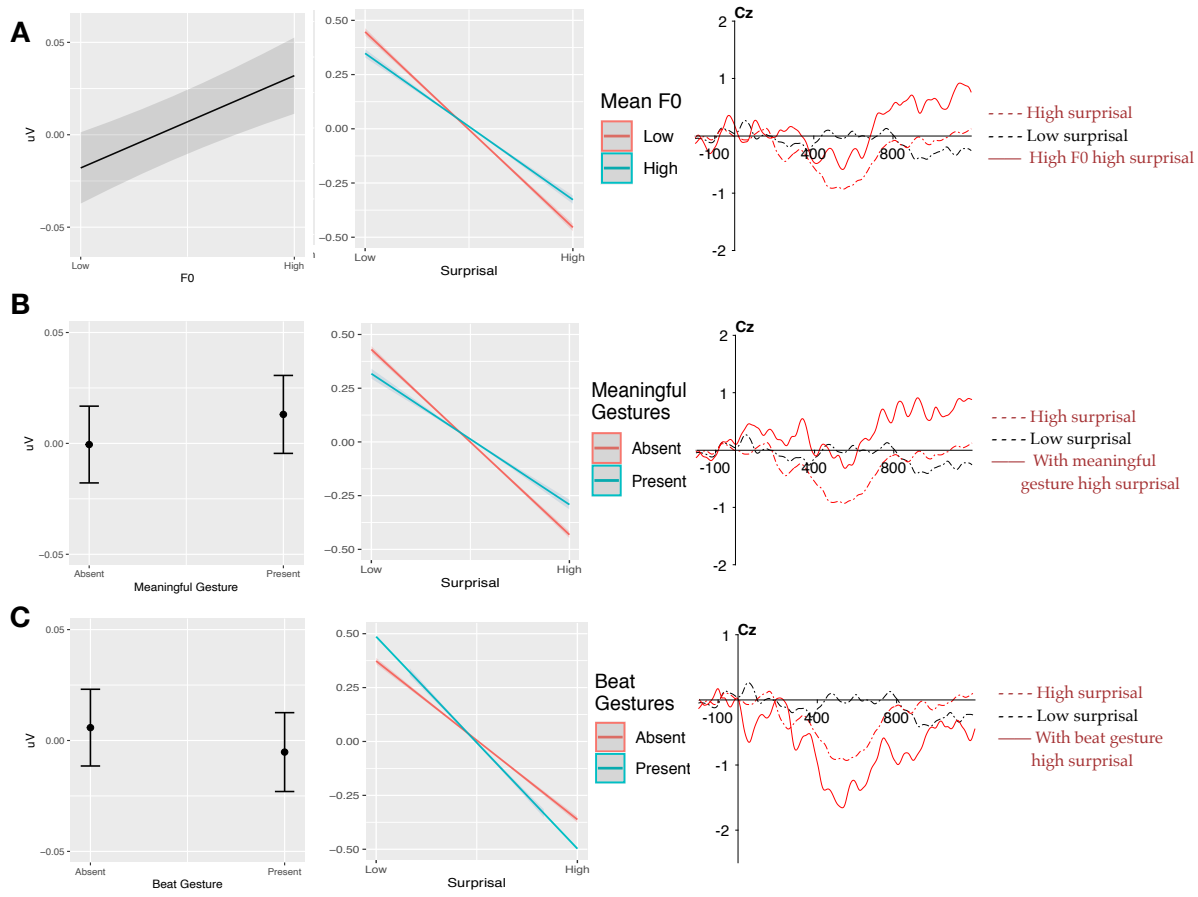


Figure 6.

Multimodal cues each modulate N400 and the impact of surprisal. Replicating the results from Exp.1, A). Higher F0 induce less negative N400, especially for high surprisal words. B). Meaningful gestures induce less negative N400, especially for

high surprisal words. C) Beat gestures induce more negative N400, especially for high surprisal words.

4.3.2.2 Robustness of interaction between cues

We replicated the majority of the significant interactions between multimodal cues (Figure 7), including the positive interaction between prosody and meaningful gestures (Panel A, Exp2: β=0.005, p<.001); between mouth informativeness and meaningful gestures (Panel B, Exp2: β=0.007, p<.001) and between mouth informativeness and beat gestures (Panel C, Exp2: β=0.004, p=0.001), indicating that the co-occurrence of these cues induced even larger N400 reduction. However, we failed to replicate the three way interaction between mouth informativeness, meaningful gestures and surprisal.

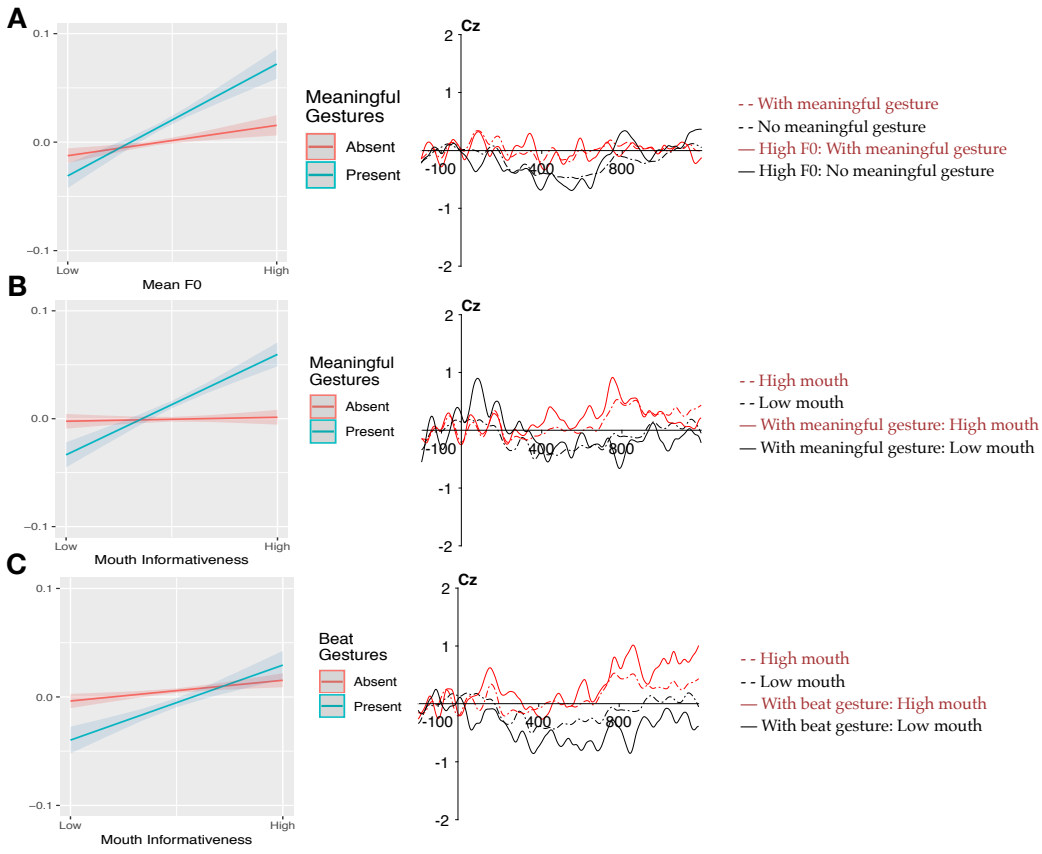


Figure 7.

Multimodal cues interact to affect N400. Replicating the results from Exp.1, A) The positive effect of meaningful gestures is larger for higher F0 words; B) The positive effect of mouth movements is larger for words with meaningful gestures; C) The positive effect of mouth movements is larger for words with beat gestures.

4.3.3 Discussion

Experiment 2 investigated whether multimodal cues jointly modulate N400 responses with more naturalistic materials, and replicated most of the effects reported in Experiment 1. First, multimodal cues individually modulate N400 and the effect of surprisal. While higher prosodic pitch and meaningful gestures induce less negative N400, especially for less predictable words, beat gestures induce the opposite effect. Second, the co-occurrence between higher pitch and meaningful gestures, as well as higher mouth informativeness with both gesture types, elicit even less negative N400. We did not replicate the three way interaction between mouth informativeness, meaningful gestures and surprisal. Overall, these results suggested that multimodal cues reliably modulate the processing of linguistic predictability, both individually and interactively.

4.4 General discussion

We investigated for the first time the electrophysiological signature of naturalistic multimodal language comprehension, containing naturally occurring gestures, prosody and mouth movements. We found that linguistic predictability, indexed by surprisal values, is associated with N400 amplitude per word. However, more crucially, this effect of linguistic predictability is also modulated by multimodal cues and their interactions.

Our first question is whether multimodal cues always modulate language comprehension in the multimodal context. This question is addressed by measuring whether the effect of linguistic predictability, indexed by N400, is modulated by other cues. We found that prosodic accentuation (marked by higher pitch prosody) and meaningful gestures reduced N400 overall, especially for less predictable words. Whereas beat gestures showed the opposite effect, inducing more negative N400, especially for less predictable words. Mouth movement did not modulate the surprisal effect in itself but instead participated in the interaction with gestures to modulate N400. Thus, our results suggest that word predictability calculated only on linguistic context can over-estimate (when not accounting for the effects of prosody and meaningful gestures) or underestimate (when not accounting for beat gestures) the cognitive load during comprehension, indexed by N400 amplitude. While this does not mean linguistic predictability no longer have an effect in multimodal communication (due to the existance of the robust surprisal effect on N400), this pattern supports that language comprehension in the face-to-face environment involves linguistic and other multimodal information which cannot be neglected either.

Our second question is, what are the dynamics of multimodal language comprehension. We address this question by investigating whether and how multimodal cues interact with each other. We found that higher pitch prosody enhances the facilitatory effect of meaningful gestures and that the co-occurrence of more informative mouth movements and any gestures (meaningful or beat) leads to an even larger N400 reduction. Therefore, the processing of each multimodal cue at any time is dependent on the presence of other cues. Thus, investigating individual

multimodal cues while excluding other cues does not provide the complete picture of multimodal language processing.

Our results confirmed and extended previous studies investigating the impact of each multimodal cue. Prosodic accentuations are considered a marker of new information (Cruttenden, 2006), as speakers are more likely to produce accentuation when words convey new information. The incongruent pairing of newness and prosodic accentuation elicit more negative N400, indicating enhanced processing difficulty (e.g. Magne et al., 2005). Our findings complement previous studies by showing that naturally occurring pitch prosody also favours the processing for less predictable words, indexed by less negative N400 amplitude. We also found that meaningful gestures facilitate the processing of words, especially if the linguistic predictability is low. 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 (Holle and Gunter, 2007; Obermeier et al., 2011; Obermeier et al., 2012), and previous work that showed that words produced with incongruent gestures induce larger N400 (see review in Özyürek, 2014). More crucially, our results also suggest that meaningful gestures play a more general role in face-to-face communication, as they always support word processing, not just in the case of incongruence or ambiguity.

However, beat gestures did not show the same effect. Instead, beat gestures elicited an even larger N400 effect, especially for high surprisal words. This effect may be due to beat gestures enhancing the saliency of a specific word (Krahmer and Swerts, 2007), thus highlighting its low predictability in the context. Alternatively, listeners might extract meaning from all gestures and integrate it with the speech by default. Since beats are not meaningful, integration fails, inducing processing

difficulties. Previous studies failed to find the same N400 effects of beat gestures (Wang & Chu, 2013; Morett et al., 2021). One possible reason for the discrepancy is that these studies manipulated the presence of beat gestures to have one single stroke per sentence. This is different from the naturally occurring, more continuous beat gestures, which were shown to have different effects (Rohrer et al., 2020). Alternatively, as these studies only presented beat gestures but not any meaningful hand movements, participants may not pay enough attention to the hand gestures. Shifts in the weight attributed to different cues based on specific tasks are documented in the literature (e.g., Gunter and Weinbrenner, 2017; Holle and Gunter, 2007; Obermeier et al., 2015), highlighting the importance of using ecologically valid paradigms.

We did not find any reliable effect of mouth movements in itself, either as a main effect modulating N400 or as an interaction with surprisal (thus modulating the effect of linguistic predictability). Most previous studies focused on mouth movements and language perception and found that the presence of mouth reduced early N1-P2 amplitude, indicating that mouth movements modulate the early processing of words. Only two studies so far investigated the impact of mouth in the N400 window and reported different results. While Brunellière and colleagues (2013) reported that informative mouth movements induced a more negative N400, Hernández-Gutiérrez and colleagues (2018) did not find any impact of mouth movements on the N400. Further research is needed to clarify the discrepancies. However, our results suggested that mouth movements do not modulate N400 in itself but may reduce the N400 amplitude with other cues (i.e. gestures) when presented in a multimodal context.

Finally, our results extended previous literature by showing that multimodal cues interact with each other in more naturalistic material. We reported a robust interaction between meaningful gestures and prosody for the first time: higher pitch prosody enlarges the facilitatory effect of meaningful gestures. fMRI studies suggested that prosodic accentuation may activate the more domain-general attention network (Kristensen et al., 2013). Therefore, the higher pitch may heighten the attention to meaningful gestures that co-occur and therefore enlarge its effect. Alternatively (or additionally), as Holler & Levinson (2020) argued, multimodal cues are automatically bundled together based on their natural correlation. As meaningful gestures usually co-occur with higher pitch (Brentari et al., 2013; Esteve-Gibert & Prieto, 2013), their combination may be easier to process, thus offering even larger processing benefit indexed by N400 reduction. Moreover, we found that the effect of mouth movements is enhanced by the presence of any gestures (both meaningful and beat). While one may expect the effect of mouth movements to be smaller when gestures are present, as both cues occupy the visual channel thus may compete for attentional resources, we found replicable evidence that mouth movements show a larger facilitatory effect when gestures are present. While comprehenders tend to gaze at the eye-areas in general (the “eye-primacy effect”), some studies found that listeners focus on chin areas while processing hand movements (e.g. Beattie, Webster & Ross, 2010). Therefore, it is possible that mouth movements fall within the focus of visual attention more easily when attention is also drawn to gestures. Future study may further investigate this possibility by adding eye-tracking measures to capture visual attention.

4.4.1 Toward a neurobiological model of natural language use

In probabilistic-based predictive accounts, the N400 is taken as an index of the processing demands associated with low predictability (e.g. Kuperberg and Jager, 2015). 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 (Levy, 2008). Thus, the N400 is linked to the process of updating the distribution of hypotheses: smaller N400 is associated to more accurate prior distributions/predictions (Kuperberg and Jager, 2015). Our work shows that these mechanisms do not operate only on linguistic information but crucially, they weight 'non-linguistic' multimodal cues. Higher pitch prosody may prepare comprehenders for lower 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 & Levinson, 2019).

Conventional dual-stream neurobiological models do not typically concern themselves with face-to-face language and are mostly localised to perisylvian and inferior frontal regions (e.g., Hickok and Poeppel, 2007). As such, they cannot easily accommodate the results here, and the naturalistic combination of multiple cues during speech comprehension may engage a wider set of brain regions. A better fit are those models in which language comprehension is considered in context and associated with many interconnected networks distributed throughout the whole brain (Hasson et al., 2018; Skipper, 2015). 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 (Skipper, 2015). 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 (Skipper, 2007, 2009). 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 (e.g., Skipper, van Wassenhove, Nusbaum, and Small, 2007). These models predict the involvement of a wider range of networks, such as motor areas (which has been found to reflect processing of mouth movements, e.g. Skipper et al., 2007); or parietal areas (which has been shown to be sensitive to prosodic processing, e.g. Kristenson et al., 2013). Our finding of multiple interactions between cues is more compatible with this view, suggesting that multimodal prediction processes are dynamic, re-weighting each cue based on the status of other cues.

To conclude, our study assessed language processing in the naturalistic multimodal environment for the first time and provided evidence that multimodal cues constantly and dynamically interact to construct predictions. Thus, our study provides a new, more ecologically valid way to understand the neurobiology of language, in which multimodal cues are dynamically orchestrated. In the next chapter, we will apply the same paradigm to investigate how non-native speakers process multimodal materials.

Reference

1. 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).

2. Arbib, M. A., Liebal, K. & Pika, S. Primate Vocalization, Gesture, and the Evolution of Human Language. *Curr. Anthropol.* **49**, 1053–1076 (2008).

3. 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).

4. Arnold, P. & Hill, F. Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *Br. J. Psychol.* **92**, 339–355 (2001).

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

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

7. Beattie, G., Webster, K. & Ross, J. The Fixation and Processing of the Iconic Gestures That Accompany Talk. *J. Lang. Soc. Psychol.* **29**, 194–213 (2010).

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

9. BNC Consortium. The British national corpus, version 3 (BNC XML Edition). *Distrib. Oxf. Univ. Comput. Serv. Behalf BNC Consort.* **5**, 6 (2007).

10. Bock, J. K. & Mazzella, J. R. Intonational marking of given and new information: Some consequences for comprehension. *Mem. Cognit.* **11**, 64–76 (1983).

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

12. 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).

13. Buchwald, A. B., Winters, S. J. & Pisoni, D. B. Visual speech primes open-set recognition of spoken words. *Lang. Cogn. Process.* **24**, 580–610 (2009).

14. Calvert, G. A. *et al.* Activation of Auditory Cortex During Silent Lipreading. *Science* **276**, 593–596 (1997).

15. 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).

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

17. 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).

18. 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).

19. 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).

20. Fernald, A. *et al.* A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants*. *J. Child Lang.* **16**, 477–501 (1989).

21. Fort, M. *et al.* Seeing the initial articulatory gestures of a word triggers lexical access. *Lang. Cogn. Process.* **28**, 1207–1223 (2013).

22. 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).

23. Friston, K. & Kiebel, S. Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. B Biol. Sci.* **364**, 1211–1221 (2009).

24. 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).

25. Hagoort, P. & Brown, C. M. ERP effects of listening to speech: semantic ERP effects. 13 (2000).

26. 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).

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

28. 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).

29. Hobaiter, C., Byrne, R. W. & Zuberbühler, K. Wild chimpanzees' use of single and combined vocal and gestural signals. *Behav. Ecol. Sociobiol.* **71**, 96 (2017).

30. Holle, H. & Gunter, T. C. The Role of Iconic Gestures in Speech Disambiguation: ERP Evidence. *J. Cogn. Neurosci.* **19**, 1175–1192 (2007).

31. 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).

32. Holler, J. & Levinson, S. C. Multimodal Language Processing in Human Communication. *Trends Cogn. Sci.* **23**, 639–652 (2019).

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

34. 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).

35. Iverson, J. M. & Goldin-Meadow, S. Gesture Paves the Way for Language Development. *Psychol. Sci.* **16**, 367–371 (2005).

36. Jouravlev, O. *et al.* Speech-accompanying gestures are not processed by the language-processing mechanisms. *Neuropsychologia* **132**, 107132 (2019).

37. Kelly, S. D., Kravitz, C. & Hopkins, M. Neural correlates of bimodal speech and gesture comprehension. *Brain Lang.* **89**, 253–260 (2004).

38. 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).

39. 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).

40. Kuperberg, G. R. & Jaeger, T. F. What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* **31**, 32–59 (2016).

41. Kushch, O., Igualada, A. & Prieto, P. Prominence in speech and gesture favour second language novel word learning. *Lang. Cogn. Neurosci.* **33**, 992–1004 (2018).

42. 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).

43. Levinson, S. C. & Holler, J. The origin of human multi-modal communication. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130302 (2014).

44. Lewkowicz, D. J. & Hansen-Tift, A. M. Infants deploy selective attention to the mouth of a talking face when learning speech. *Proc. Natl. Acad. Sci.* **109**, 1431–1436 (2012).

45. Loehr, D. Aspects of rhythm in gesture and speech. *Gesture* **7**, 179–214 (2007).

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

47. MacCallum, R. C., Zhang, S., Preacher, K. J. & Rucker, D. D. On the practice of dichotomization of quantitative variables. *Psychol. Methods* **7**, 19 (2002).

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

49. Massaro, D. W. & Jesse, A. Audiovisual speech perception and word recognition. *Oxf. Handb. Psycholinguist.* (2007) doi:10.1093/oxfordhb/9780198568971.013.0002.

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

51. Morett, L. M., Landi, N., Irwin, J. & McPartland, J. C. N400 Amplitude, Latency, and Variability Reflect Temporal Integration of Beat Gesture and Pitch Accent During Language Processing. *Brain Res.* 147059 (2020) doi:10.1016/j.brainres.2020.147059.

52. Mortensen, D. R. *et al.* PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. in *COLING* (2016).

53. 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).

54. 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).

55. 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).

56. 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).

57. Pilling, M. Auditory Event-Related Potentials (ERPs) in Audiovisual Speech Perception. *J. Speech Lang. Hear. Res.* **52**, 1073–1081 (2009).

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

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

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

61. Seyfarth, R. M. & Cheney, D. L. Production, usage, and comprehension in animal vocalizations. *Brain Lang.* **115**, 92–100 (2010).

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

63. 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).

64. 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.

65. 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).

66. 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).

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

68. Smith, N. J. & Levy, R. The effect of word predictability on reading time is logarithmic. *Cognition* **128**, 302–319 (2013).

69. Sumby, W. H. & Pollack, I. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* **26**, 212–215 (1954).

70. Ter Bekke, M., Drijvers, L. & Holler, J. The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech. (2020) doi:10.31234/osf.io/b5zq7.

71. 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).

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

73. Vigliocco, G., Krason, A., Stoll, H., Monti, A. & Buxbaum, L. J. Multimodal comprehension in left hemisphere stroke patients. *Cortex* **133**, 309–327 (2020).

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

75. Wilke, C. *et al.* Production of and responses to unimodal and multimodal signals in wild chimpanzees, Pan troglodytes schweinfurthii. *Anim. Behav.* **123**, 305–316 (2017).

76. 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).

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

78. Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J. I., & Vigliocco, G. (2021). More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension. *Proceedings of the Royal Society B*, *288*(1955), 20210500.

Chapter 5

5 Multimodal language comprehension in L2

5.1 Introduction

Almost 1 billion people around the world speak English as a second language (L2; Ethnologue 24th edition, 2021). Despite the large number of L2 users, to what extent they benefit from multimodal information such as prosody, gestures and mouth movements in daily face-to-face conversations remains less understood. It is possible that the multimodal nature of naturalistic speech adds to the complexity of comprehension, as listeners have to process information from different sources simultaneously. This challenge may be especially pronounced for L2 users, as comprehension in a non-native language can be cognitively taxing (e.g. Hopp, 2010). On the other hand, the extra non-linguistic multimodal information may make up for the lower ability to comprehend linguistic information in L2 (e.g. listeners may access semantic information from meaningful gestures even if they failed to catch on the speech itself). In this chapter, we use the same experimental design as in Chapter 4 to investigate how L2 listeners comprehend naturalistic style speech where multimodal cues co-occur. This study has been published as Zhang, Ding, Frassinelli, Tuomainen, Klavinskis-whiting & Vigliocco (2021).

A handful of studies investigated how L2 listeners process each multimodal cues, and these studies indicate that L2 listeners are typically less likely to benefit from these cues compared to L1 listeners. For example, it has been found in L1 that

prosodic accentuation enhances listeners' attention to stressed information, possibly via activation of the attentional network (Kristensen et al., 2012). Therefore, accentuated information is processed faster (e.g. Cutler & Foss, 1977), and any mismatching information is highlighted, indexed by larger N400 associated with incongruent linguistic information (e.g. Li, Hagoort & Yang, 2008). However, apart from the attentional functions, prosodic accentuation also marks new and less predictable information, with new information more likely to be accompanied by accentuation (e.g. Cruttenden, 2006). L1 listeners encounter larger processing difficulty when such a pattern is violated, either for new information without accentuation or old information with accentuation, as indexed by more negative N400 (e.g. Magne et al., 2005). It is usually found that N400 amplitude is reduced for the naturally occurring (thus congruent) prosody where less predictable linguistic information is accompanied with prosodic accentuation (see Chapter 4). For L2 listeners, however, although accentuated information in is also easier to process in general just like in L1 listeners (Akker & Cutler, 2003; Takahashi et al., 2018), some evidence suggests that the (in)correct mapping between the prosodic and the linguistic information only showed smaller (if any) effect. For example, Akker and Cutler (2003) found that both L1 and L2 listeners showed faster phoneme detection when prosodic accentuation was present, supporting that accentuation enhances attention to target information (as mentioned above). However, while words in focused position (induced by a preceding question) showed smaller effect of prosody in L1, such interaction was absent in L2. This suggest that the impact of prosodic accentuation is modulated by semantic information (namely semantic focus) in L1 listeners but not L2. The authors argued that L2 listeners cannot rapidly incorporate prosodic information with semantic structure, possibly due to insufficient cognitive

resource as L2 comprehension is more computationally taxing. Along similar lines, an eye-tracking study found that while both L1 and L2 participants tend to look at the corresponding new/old object based on the presence of prosodic accentuation, only L1 participants predictively restricted their attention to the upcoming referent (i.e. prior to the utterance of the objects' label, Perdomo & Kaan, 2019), possibly indicating that L1 users are more efficient in linking prosodic pattern with newness of an object. Lee, Perdomo and Kaan (2019) further found that while L1 participants showed larger N400 amplitude when prosodic accentuation is inappropriately paired with the linguistic context, L2 participants did not show the same effect (Lee et al., 2019), indicating that the link between prosodic and linguistic information in L2 is weaker. These studies suggest that while prosodic accentuation may modulate L2 listeners' attention to specific linguistic information in general, L2 listeners are less capable of mapping the prosodic pattern with linguistic information.

Compared with prosody, studies comparing L1 and L2 processing of meaningful gestures have produced mixed results. Meaningful gestures (i.e. iconic gestures or pointing gestures) directly convey properties of the referent and therefore offers a processing advantage in general (see meta-analysis in Hostetter, 2011; Dargue, Sweller, & Jones, 2019). However, when recognising single word in noise produced either with or without iconic gestures, L2 comprehenders have been found to benefit less than L1 comprehenders from these gestures (Drijvers, Vaitonytė, et al., 2019; Drijvers & Özyürek, 2019). An MEG study using the same task further found that while both L1 and L2 participants showed alpha/beta power suppression when gestures were present, indexing facilitatory effect of gestures, L2 comprehenders showed smaller power suppression in inferior and anterior temporal areas (Drijvers, van der Plas, Ozyurek & Jensen, 2019). As these areas are

associated with the access and integration of semantic information (e.g. Ralph, Jefferies, Patterson & Rogers, 2017), the authors argued that L2 listeners might experience more problems unifying linguistic and gestural information, which may explain the smaller behavioural improvements associated with gestures. However, when measuring the impact of meaningful gestures on comprehension (rather than recognition), studies found that L2 comprehenders might benefit more than L1 comprehenders. Electrophysiological studies found that incongruent meaningful gestures elicit more negative N400, indicating enhanced difficulty in semantic processing, in both L1 (e.g. Özyürek et al., 2007) and L2 comprehenders (Ibáñez et al., 2010), but this N400 effect is even larger in L2 than L1 participants (Drijvers & Özyürek, 2018). This indicates that L2 listeners are more sensitive to the semantic mismatch between gestures and speech. Along similar line, when comprehending longer stories, L2 but not L1 participants showed more accurate recall when gestures were present (Dahl & Ludvigsen, 2014). This advantage has been found especially for more complex speech (Lin, 2021) and especially for less proficient L2 comprehenders (Sueyoshi & Hardison, 2005). These findings are compatible with the theory that meaningful gestures scaffold language comprehension by providing extra semantic information in the visual domain, which is especially helpful when the listeners are less proficient and/or when the speech is more challenging. Alternatively or additionally, meaningful gestures may potentially even bypass linguistic processing by providing direct semantic information. This effect may only be helpful for L2 listeners in comprehension tasks (versus recognition), possibly because the semantic information from gestures can be directly incorporated into comprehension, but has to be first transformed into phonological information to be used in the recognition processes.

We know a lot less about whether and how beat gestures impact L2 comprehension. It has been found in L1 that words accompanied with beat gestures are more salient (e.g. Krahmer & Swerts, 2007), and therefore information presented with beat gestures may be learnt better (e.g. Kushch & Prieto, 2016, but the results are mixed, see e.g. Feyereisen, 2006; Macoun & Sweller, 2016). Similar facilitatory effect of beat gestures on word learning has also reported in L2 (Pi et al., 2021, but see Lin, 2021, who reported no effect of beat gestures), and the effect is larger when prosodic accentuation is also present (Kushch et al., 2018), as in natural communication. However, the effect of beat gestures may differ based on how the gestures are performed. Rohrer and colleagues (2020) found that more continuous beat gestures produce no effect in L1 participants and even worse memory performance (than without gestures) in L2 comprehenders (Rohrer et al., 2020), in contrast to beat gestures performed with a single stroke (e.g. Kushch et al., 2018). The authors attributed this difference to the fact that the more continuous beat gestures are less visually salient, making it more difficult to integrate them with speech. This difficulty may be especially pronounced in L2 comprehenders, therefore inducing worse memory than when no gestures are used.

Finally, mouth movements have long been known to facilitate word recognition of L1 words (e.g. Sumby and Pollack, 1954) and this is also true for L2 words (Navarra & Soto-Faraco, 2007). Some studies found in single word recognition tasks that visible mouth movements induce smaller improvements in L2 comprehenders than L1 (Drijvers, Vaitonytė, et al., 2019; Drijvers & Özyürek, 2019). However, one study reported that, when listening to longer connected speech, L2 listeners are more likely to look at the mouth area of the speaker (relative to the eye areas) compared with the L1 comprehenders (Birulés et al., 2020). Therefore, it is

possible that the impact of mouth movements is again task dependent: while L2 users may be less efficient in using mouth movements for single words, it is also possible that they search for sensory information more actively to aid their comprehension in longer or more complex materials.

Finally, only a handful of studies investigated how multiple cues jointly affect L2 processing. Drijvers and Özyürek (2019) presented participants with single words embeded in noise and accompanied by meaningful gestures, visible mouth movements or both cues. The task was to type down the word the actress said. They found that when the auditory input was heavily degraded, L1 but not L2 users benefited from the combination of visible mouth movement and meaningful gestures on top of one single cue, indicating that L2 users experience more problems combining mouth and gestural information when the recognition is highly challenging. In an eye-tracking study using a similar design, Drijvers, Vaitonytė and Özyürek (2019) further found that although both L1 and L2 users gazed more at the face area in general, L2 users gazed more often at hand gestures than L1 comprehenders. Interestingly, only L1 but not L2 users' gazes to the gestures predicted their actual benefit from gestures at a later recall task. Therefore, L2 users might pay more attention to gestures, but are less efficient in using them to recognize words embedded in noise. To note, although these studies presented both mouth movements and meaningful gestures, the goals were to establish the double enhancement (i.e. whether the presence of two cues induce larger facilitatory effect than one cue) or to investigate the relative importance of cues. Therefore, they cannot elucidate whether the two cues interacts (i.e. whether the presence of one cue enhances/decreases the effect of another, which can happen on top of the double enhancement). Finally, one study investigated the interaction between

prosodic accentuation and beat gestures. Kushch, Igualada and Prieto (2018) asked participants to learn new L2 words with prosodic accentuation and/or beat gestures, and found that participants were more likely to remember words that were accompanied by both cues. Conversely, when words were accompanied by beat gestures but not prosodic accentuation, participants' performance was worse than when there was only prosodic accentuation, indicating that only the naturally occurring beat gesture (that typically co-occur with prosodic accentuation) facilitate learning of L2 words.

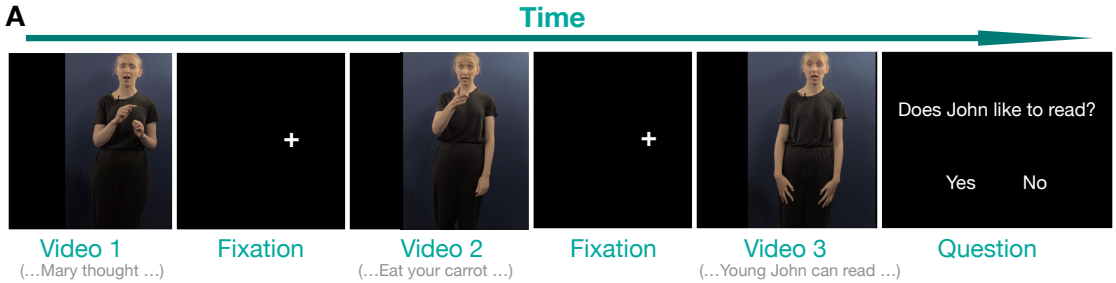
In summary, previous studies suggest that L2 comprehension is modulated by each multimodal cue (e.g. Akker and Cutler, 2003; Dahl & Ludvigsen, 2014; Pi et al., 2021; Navarra & Soto-Faraco, 2007), with potential interactions between cues (e.g. prosodic accentuation and beat gestures, Kushch, Igualada & Prieto, 2018). While the majority of these studies found that L2 listeners benefit less from multimodal cues than L1, the pattern may differ across cues and tasks. For example, while studies on prosodic accentuation and beat gestures uniformly reported smaller facilitatory effect in L2 than L1, some studies reported L2 users paying more attention to mouth movements (Birulés et al., 2020) or meaningful gestures (Drijvers & Özyürek, 2018) and actually benefiting more from them (Dahl and Ludvigsen, 2014). As mouth movements and meaningful gestures directly convey sensory and semantic information, in contrast to beats and prosodic accentuation that have a role in drawing attention to specific parts of the speech, it is possible that L2 listeners pay more attention to mouth movements and meaningful gestures, and are more likely to make use of them to aid linguistic processing. Moreover, studies reporting a larger effect of multimodal cues in L2 listeners typically presented participants with longer connected speech (Birulés et al., 2020, Dahl and Ludvigsen, 2014; Sueyoshi and

Hardison, 2005) rather than single words. These longer passages may contain more challenging linguistic information as well as richer multimodal cues, which means multimodal cues can be especially helpful.

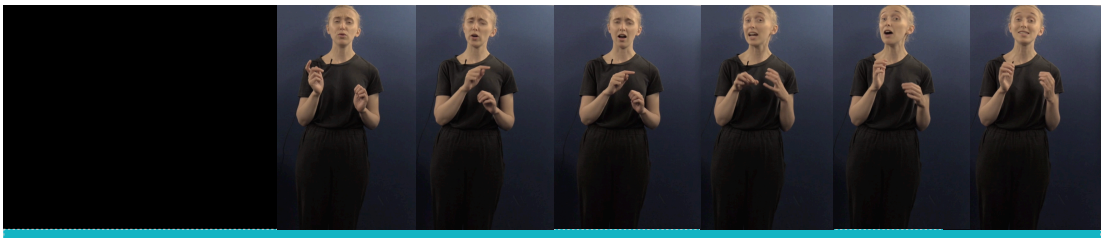
What remains unknown is the electrophysiological pattern of L2 comprehension in a more naturalistic context, where linguistic information and multimodal cues co-occur and interact. Previous studies in L2 predominantly investigated each cue individually in artificial tasks and conditions (e.g. in single word recognition tasks, such as Drijvers & Ozyurek, 2018; or when hiding away other multimodal cues, such as Akker & Cutler, 2003). These experimental manipulations may potentially enhance participants' attention to the cue present, thus enlarging the effect of individual cue being investigated. Therefore, it is unknown whether all cues continue to have an impact in the naturalistic context, with the presence of other cues. Moreover, as previous L2 studies primarily focused on individual cues, interactions between multimodal cues are largely unexplored (except for Kushch et al., 2018), with some interactions never studied (e.g. prosody and meaningful gestures). These interactions are likely to affect L2 comprehension as well, based on the pattern observed in L1 (as was shown in Chapter 4).

Here, we present an electrophysiological study of L2 processing of naturalistic-style audio-visual materials. We used the exact same design and stimuli as in Experiment 2, Chapter 4 (see Figure 1 for an example). An actress produced passages chosen from BBC TV scripts with naturally occurring prosodic change, mouth movements and gestures. Twenty highly proficient non-native English speakers (Mandarin-English) watched these videos while their EEG was recorded. We first identified in L2 listeners the electrophysiological component sensitive to linguistic predictability and its time window. We then quantified multimodal cues per

word, and analysed how these cues individually and jointly modulated the EEG response in the time window in which linguistic predictability had an effect. Finally, we compared the responses between L2 participants with L1 participants (from Experiment 2, Chapter 4) to investigate differences across the groups.



B



	Mary	thought	of	joining	Clara.	They	might	pick	flowers	together.
Surprisal	-	6.73	-	9.02	-	-	-	8.63	9.41	7.40
Pitch 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.

Illustration of experiment design. (A) Participants watched videos of an actress narrating short passages in naturalistic style, and answered comprehension questions following some videos. (B) We quantified informativeness of surprisal, pitch prosody, gestures and mouth movements per each content word.

5.2 Methods

5.2.1 Participants

Twenty (16 females, 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 an 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.

5.2.2 Materials

Materials were taken from Exp.2, Chapter 4. Participants of the EEG study (both in Exp.2 Chapter 4 and the current experiment) rated the difficulty of each passage after the experiment on a 1-5 scale in order to determine whether any stimulus was too difficult. The average difficulty score of the 79 passages was not significantly different across L1 participants (Exp.2, Chapter 4) and 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., which indicated that no stimulus in our study was extremely difficult for all the L2 participants to follow. Therefore, all the 79 passages were included in further analyses.

5.2.3 Procedure

The recording procedure is identical with Exp.2 Chapter 4.

5.2.4 Quantification of cues

Quantification of cues are taken from Exp.2, Chapter 4.

5.2.5 Preprocessing of EEG data

The data was pre-processed according to Exp.2, Chapter 4. The data cleaning process (using window peak-to-peak analysis and step-like artifact analysis) resulted in an average rejection of 8.69% (SD=14.12) of the trials.

5.2.6 Hierarchical linear modelling analysis

Same Hierarchical linear modelling analysis as in Exp.2 Chapter 4 was conducted to identify the time window where linguistic surprisal have an impact on L2 comprehenders.

5.2.7 Linear mixed effect regression analysis

We conducted LMER analyses using lme4 package (Bates, Maechler, Bolker, & Walker, 2015) under RStudio (version 4.0.4). For each participant, mean ERPs (without baseline correction) 500-800 ms time windows (where a significant negativity was detected in LIMO analysis, see 5.3.1 below for the result) were extracted from 32 electrodes for all the content words and were used as a dependent variable. Mean ERPs in the time window of -100 to 0ms were extracted as well as a baseline for use as a control predictor in the LMER models. For all models below, we only include the words with gestures (for videos with gestures) and the corresponding words without gesture (for videos without gestures) to balance the number of observations between groups.

Analysis 1: How do multimodal cues affect L2 processing? Here we analyzed how multimodal cues interact to affect N400 of L2 participants. The independent variables included in the LMER models were: 1) main effects of: surprisal, mean F0, meaningful gesture, beat gesture, 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: extracted baseline (-100 to 0ms), word length, word order in the passage, passage order in the experiment, x, y and z coordinates of electrode. No main or interaction effects showed multicollinearity, with variance inflation factor (VIF) less than 2.4, kappa=5.63. All continuous variables, including ERP, surprisal, mean F0, mouth informativeness, baseline, 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 was 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). 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: Do multimodal cues show the same effects in L1 and L2? Here we compared results from L2 participants to those of L1 participants who were tested with the same materials (Exp.2 Chapter 4). The EEG responses within 500-800ms from the 20 L1 participants reported in Chapter 4 (Exp 2) were combined with the L2 data described above. As 500-800ms did not cover the full N400 window for L1 participants, we also compared 350-800ms for L1 (identified in LIMO analysis on L1 participants) with 500-800ms for L2 participants. As the results are very similar, we only report the results from 500-800ms across both groups below. 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).

5.3 Results

5.3.1 L2 comprehension is sensitive to surprisal

First, we established the precise time window in which linguistic surprisal has an effect in L2 users with hierarchical LInear MOdeling (LIMO toolbox) instead of specifying N400 window a priori, as previous studies never investigated the effect of surprisal in L2 comprehenders in multimodal context. LIMO toolbox performs a

regression-based EEG analysis, which decomposes ERP signal into time-series of beta coefficient waveforms associated with each continuous variable. As shown in Figure 2, EEG responses for words with higher surprisal were significantly more negative in the 500-800ms time window post-stimulus. Therefore, we focused on 500-800ms in all following analyses.

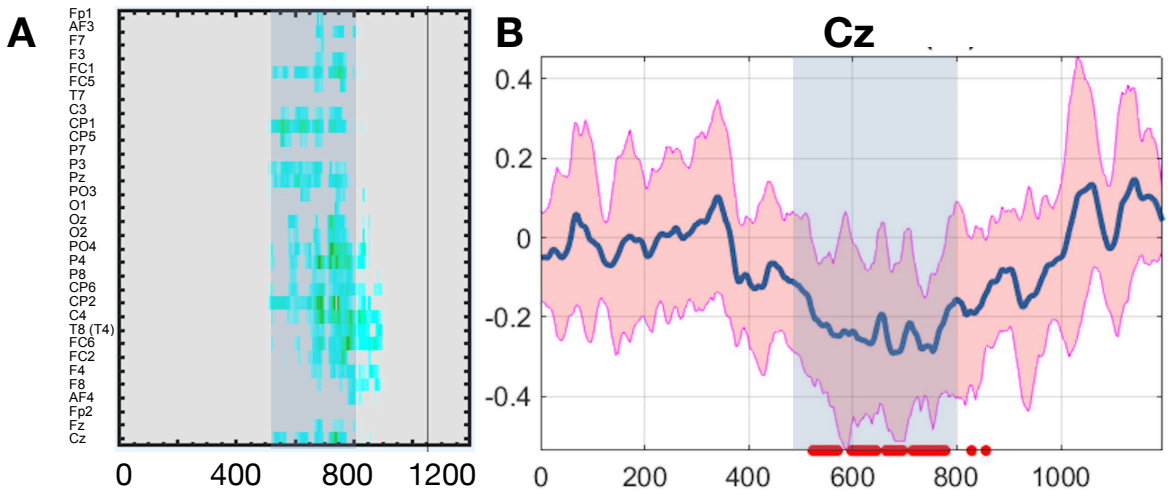


Figure 2.

Hierarchical linear modelling results showing the ERP sensitive to surprisal (one-sample t-test P<0.05, cluster-corrected). (A) Time window (500-800ms) showing increased significant negativity associated with surprisal (in green). Grey areas are not statistically significant. (B) Averaged beta plot for electrode Cz illustrating that beta values for surprisal were significantly negative compared with 0 (flat waveform) in around 500-800ms. 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.

5.3.2 Analysis 1: How do multimodal cues affect L2 processing?

The first LMER analysis investigates whether L2 users also make use of multimodal cues and how these cues interact during comprehension of naturalistic style speech. A positive effect of a cue or an interaction indicates that the cue or

combination induced less negative (smaller amplitude) N400, suggesting easier processing. Conversely, a negative effect or interaction indicates that the cue induce more negative N400 than without the cue or the combination induce smaller N400 reduction than the numerical addition of two effects. See full results in Table 1.

Table 1.

Full results for analysis 1: linear mixed effect regression analysis on N400 (500-800ms) in L2 participants.

Fixed Effects	**β**	**SE**	**t**	**p**
(Intercept)	*0.004*	*0.011*	*0.341*	*0.736*
Predictor Variables				
Surprisal	***-0.008***	***0.002***	***-5.191***	***<.001****
Mean F0	***0.004***	***0.002***	***2.535***	***0.011****
Mouth Informativeness	***0.007***	***0.001***	***4.957***	***<.001****
Meaningful Gesture (Present)	*0.002*	*0.001*	*1.483*	*0.138*
Beat Gesture (Present)	*0.002*	*0.001*	*1.552*	*0.121*
Surprisal:Mean F0	***-0.006***	***0.002***	***-3.526***	***<.001****
Surprisal:Mouth Informativeness	***0.010***	***0.002***	***6.186***	***<.001****
Surprisal:Meaningful Gesture (Present)	***0.019***	***0.001***	***15.151***	***<.001****
Surprisal:Beat Gesture (Present)	*0.001*	*0.001*	*0.447*	*0.655*
Mean F0:Mouth Informativeness	***-0.003***	***0.001***	***-2.604***	***0.009****
Mean F0:Meaningful Gesture (Present)	***0.003***	***0.001***	***3.067***	***0.002****
Mean F0:Beat Gesture (Present)	*0.001*	*0.001*	*0.640*	*0.522*
Mouth Informativeness:Meaningful Gesture (Present)	***0.004***	***0.001***	***3.549***	***<.001****
Mouth Informativeness:Beat Gesture (Present)	***-0.006***	***0.001***	***-4.400***	***<.001****
Surprisal:Mean F0:Mouth Informativeness	*-0.002*	*0.006*	*-0.419*	*0.680*
Surprisal:Mean F0:Meaningful Gesture (Present)	***0.012***	***0.005***	***2.323***	***0.031****
Surprisal:Mean F0:Beat Gesture (Present)	*0.000*	*0.005*	*-0.053*	*0.958*
Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	*0.007*	*0.004*	*1.783*	*0.089*
Surprisal:Mouth Informativeness:Beat Gesture (Present)	*0.003*	*0.005*	*0.501*	*0.621*
Control Variables				
Word Order	*-0.002*	*0.001*	*-2.031*	*0.042*
Word Length	***0.007***	***0.001***	***6.291***	***<.001****
Sentence Order	*-0.002*	*0.001*	*-1.766*	*0.077*

Baseline	***0.759***	***0.001***	***768.112***	***<.001******
Electrode X	***-0.006***	***0.001***	***-5.954***	***<.001******
Electrode Y	***0.003***	***0.001***	***2.810***	***0.005***
Electrode Z	***-0.004***	***0.001***	***-4.399***	***<.001******

Random Effects		Variance	Std.Dev.
Participant ID	*(Intercept)*	*0.002*	*0.048*
	Surprisal:Mean F0:Mouth Informativeness	*0.001*	*0.024*
	Surprisal:Mean F0:Meaningful Gesture (Present)	*0.000*	*0.022*
	Surprisal:Mean F0:Beat Gesture (Present)	*0.001*	*0.023*
	Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	*0.000*	*0.016*
	Surprisal:Mouth Informativeness:Beat Gesture (Present)	*0.001*	*0.023*

Model: analysis 1

AIC	*BIC*	*logLik*	*deviance*	*df.resid*
892325	*892864*	*-446113*	*892227*	*448591*

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

We found a main effect of surprisal: less predictable words induced more negative N400 (β= -0.008, SE=0.002, p<.001). Crucially, multimodal cues modulated the 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), related to overall less negative N400 component when words were presented with higher pitch or more informative mouth movement. Further, 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: less predictable words showed less negative N400 when accompanied by more informative mouth movements and meaningful gestures. In contrast, mean F0 showed a negative interaction with surprisal (β=-0.006, SE=0.002, p<.001), as less predictable words actually showed larger N400 with higher pitch prosody.

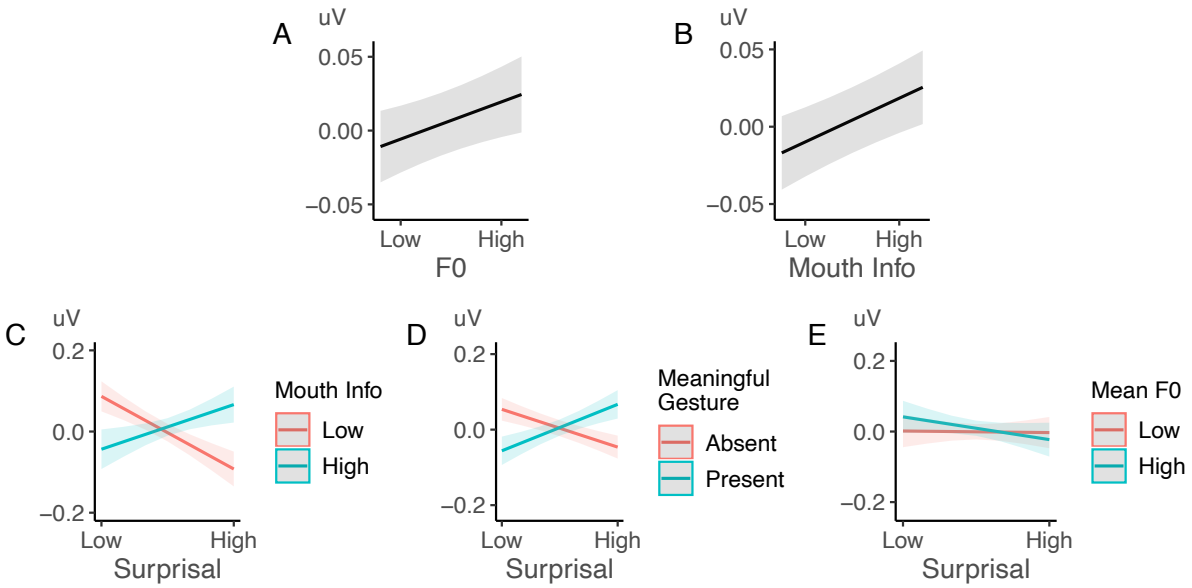


Figure 3.

Multimodal cues each modulate L2 processing. (A) Positive main effect of prosodic accentuation (mean F0). (B) Positive main effect of mouth movement. (C) Positive interaction between mouth informativeness and surprisal. (D) Positive interaction between meaningful gestures and surprisal. (E) Negative interaction between prosodic accentuation and surprisal. Plots depict the predicted value of the mean amplitude of the ERP within 500-800 ms (grey areas = confidence intervals). All following conventions are the same.

We also 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 higher mouth informativeness and higher pitch words. Conversely, there was a positive interaction between F0 and meaningful gesture (β=0.003, SE=0.001, p=.002), but further interact with surprisal (β=0.012, SE=0.005, p=.031). Meaningful gestures elicited even less negative N400 when co-occurring with higher pitch, especially for higher surprisal words. While the interaction between mouth and meaningful gestures is

positive (β=0.004, SE=0.001, p<.001), the interaction between mouth and beat gesture was negative (β=-0.006, SE=0.001, p<.001). More specifically, meaningful gestures induced less negative N400 while beat gestures induced more negative N400 for words with informative mouth movement.

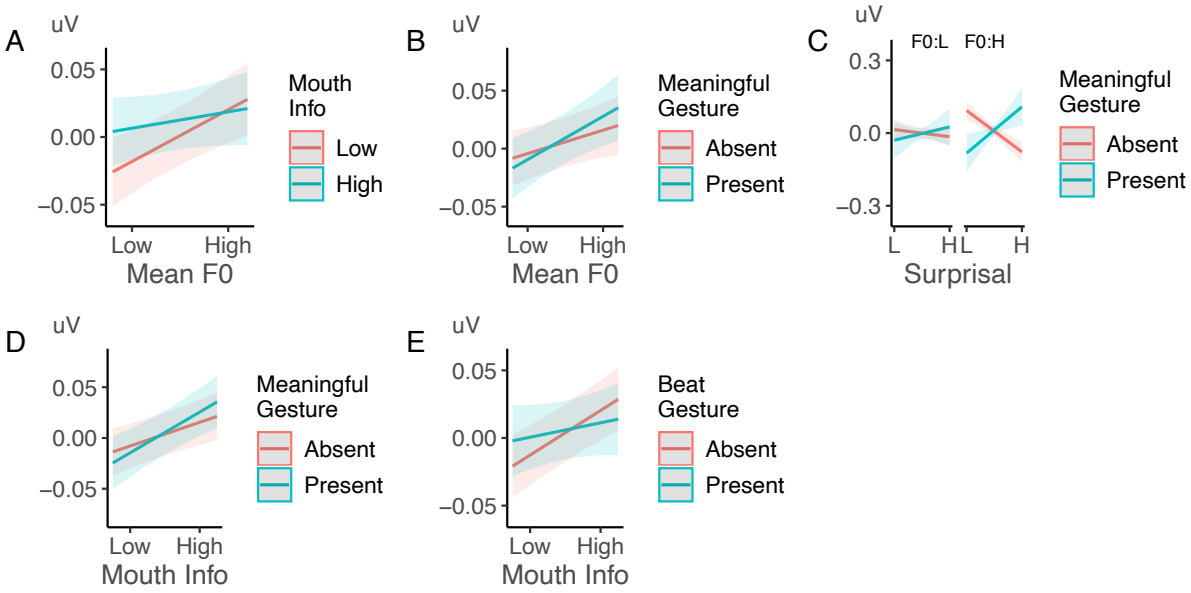


Figure 4.

Multimodal cues interact during L2 processing. (A) Negative interaction between prosodic accentuation and mouth informativeness. (B) Positive interaction between prosodic accentuation and meaningful gestures. (C) Positive interaction between prosodic accentuation, meaningful gestures and surprisal. (D) Positive interaction between meaningful gestures and mouth informativeness. (E) Negative interaction between beat gestures and mouth informativeness.

5.3.3 Analysis 2: Do multimodal cues show the same effect in L1 and L2?

Analysis 2 compared the results from L2 participants to those of L1 participants who were tested with the same materials (acquired from Chapter 4, Exp.2) to investigate any differences between L2 and L1. Full results are reported in Table 2. Positive interaction with L1 (indicating native status being L1) indicates that

L1 listeners showed larger N400 reduction associated with cues or the combination between cues.

Table 2.

Full results for analysis 2: linear mixed effect regression analysis on N400 (500-800ms) in L1 + L2 participants.

Fixed Effects	**β**	**SE**	**t**	**p**
(Intercept)	*-0.002*	*0.007*	*-0.300*	*0.766*
L1	*-0.003*	*0.007*	*-0.407*	*0.686*
Multimodal cues across L1 + L2				
Surprisal	***-0.016***	***0.001***	***-13.866***	***<.001******
Mean F0	***0.008***	***0.001***	***7.188***	***<.001******
Mouth Informativeness	***0.011***	***0.001***	***10.614***	***<.001******
Meaningful Gesture (Present)	***0.004***	***0.001***	***4.977***	***<.001******
Beat Gesture (Present)	***-0.004***	***0.001***	***-3.715***	***<.001******
Surprisal:Mean F0	***-0.003***	***0.001***	***-2.098***	***0.036****
Surprisal:Mouth Informativeness	***0.003***	***0.001***	***3.158***	***0.002****
Surprisal:Meaningful Gesture (Present)	***0.010***	***0.001***	***11.340***	***<.001******
Surprisal:Beat Gesture (Present)	*-0.001*	*0.001*	*-1.015*	*0.310*
Mean F0:Mouth Informativeness	***-0.002***	***0.001***	***-3.118***	***0.002****
Mean F0:Meaningful Gesture (Present)	***0.004***	***0.001***	***4.981***	***<.001******
Mean F0:Beat Gesture (Present)	***0.006***	***0.001***	***5.843***	***<.001******
Mouth Informativeness:Meaningful Gesture (Present)	***0.007***	***0.001***	***8.355***	***<.001******
Mouth Informativeness:Beat Gesture (Present)	*0.001*	*0.001*	*1.000*	*0.317*
Surprisal:Mean F0:Mouth Informativeness	*-0.005*	*0.004*	*-1.266*	*0.213*
Surprisal:Mean F0:Meaningful Gesture (Present)	*0.003*	*0.003*	*0.887*	*0.380*
Surprisal:Mean F0:Beat Gesture (Present)	*-0.002*	*0.004*	*-0.405*	*0.688*
Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	*0.001*	*0.003*	*0.256*	*0.800*
Surprisal:Mouth Informativeness:Beat Gesture (Present)	*0.004*	*0.004*	*1.140*	*0.261*
Comparison between L1 and L2				
L1:Surprisal	***-0.009***	***0.001***	***-8.349***	***<.001******
L1:Mean F0	***0.004***	***0.001***	***3.586***	***<.001******
L1:Mouth Informativeness	***0.004***	***0.001***	***4.219***	***<.001******
L1:Meaningful Gesture (Present)	***0.002***	***0.001***	***2.800***	***0.005****
L1:Beat Gesture (Present)	***-0.006***	***0.001***	***-5.952***	***<.001******
L1:Surprisal:Mean F0	***0.003***	***0.001***	***2.683***	***0.007****
L1:Surprisal:Mouth Informativeness	***-0.007***	***0.001***	***-6.090***	***<.001******

L1:Surprisal:Meaningful Gesture (Present)	***-0.008***	***0.001***	***-8.644***	***<.001******
L1:Surprisal:Beat Gesture (Present)	*-0.002*	*0.001*	*-1.584*	*0.113*
L1:Mean F0:Mouth Informativeness	*0.000*	*0.001*	*0.631*	*0.528*
L1:Mean F0:Meaningful Gesture (Present)	*0.001*	*0.001*	*1.325*	*0.185*
L1:Mean F0:Beat Gesture (Present)	***0.005***	***0.001***	***4.481***	***<.001******
L1:Mouth Informativeness:Meaningful Gesture (Present)	***0.003***	***0.001***	***3.619***	***<.001******
L1:Mouth Informativeness:Beat Gesture (Present)	***0.006***	***0.001***	***6.805***	***<.001******
L1:Surprisal:Mean F0:Mouth Informativeness	*-0.002*	*0.004*	*-0.593*	*0.557*
L1:Surprisal:Mean F0:Meaningful Gesture (Present)	***-0.009***	***0.003***	***-2.562***	***0.014****
L1:Surprisal:Mean F0:Beat Gesture (Present)	*-0.001*	*0.004*	*-0.300*	*0.765*
L1:Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	***-0.006***	***0.003***	***-2.313***	***0.026****
L1:Surprisal:Mouth Informativeness:Beat Gesture (Present)	*0.002*	*0.004*	*0.447*	*0.657*
Control Variables				
Word Order	***0.002***	***0.001***	***3.128***	***0.002****
Word Length	***0.005***	***0.001***	***5.624***	***<.001******
Sentence Order	*0.000*	*0.001*	*-0.551*	*0.582*
Baseline	***0.746***	***0.001***	***1045.731***	***<.001******
Electrode X	***-0.006***	***0.001***	***-8.907***	***<.001******
Electrode Y	***0.002***	***0.001***	***3.002***	***0.003****
Electrode Z	***-0.004***	***0.001***	***-6.200***	***<.001******

Random Effects		**Variance**	**Std.Dev.**
Participant ID	*(Intercept)*	*0.002*	*0.041*
	Surprisal:Mean F0:Mouth Informativeness	*0.000*	*0.022*
	Surprisal:Mean F0:Meaningful Gesture (Present)	*0.000*	*0.021*
	Surprisal:Mean F0:Beat Gesture (Present)	*0.001*	*0.025*
	Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	*0.000*	*0.016*
	Surprisal:Mouth Informativeness:Beat Gesture (Present)	*0.001*	*0.023*

Model: analysis 2

AIC	*BIC*	*logLik*	*deviance*	*df.resid*
1789613	*1790419*	*-894737*	*1789475*	*883515*

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

Overall, L2 participants showed smaller effects: Surprisal had a smaller negative effect in L2 than L1 (β=-0.009, SE=0.001, p<.001), such that the effect of linguistic predictability induced smaller N400 changes in L2 users. In addition, L2

participants also showed smaller facilitatory effect of each multimodal cue (indexed by N400 reduction), including higher pitch (β=0.004, SE=0.001, p<.001) especially for higher surprisal words (β=0.003, SE=0.001, p=.007); higher mouth informativeness (β=0.004, SE=0.001, p<.001) and presence of meaningful gestures (β=0.002, SE=0.001, p=.012). Finally, L2 users showed a smaller negative effect of beat gestures (β=-0.006, SE=0.001, p<.001).

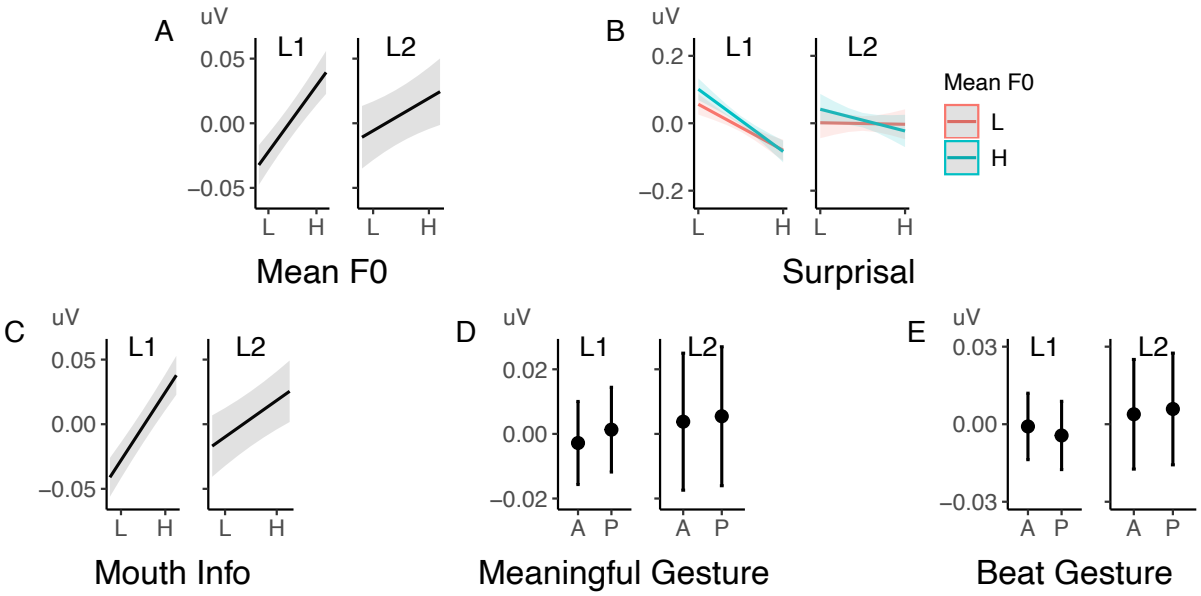


Figure 5.

L2 users showed smaller N400 response to individual cues. L2 participants showed smaller effect for: (A) (positive) prosodic accentuation; (B) (positive) interaction between prosodic accentuation and surprisal; (C) (positive) mouth informativeness; (D) (positive) meaningful gesture; (E) (negative) beat gestures.

L2 participants were also less affected by the interaction between cues. Three-way interactions between native status and the combination of two cues, including pitch and beat gestures (β=0.005, SE=0.001, p<.001) and mouth informativeness and meaningful gestures (β=0.003, SE=0.001, p<.001). L2 (but not L1) participants showed negative interaction between mouth informativeness and

beat gestures (β=0.006, SE=0.001, p<.001). Therefore, N400 reduction associated with the combination between these cues were smaller in L2 than L1.

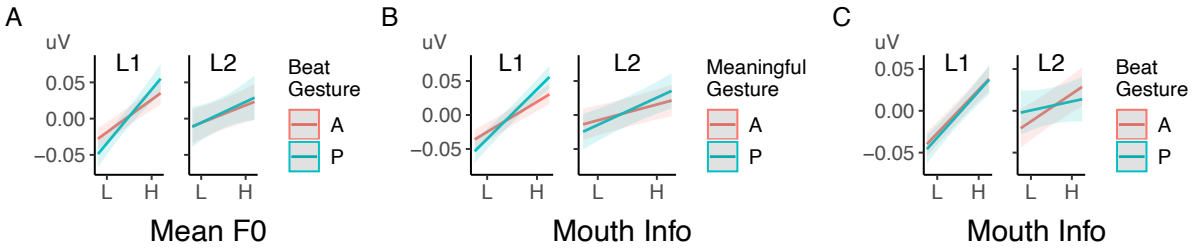


Figure 6.

L2 users showed smaller N400 reduction to combination of cues, including (A) prosodic accentuation and beat gesture; (B) mouth informativeness and meaningful gestures; and (C) mouth informativeness and beat gestures (a negative interaction).

However, despite this general pattern, L2 participants actually showed a larger N400 reduction than L1 speakers for higher surprisal words with meaningful gestures (β=-0.008, SE=0.001, p<.001) or more informative mouth movements (β=-0.007, SE=0.001, p<.001). Moreover, 4-way interactions between native status and surprisal, prosody, meaningful gesture (β=-0.009, SE=0.003, p=.014) and surprisal, mouth informativeness, meaningful gesture (β=-0.006, SE=0.003, p=0.026) showed that when words were less predictable, L2 users benefited more than L1 users from the combination of higher pitch and meaningful gesture as well as the combination of more informative mouth movement and meaningful gesture.

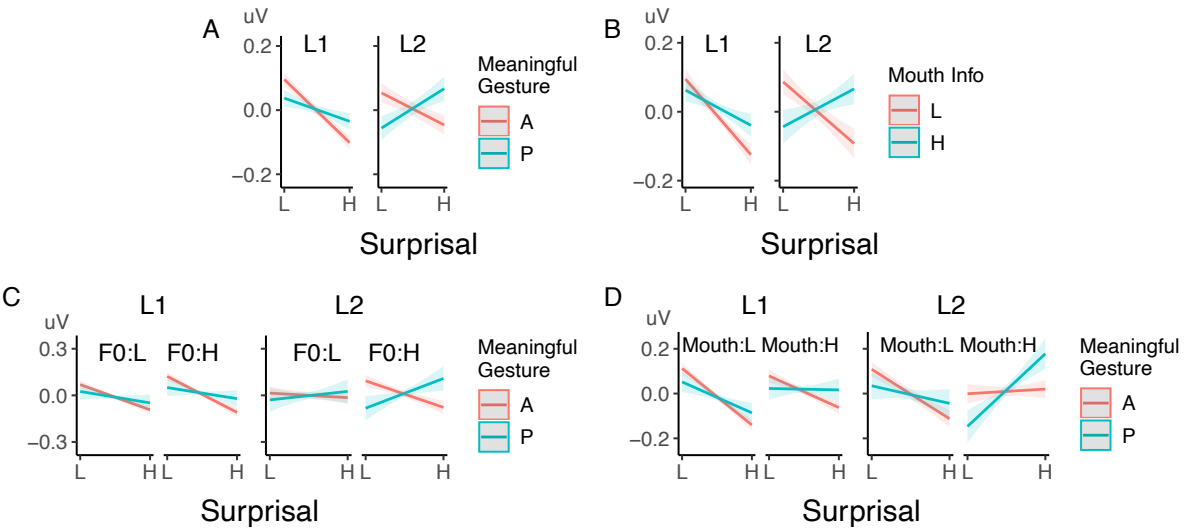


Figure 7.

For less predictable words, L2 users showed larger N400 reduction than L1 for (A) meaningful gestures; (B) mouth informativeness: (C) meaningful gestures with higher pitch prosody; (D) high mouth informativeness with meaningful gestures.

5.4 Discussion

Our study characterised the electrophysiological pattern of highly proficient L2 listeners in naturalistic audiovisual comprehension, with multimodal cues including prosodic accentuation, gestures and mouth movements. First, we established that L2 users are sensitive to linguistic predictability (surprisal), indexed by the N400 effect. We found that the EEG responses of L2 participants are sensitive to linguistic surprisal, inducing more negative EEG responses around 500-800ms post stimulus (later than the L1 participants in Chapter 4). We then characterised how multimodal cues such as prosodic accentuation, gestures and mouth movements modulate N400, to establish the impact of multimodal information on linguistic processing. We found that words with higher pitch induce less negative N400 overall but especially for lower surprisal words, while more informative mouth movements and meaningful gestures elicit less negative N400 for higher surprisal words. This suggests that

while all three cues facilitate comprehension in general, the effect differs for words with higher/lower linguistic predictability. We further found a number of interactions among the cues: higher pitch enhanced the facilitatory effect (indexed by N400 reduction) of meaningful gesture (especially for high surprisal words) but decreased the same effect for mouth movement. Co-occurrence between mouth informativeness and meaningful gestures induced less negative N400 while co-occurrence between mouth informativeness and beat gestures induce more negative N400. This suggests that L2 users are also sensitive to the combination of cues.

Compared to L1 comprehenders reported in Chapter 4, L2 comprehenders showed overall similar effects of multimodal cues and their combinations. L1 and L2 both showed less negative N400 for words with higher pitch, although the N400 reduction is especially large for higher surprisal words in L1 and for lower surprisal words in L2. L1 participants didn't show any replicable effect of mouth informativeness in itself, but L2 participants showed less negative N400 for words with high mouth informativeness, especially for higher surprisal words. Both L1 and L2 participants benefit from the presence of meaningful gestures, and the effect is especially large for higher surprisal words. Finally, while L1 participants reported more negative N400 associated with beat gestures especially when surprisal is higher, such a effect and interaction are not found in L2. In terms of the interactions between multimodal cues, while both L1 and L2 participant showed a positive interaction between prosody and meaningful gestures (mediated by surprisal in L2, so that the combination of higher pitch and meaningful gestures elicit even larger N400 reduction for less predictable words), L2 participants additionally showed negative interaction between prosody and mouth informativeness, so that words with higher pitch but lower mouth informativeness benefit the most. Additionally, L1

participants reported positive interaction between mouth informativeness and both gesture types (meaningful and beat), while L2 comprehenders only showed positive interaction between mouth informativeness and meaningful gestures (further interacted with surprisal, so that the N400 reduction associated with the combination of meaningful gestures and informative mouth movements is larger for higher surprisal words), but negative interaction between mouth informativeness and beat. Further statistical comparison showed that L2 comprehenders overall benefit less from multimodal cues (namely smaller main effect of prosody, mouth informativeness, meaningful gestures and beat gestures, although the main effect of mouth informativeness was not replicable in L1) and their interactions (namely smaller/more negative interactions between prosody and beat gestures and between mouth informativeness and both gesture types). This pattern is generally in line with previous studies. Crucially, however, when words are less predictable based on linguistic context, L2 users do benefit more than L1 from meaningful gestures (especially when co-occurring with higher pitch prosody) and informative mouth movement (especially when co-occurring with meaningful gestures).

Our study provides the first evidence that L2 users, just like L1, also benefit from each multimodal cues and their combinations in a naturalistic context, where multimodal cues co-occur. Although such impact of multimodal cues is smaller in L2 in general, when linguistic predictability is low, L2 users actually benefit more from some cues and interactions (i.e. meaningful gestures & informative mouth movements).

5.4.1 L2 processing is modulated by multimodal cues and their combinations

The first main finding of our study is that different multimodal cues impact L2 processing. In line with previous behavioural studies (e.g. Akker & Cutler, 2003;

Takahashi et al., 2018), we found that prosodic accentuation facilitates L2 comprehension, indexed by smaller N400. However, this effect is smaller for less predictable words, in contrast to L1 listeners. This indicates that while accentuation marks new and less predictable words (Cruttenden, 2006), L2 users are less capable of using this link to facilitate comprehension for words that are less predictable based on linguistic context. One may wonder whether such difference is driven by the typological difference between the native language of our L2 participants, Mandarin, and English (e.g. Mandarin is a tonal language where pitch countour is also used to differentiate meanings), and whether such results is generalizable to L2 users from other language backgrounds. While this question deserves further empirical investigation, we believe this pattern is not solely due to the impact of their first language. First, Mandarin speakers can produce and perceive in English appropriate prosodic accentuation based on newness of objects (Takahashi et al., 2018), and thus should in theory be able to make use of prosodic information in our task. Moreover, L2 listeners across different native languages reported similar patterns of being less capable of mapping prosodic patterns with semantic information (Dutch: Akker & Cutler, 2003; Chinese: Perdomo & Kaan, 2019; Lee et al., 2019). Therefore, it is more likely that L2 users in general are more likely to encounter problems identifying prosodic prominence in online processing of connected speech (e.g. Rosenberg et al., 2010). Alternatively (or additionally), L2 listeners may face limited cognitive resources that constraints their ability to integrate information across channels (e.g. Hopp, 2010; Sorace, 2011).

We also found a facilitatory effect of meaningful gestures. Previous studies found that incongruent meaningful gestures induced larger N400 in L2 comprehenders (Drijvers & Özyürek, 2018; Ibáñez et al., 2010). Our finding further

indicated that naturally occurring congruent meaningful gestures make comprehension easier, as indexed by smaller N400. Previous studies found that L2 comprehenders benefit more from meaningful gestures in comprehension tasks (e.g. Dahl & Ludvigsen, 2014; Sueyoshi & Hardison, 2005) but not recognition tasks (e.g. Drijvers & Özyürek, 2019), potentially because meaningful gestures carry semantic information that can be directly incorporated in meaning processing, but requires additional activation of phonological forms based on meaning to facilitate recognition. As N400 has been known to reflect semantic processing (e.g. Kutas & Federmeier, 2011), our findings are in line with the studies showing that meaningful gestures facilitate L2 comprehension. This effect is especially strong for higher surprisal words, suggesting that semantic information conveyed by meaningful gestures is used by L2 users when linguistic information is difficult. This finding supports the idea that gestures facilitate L2 comprehension by providing extra visual information, either supporting the comprehension of linguistic information in L2, or even potentially by-passing linguistic processing.

We report for the first time that informative mouth movements also modulate N400 in L2, showing a facilitation of L2 comprehension. While previous studies found that seeing the mouth lead 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 in L2, likely by enhancing the recognizability of words.

Our study quantified multimodal cues in their natural co-occurrence, therefore allowing us to assess how these multimodal cues interact in L2 comprehension. Prosodic accentuation enhances the facilitatory effect of meaningful gestures (especially for less predictable words). This may suggest that higher pitch

encourages the attention to other cues present (e.g. Kristensen et al., 2012), thus encouraging the access and integration of gestural information. Alternatively, it may occur because of "local" binding of the cues that can arise as accentuation often co-occur with gestures (Holler & Levinson, 2019). Whereas, prosody showed a negative interaction with mouth informativeness, suggesting that higher pitch facilitates the processing of words with lower mouth informativeness. It is possible that these words are pronounced more clearly with prosodic accentuation, thus the facilitatory effect of mouth is larger when they are accentuated. 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. One possible explanation is that the presence of hand movements may draw participants' visual attention away from the mouth (Drijvers et al, 2019). This shift of attention can lead to different effects based on the type of gestures: meaningful gestures carry additional semantic information which may compensate for such averted attention, while beat gestures cannot and therefore resulting in enhanced processing difficulty.

5.4.2 Different patterns between multimodal processing in L2 and L1

Compared with the pattern of multimodal comprehension in L1 comprehenders reported in Chapter 4, L2 participants showed similar patterns in general. Higher pitch, informative mouth movements and meaningful gestures facilitate the comprehension of both groups, indexed by less negative N400. Moreover, both group benefit from the combination of higher pitch and meaningful gestures, as well as meaningful gestures and more informative mouth movements. Therefore, the overall function of multimodal cues are similar across both L1 and L2 comprehenders.

However, statistical analysis suggested that compared with L1 users, L2 comprehenders show smaller effects of the multimodal cues (in isolation or in combination) in general, indicating that L2 users overall benefit less from multimodal information. 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 broken down, potentially because L2 users are less capable of accessing and integrating multimodal information. This is in line with previous studies reporting smaller facilitatory effect of each cue in L2 (prosody: Akker & Cutler, 2003; Perdomo & Kaan, 2019; Lee et al., 2019; gestures: Drijvers, Vaitonytė, et al., 2019; Drijvers & Özyürek, 2019; Rohrer et al., 2020; mouth: Drijvers, Vaitonytė, et al., 2019; Drijvers & Özyürek, 2019). It is possible that L2 users are less proficient with the patterns of multimodal cues in their non-native language and thus suffer more problem extracting information from multimodal cues when they are presented simultaneously in online comprehension. Another possibility is that the processing of complex connected speech occupies larger cognitive resources in L2 listeners and thus leaves little for the processing and integration of multimodal information (e.g. Hopp, 2010)

Contrary to this general pattern, however, we found that 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). Compared 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. Therefore, when linguistic information is hard, the additional visual

information can aid comprehension by either scaffolding linguistic processing or by-passing linguistic processing, which is especially helpful for L2 users. The contrast between L2 users' general lower ability to benefit from multimodal cues and their even larger benefit from meaningful gestures and mouth may indicate that L2 users are regulating cognitive resources in an efficient way: by adding more weight to the more informative multimodal cues when linguistic information is difficult to process, L2 users may potentially be more likely to achieve successful comprehension. This possibility is supported by previous literatures reporting that L2 listeners tend to look more at the hand (when meaningful gestures are performed, Drijvers, Vaitonytė, et al., 2019) and mouth (in longer connected speech, Birules et al., 2020) than L1 speakers, indicating that L2 users actively seek information from hands and mouth. In general, it has been found that listeners tend to prefer the more reliable source of information. For example, comprehenders has been found to rely more on gestural information when it is always consistent with speech, compared with when it is sometimes incongruent (Holler & Gunter, 2007; Obermeier, Kelly & Gunter, 2015). For another instance, in word recognition task with both auditory and visual information, participants were found to systematically rely more on the noise-free modality (Fourtassi & Frank 2020). The meaningful gestures and informative mouth movements in our experiment may present similar reliably helpful information, which may further encourage L2 comprehenders to increase weights on these cues. 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 context for such adjustments to occur.

Our results call for revisions and specifications of current theories of L2 processing. Current theories of 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. However, while our results are broadly in line with the multimodal frameworks outlined by these theories, in that multimodal cues dynamically modulate comprehensions, these theories are typically underspecified and thus cannot predict individual findings from our studies (e.g. the interaction between mouth informativeness and meaningful gestures is positive, but the interaction between mouth informativeness and beat gestures is negative). Our studies can pose new constraints for these theories to further specify the mechanism underlying multimodal comprehension in L1 and L2.

To conclude, our study provides the first electrophysiological investigation of L2 processing in more naturalistic contexts where more than one cue co-occur. We characterised how multimodal cues jointly modulate L2 comprehension, and highlighted those cues that can be most useful for L2 listeners (i.e. meaningful gestures and mouth movements). Our findings call for a broader focus of current experimental and theoretical works in L2 processing, as our results clearly show that L2 listeners always use multimodal cues that occur in naturalistic settings and actively weight them based on linguistic and multimodal context.

Reference

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

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

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. 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.

10. 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.

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

12. 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.

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

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

15. 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.

16. 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.

17. 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.

18. 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.

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

20. 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.

21. 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.

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

23. 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.

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

25. 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.

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

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

28. 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*.

29. 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.

30. 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.

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

32. 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.

33. 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.

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

35. Zhang, Y., Ding, R., Frassinelli, D., Tuomainen, J., Klavinskis-whiting, S., & Vigliocco, G. (2021). Electrophysiological signatures of multimodal comprehension in second language. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43).

36. Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J. I., & Vigliocco, G. (2021). More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension. *Proceedings of the Royal Society B*, *288*(1955), 20210500.

Chapter 6

6 The role of multimodal cues in concept learning

6.1 Introduction

One crucial finding from the EEG studies in Chapter 4 and 5 is that the impact of each multimodal cue is modulated by the other cues present. For example, the interaction between prosody and meaningful gestures (iconic and deictic) is reliably found across different populations, both native and L2 users of English. Meaningful gestures showed a larger facilitatory effect, indexed by smaller N400, when co-occurring with higher pitch prosody, indicating that prosody enlarges the impact of meaningful gestures in language comprehension. However, due to the naturalistic nature of the paradigm, prosodic accentuation and gestures may well correlate with each other (e.g. Krahmer & Swerts, 2007; Brentari, Marotta, Margherita, & Ott, 2013; Esteve-Gibert & Prieto, 2013), thus this positive interaction may be driven either by the modulation of one cue on the other or by the pure tendency for the two cues to co-occur. Therefore, to further elucidate whether and how gestures and prosody interact in language comprehension, we need to manipulate them. Here we report a controlled experiment in which we manipulate the two cues to investigate their joint impact on learning new concepts and labels for these concepts. Note that the offline measure of memory is not directly comparable with the EEG studies measuring the online processing of individual words. However, due to the restrictions posed by the COVID-19 pandemic, we only performed behavioural experiment online, measuring

whether and how participants' memory of new concepts is affected by the two cues. The procedure has been pre-registered at https://osf.io/c27m9/?view_only=c99fe6012a2e46c082cc9d908bc4511d.

Information encoded with Iconic gestures, gestures that imitate word meaning (e.g. "drinking" – shaping hands as if holding a cup and moving towards mouth), has been associated with better memory (e.g., Riseborough, 1981; So et al., 2012; Huang et al., 2019). For example, So, Sim Chen-Hui, & Low Wei-Shan (2012) presented adults and 4-5-year-old children with videos of single words and asked them to recall the words (without moving their hands). They found that both adults and children recalled more words that were encoded with iconic gestures, compared with the no gesture condition. Further, L2 words accompanied with iconic gestures are also learnt better, measured by higher recognition accuracy (Kelly, McDevitt & Esch , 2009; Huang et al., 2019). Similar mnemonic effects have been found for unconnected sentences (e.g. Cohen & Otterbein, 1992; Feyereisen, 2006) and longer passages (e.g. Dargue & Sweller, 2018a, Dargue & Sweller, 2018b), such that linguistic information with different lengths encoded with iconic gestures is more likely to be remembered. Finally, teaching sessions where the teacher used meaningful gestures, compared with the verbal only condition, are associated with better learning outcomes for children in different subjects (measured by better testing results in, e.g. geometry, Valenzeno Alibali, Klatzky, 2003; math: Church Ayman-Nolley & Mahootian, 2004; conservation of quantity: Ping & Goldin-Meadow, 2008). In an EEG study, Kelly, McDevitt & Esch (2009) found that L2 words that were previously learnt with iconic gestures, compared with those without, elicited a stronger Late Positive Complex (LPC, an event related potential peaking ~600ms post-stimulus), interpreted as an index of stronger recollection. One potential

mechanism by which the iconic gestures improve memory of speech is by providing additional sensory-motor information, therefore, the speech content is encoded in multiple neural pathways, producing more durable memory traces (e.g. Church, Garber & Rogalski, 2007). This possibility is supported by a longitudinal fMRI study (Macedonia, Muller & Friederici, 2011), which found that L2 words learnt with iconic gestures (compared with meaningless self-adjusting gestures, e.g. touching one's face/knee) are more likely to be recalled even after ~60 days post training. These words learnt with iconic gestures are associated with more activations in premotor cortices, indicating that iconic gestures enhance the motor representations of learnt words, which in turn, it is argued, lead to better memory performance.

Compared with iconic gestures that enhance the encoding of information potentially by providing additional motor representations, prosodic accentuations highlight certain parts of the speech thus increasing the prominence of the corresponding information (Kramer & Swerts, 2010). Developmental studies suggested that accentuated words are usually learnt better by children (e.g. Mannel & Friederici, 2013; Filippi, Laaha & Tecumseh Fitch, 2017). For example, in an EEG study, Mannel and Friederici (2013) first familiarized infants with unknown words by presenting them with utterances containing these target words, produced either with or without prosodic accentuations. Then, they tested whether these infants show different EEG signals when presented with these familiarised target words, compared with unfamiliarised unknown words. They found that for 6 month-old infants, the EEG responses (starting from ~300ms post stimulus) are different between unfamiliarised words and words familiarised with accentuations, but not the words familiarised without accentuations. Similarly, adults also learn new words better when they are presented with prosodic accentuations (non-words: Filippi,

Gingras & Tecumseh Fitch, 2014; L2 words: Kushch, Igualada & Prieto, 2017), and are more likely to recall sentences with prosodic changes compared with the monotone version (Harriman & Buxton, 1979). These findings suggested that prosodic accentuation can enhance the memory of linguistic information, possibly by enhancing its saliency, thus attracting more cognitive resources during the processing.

Whereas, although presence of beat gestures also enhance word prominence (Krahmer & Swerts, 2007), their impact on memory and learning is mixed. Austin & Sweller (2014) presented adults and children (3-4 years old) with verbal description of routes and tested their recall of spatial information. They found that while children showed better performance when beat gestures were performed during the description (compared with no gestures), adults did not show such effect. Conversely, So, Sim Chen-Hui, & Low Wei-Shan (2012) reported that only adults but not children (4-5 years old) showed better recall performance when memorizing single words produced with beat gestures. Moreover, other studies reported no mnemonic effect of beat gestures (despite positive effects of iconic gestures) for adults (recalling unconnected sentences, Feyerisen, 2006; recalling longer narratives, Rohrer, Delais-Roussarie & Prieto, 2020) or children (3-6 years old, recalling longer narratives, Macoun & Sweller, 2016). One study by Rohrer, Delais-Roussarie & Prieto (2020) presented L1 and L2 participants with videos of an actress describing events in comic strips and asked participants to re-create the comic based on their recall. They found that while L1 participants showed no gestural effect, L2 participants showed worse recall when beat gestures were produced. Apart from the varying tasks and participants' characteristics, one potential explanation for the divergent findings was proposed by Igualada Esteve-Gibert &

Prieto (2017). They argued that the impact of beat gestures may be strictly local, promoting the recall of the highlighted word exclusively but not the information around it. In a storytelling task, they found that that children recalled the words directly accompanied by a beat gesture better than without beat gestures, but the adjacent non-target words did not differ. This may explain the divergent pattern in the literature to an extent, as studies reporting a positive effect of beat gestures usually measure recall of single words (So et al., 2012, but see Austin & Sweller, 2014, which measured recall of spatial information instead of specific words) while studies reporting null effects typically test for recall of sentence or discourse level information (Feyereisen, 2006; Macoun and Sweller, 2016; however, Rohrer et al., 2020 did not find any mnemonic effect even when focusing on the specific words produced with beat gestures). Alternatively, it is also possible that beat gestures may have overall smaller effect than prosodic accentuation. For example, Kushch, Igualada, & Prieto (2018) found that recall and recognition performance of second language words improved to a smaller extent for beat gestures than prosodic accentuation alone.

In general, previous studies suggested that iconic gestures, prosodic accentuations and beat gestures contribute to memory and learning, although potentially to different extents and at different levels. The facilitatory effect of gestures was found across various aspects of memory and learning: single words with iconic gestures are remembered better (e.g. So et al., 2012), indicating a local effect, so that the information directly associated with iconic gestures has a deeper memory trace, possibly due to additional motor encoding. Sentences and passages with iconic gestures are also remembered better (e.g. Feyereisen, 2006; Dargue & Sweller, 2018a), suggesting the presence of a more global effect, so that information surrounding the iconic gesture may also be remembered better, potentially because

the entire passage is understood better. Apart from the information that is already known by participants, iconic gestures also promote learning of L2 words (e.g. Kelly, McDevitt & Esch, 2009), which involves the creation of a new label for a lexical representation (either pre-existing, e.g. learning the pronunciation of something in a different language, or newly constructed, e.g. learning something unique in another culture). Similarly, iconic gestures are also associated with improved learning outcomes for e.g. geometry (Valenzeno Alibali, Klatzky, 2003), which require the creation of a new concept. Therefore, iconic gestures may have a broad facilitatory effect for memory and learning, potentially because iconic gestures providing additional motor and semantic information, which leads to better understanding of the linguistic input in general. This may be able to result in memory advantage across various levels. In comparison, the effect of prosodic accentuation and beat gestures (if any) may be more restricted towards the specific information accompanied with them (e.g. Kushch, Igualada & Prieto, 2017; Igualada Esteve-Gibert & Prieto, 2017), potentially because both cues are primarily attentional and should only highlight the information directly co-occurring with them (note that while Harriman & Buxton (1979) reported sentences with prosodic changes are more memorable than those that are monotonous, the monotonous version is overall more boring and unnatural, thus might be less memorable). Within the two attentional cues, the effect of beat gestures may also be smaller (Kushch, Igualada & Prieto, 2017), potentially because prosodic variations are compulsory to naturally produced speech, while beat gestures may be more optional.

While gestural and prosodic information typically co-occur in face-to-face communications, very few studies investigated the interaction between prosody and gestures (iconic and beat). To our knowledge, no study investigated whether

prosody and iconic gestures interact in language learning, despite their tendency to co-occur (e.g. Brentari, Marotta, Margherita, & Ott, 2013). With regard to prosody and beat gestures, Llanes-Coromina, Vilà-Giménez, Kushch, Borràs-Comes & Prieto (2018) presented 4-year-old children with longer narratives and found that words with beat gestures and prosodic accentuation are remembered better than those with only prosodic accentuation, indicating that beat gestures have an effect on top of prosody. Kushch, Igualada, and Prieto (2018) further investigated whether the two cues interact by presenting participants with L2 words with beat gestures and/or prosodic accentuation. Apart from the finding that participants performed the best with both cues present, they also found that while the facilitatory effect of prosody was found both with and without beat gestures, the facilitatory effect of beat gestures was only found for words with prosodic accentuation. This indicates that the impact of beat gestures is smaller than prosody, and that prosodic accentuation enlarge the effect of beat gestures when learning novel L2 words. However, Morett and Fraundorf (2019) measured participants' memory of discourse when presented with prosodic accentuation and or beat gestures, and found that prosodic accentuation only facilitated memory for the items where beat gestures were present. Whereas, when the speaker never used beat gestures, prosodic accentuation always enhanced memory. The authors argued when beat gestures are used to make some information more salient, participants may interpret the absence of beat gestures as signalling the information being unimportant, thus may over-ride the effect of prosody.

In sum, there is evidence that iconic gestures, beat gestures and prosodic accentuation each individually facilitate memory and learning, although beat gestures may have smaller and less robust effect. Only a few studies have assessed

whether these multimodal cues interact with each other. Co-occurrence of prosodic accentuation and beat gestures may further highlight the target information, making it even more salient (e.g. Kushch et al., 2018; Morett & Fraundorf, 2019), whereas whether iconic gestures interact with prosodic accentuation to facilitate memory has not been investigated before. However, based on their positive interaction reported in Chapter 4 and 5, co-occurrence of prosodic accentuation and iconic gestures may highlight the meaningful gestural information alongside linguistic information, thus further enhancing its effect.

In a series of behavioural studies, we investigated whether adults benefit from multimodal cues (iconic and beat gestures, prosodic accentuation) and their interactions when learning new concepts. Due to the divergent pattern at different levels of memory reported in the previous literature, we quantified participants' memory of the specific information accompanied with multimodal cues (local effect), the overall message that is not directly accompanied with cues (global effect), and the name of the concepts (the label). We predict from previous literature that iconic gestures and prosodic accentuations should improve the learning (potentially at different levels) while the effect of beat gestures may not be present given the mixed literature (e.g. Feyereisen, 2006; Rohrer et al., 2020). We also predict positive interaction between beat gestures and prosodic accentuation, so that presence of both cues should bring largest facilitatory effect (e.g. Kushch et al., 2018; Morett & Fraundorf, 2019). Finally, we predict a positive interaction between iconic gestures and prosodic accentuation, based on EEG studies presented in Chapter 4 and 5.

6.2 Experiment 1

6.2.1 Methods

6.2.1.1 Participants

100 native English speakers (female = 54, mean age = 34.37, SD = 11.94, range = 18-73) were recruited from Prolific (www.prolific.co) to participate the experiment in Gorilla (www.gorilla.sc). All participants self-reported normal hearing, normal or corrected to normal vision, no autism, language disorder, cognitive impairment or dementia. The sample size was determined using power analysis for ANOVA, which suggested that a sample size of 100 would provide 80% of power to detect a small to moderate sized effect (d=0.3). Although we plan to carry out LMER analysis instead of ANOVA, the LMER analysis generally has larger statistical power (e.g. Baayen, Davidson & Bates, 2007; Meteyard & Davies, 2020) and therefore this sample size should give enough power for the LMER analysis. Participants were paid £3 for the study. All procedure was approved by the UCL ethics committee.

6.2.1.2 Materials

Short video clips introducing objects that are likely to be unknown for native English speakers were used in the current study. First, 15 objects were selected out of 35 unfamiliar objects based on the results of an online norming study, in which 96 native English speakers rated the familiarity of each object. The objects with the lowest familiarity were selected. The resulting objects falls into 4 groups: 1) tools, including strigil, chatelaine, dethorner; 2) musical instruments, including caxixi, cristal baschet, hulusi, shekere; 3) animals, including tarsier, anhinga, axolotl, cassowary, okapi; and 4) fruits, including kiwano, mangosteen, rambutan. Then, for each object, we wrote a short introduction passage (see Figure 1 for example passage and experimental conditions). All passages followed an identical structure: first, an

introductory sentence (e.g. “A strigil is a body cleaning tool that was widely used by ancient civilisations”), which is followed by six pairs of sentences, each containing a pair of keywords. These keyword-pairs describe the property of the object and are gesturable in order to implement the iconic gesture manipulations (e.g. for strigil, the first keyword-pair is “straight handle”, which is a prominent visual feature of the object that can be gestured). Each keyword-pair is embedded in two sentences, with the first sentence containing the keyword-pair (e.g. “The strigil has a straight handle on the side.”) and the second sentence expanding the content provided by the first sentence (e.g. “It was made of metal.”). The second sentence was added so that the stimuli will look more naturalistic (e.g. to avoid the speaker gesturing too frequently and being pantomime-like).

A male native English speaker then recorded each passage in six conditions (3*2), with the keyword-pairs being produced with iconic gesture (IG), beat gesture (BG) or no gesture (NG), and with prosodic accentuation (PA) or without prosodic accentuation (NPA). Iconic gestures were performed with one single stroke; whereas, beat gestures with performed with two separate strokes (corresponding to the two keywords in a pair). All gestures were hold towards the end of the second sentence. To make sure that the prosody was different across the PA and NPA conditions, we manually annotated the mean pitch (F0) and mean intensity for each key word (see Table. 1). The differences were statistically significant (2 tailed pairwise t test, $p<.001$ for both F0 and intensity measures).

Table 1.

Average F0 and intensity for keyword-pairs across conditions

	F0 (Hz)			Intensity (dB)		
	Beat Gestures (BG)	Iconic Gestures (IG)	No Gesture (NG)	Beat Gestures (BG)	Iconic Gestures (IG)	No Gesture (NG)

With prosodic accentuation (PA)	168.84	164.14	163.53	69.41	69.74	68.91
Without prosodic accentuation (NPA)	135.79	130.29	132.59	66.25	66.83	65.61

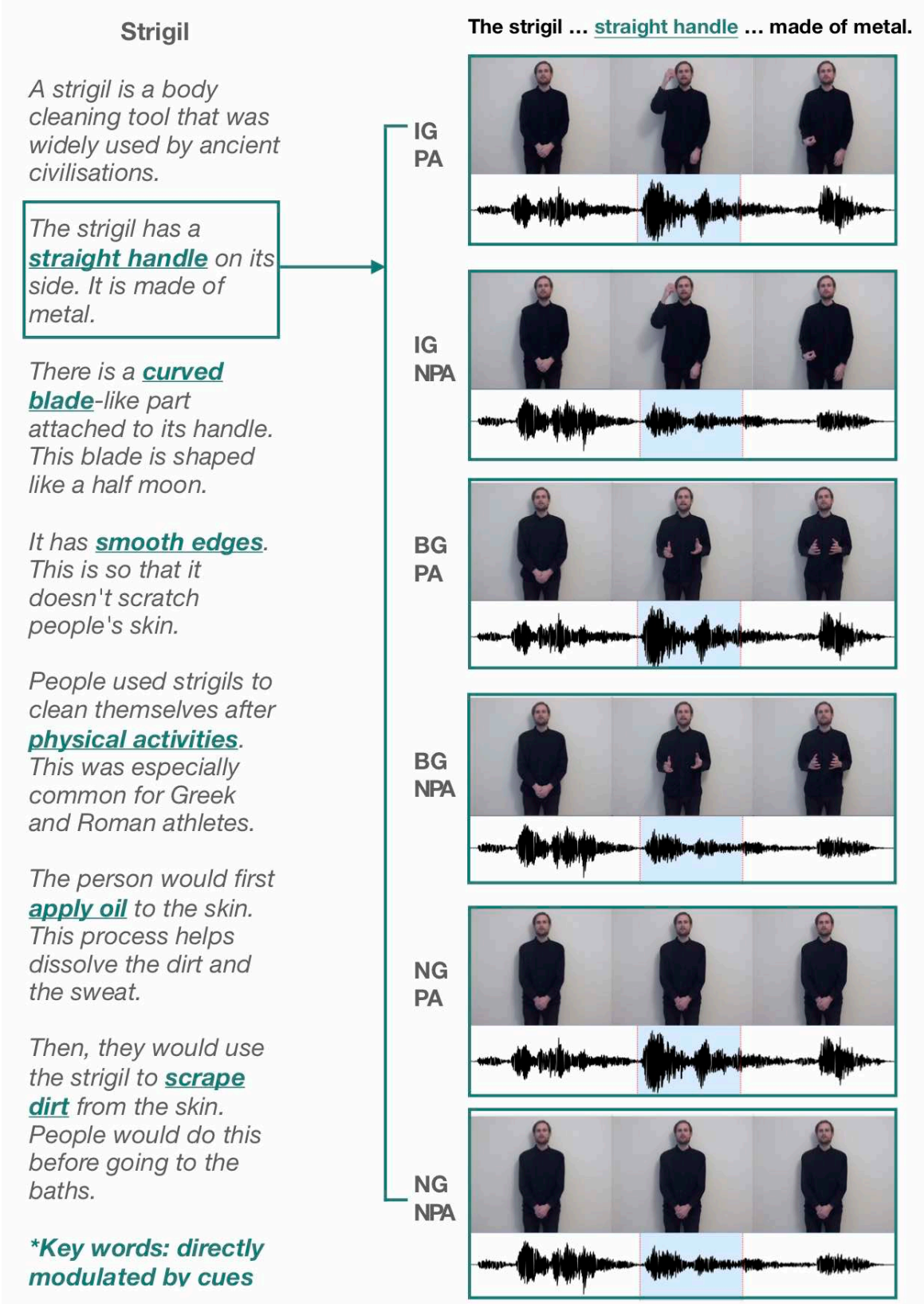


Figure 1.

Experimental material. Short passages about low frequency objects were created, with 6 pairs of key words in each passage directly modulated by gestures (IG = iconic gestures; BG = beat gestures; NG = no gestures) and prosodic accentuation (PA = with prosodic accentuation; NPA = no prosodic accentuation).

6.2.1.3 Procedure

The six conditions per object were semi-randomly assigned to six different lists, such that each list contained each object once and all conditions (with two or three objects per condition). Each participant was then randomly assigned to a list and watched all fifteen videos with randomized order. Before each video, participants were presented with the label of the object and a picture of it for 4000 ms. The videos then started automatically. To check participants' attention, four catch trials were shown randomly, where participants saw a picture and needed to respond to a yes/no question (e.g. "is this a brush?") by pressing buttons on the screen. Participants were instructed to watch the videos attentively as there will be memory tests afterwards. They were also asked not to take notes or breaks and were told that the experiment will stop automatically if they miss more than three catch trials. Participants were only able to participate with a laptop or desktop computer and were asked to complete the task in full screen to ensure that they could see the videos properly. Finally, participants were also required to ensure that their internet connection was stable for smooth streaming of the videos.

After the presentation of all the videos, participants' memory was tested via three separate tasks. Participants first completed a label recall task, where they were instructed to write down all the names of the objects that they learnt. They were encouraged to guess even if they do not remember the exact spelling. Then,

participants completed a label recognition task, where they were asked to indicate whether they learnt a certain word from the videos by pressing yes/no button on the screen. Participants were randomly presented with 30 words, with 15 being the labels of the objects in videos and the other 15 being real words matched on number of syllables, language of origin, phonology, and frequency. Frequency was calculated using ENCOW16A, an English web corpus (experiment labels: *M* = 65.27; dummy labels: *M* = 74.3). After that, participants completed an information recall task, where they were presented with the name of each object and were asked to write down everything they could remember about the object.

Finally, participants were asked to indicate whether they knew any of the items prior to the experiment by ticking the labels. The known words were removed from all following analysis. On average participants took around 35 minutes to complete the experiment. See figure 2 for illustration of the entire procedure.

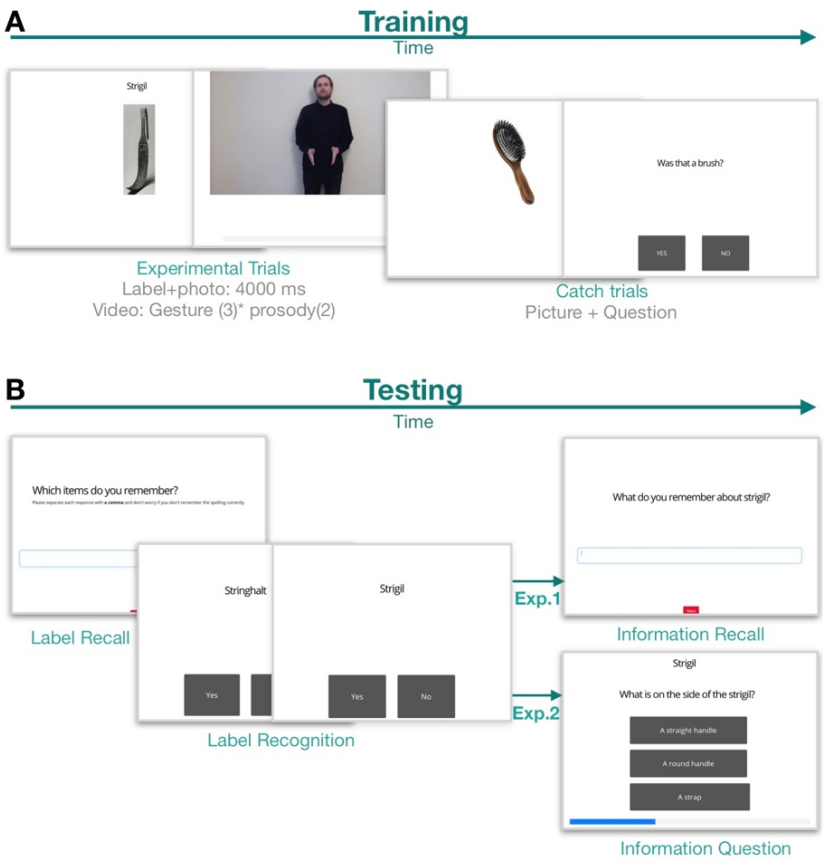


Figure.2

Experimental procedure (Exp.1 and 2). Participants first learned the objects in the training task, where they were presented with the label and photo of the object and the videos with gesture (IG, BG or NG) and/or prosodic accentuation (PA, NPA). 4 catch trials were included to establish participants' attention to the task. Then, participants' memories were measured in the testing task. Their memory of the label was measured by a label recall task ("what items do you remember?") followed by a label recognition task (indicate yes/no for whether an object was presented in the training phase). Then, Exp.1 measured participants' memory of the information associated with each item via an information recall task (e.g. "what do you remember about strigil?"), while Exp.2 measured it via forced-choice questions about each information conveyed by the keyword-pairs.

6.2.1.4 Quantification of performance

Memory of the label (label recall task & label recognition task)

We measured participants' memory of the name of the object via a label recall task and a label recognition task. Performance of the label recall task was quantified as the Levenstein distance (the minimum number of edits required to change string A into string B, e.g. the Levenstein distance between "mat" and "mall" is 2, because you need to replace the "t" into "l", and insert another "l") between participants' guesses and the target label presented in the video. Each word that a participant typed down was compared with all 15 labels. The label with the shortest distance with the response was identified as the target label and the corresponding (shortest) distance was identified as the target distance. The distance values were then reversed and transformed to a scale between 0 and 1. A distance of 0 (i.e. an

identical response) would be transformed into label recall score of 1, indicating a perfect score, while a distance over 5 is classified as incorrect and thus receive a score of 0. The threshold of 4 was determined by taking the 95% threshold in the estimated distribution of Levenshtein distances between the target labels and a sample of pseudowords, randomly generated using the software Pseudo based on the original labels (Jwo & Cheng, 2010).

Performance on the label recognition task was quantified first into accuracy per condition. Additionally, we calculated d' value per participant to check for response bias as exclusion criteria.

Memory of the information (information recall task)

We measured participants' memory for the information conveyed in the passage via the information recall task. Performance on the information recall task was measured separately into local (i.e. memory for the keyword-pairs modulated by gestures/prosody) and global (i.e. memory for the entire passage) performance. For the local recall performance, we automatically identified the number of keyword-pairs that occurred in participants' response. As this measure would not be sensitive to the rephrasing of the original information, responses were also manually coded. For the global performance, we automatically calculated the cosine similarity between participants' responses and the target passages (after removing function words). Again, responses were also manually coded by counting the number of information originally occurred in the introduction (the first and second sentence are counted as two different pieces of information, because the first sentence introduces the key word-pairs, while the second introduces additional descriptions.).

Linking label with information (information recall task)

Finally, we measured how accurately participants linked the label and its information correctly. This is quantified via manual coding of the information recall data, with 1 for correct link and 0 for incorrect link.

6.2.1.5 Statistical analysis

We performed linear mixed effect analysis (LMER) for continuous measures and generalized mixed effect analysis (GLMER) for the binary measures (namely label recognition task) using the lme4 package in R (Bates et al., 2007). Significance of effects was established using the lmerTest package (Kuznetsova, 2015, using Satterthwaite's degrees of freedom). Prosody status (isPA), gesture status (isIG, isBG) and all interactions were added as fixed variables, while the order of training was entered as a control variable. The categorical variables were sum coded so that the effects are compared against the grand mean. To further explore whether there are differences between each level, we additionally conducted the analysis dummy coding the categorical variables. As R by default produce the simple effect at the reference level when categorical variables are dummy coded (instead of the main effect ignoring the reference level), we tested both PA and NPA as reference level for prosody, and NG and BG as reference level for gestures. We excluded from the analysis trials where participants indicated previous knowledge of an object (n=183). We further excluded 2 participants whose d' values from the label recognition task were 2 SD below the threshold, as well as the information recall response from one participant due to directly copying answers from online sources. Participant and object were random variables. The maximal random structure did not converge for any of the analyses even after reduction of each possible random slope, so only intercepts were allowed to vary for each model.

6.2.2 Results

Memory of the label We did not find any significant main or interaction effect between conditions for memory of the label, measured in either recall task or recognition task. Participants' memory of the spelling of the label was poor, as the mean score for the recall task is only 0.13 out of 1 (SD = 0.28), which means on average they get between 4 and 5 letters incorrect per word. Participants showed better memory for the recognition task (M=0.75, SD = 0.43). See figure 3 for distribution of responses.

In the additional analysis using dummy coding, we found a significant interaction between prosody and iconic gesture (B = 0.08, SE = 0.04, p = 0.025) for the recall task, when using beat gestures as reference. Therefore, the effect of prosody is different between beat and iconic gestures when memorizing the label of objects: while iconic gestures with prosodic accentuations yields better label recall, the performance is better when beat gestures are without prosodic accentuation. We also found a significant effect of order (p = 0.009, SE = 0.002, p<.001), so that concepts learnt later are remembered better. We did not find any significant effect of conditions for recognition accuracy.

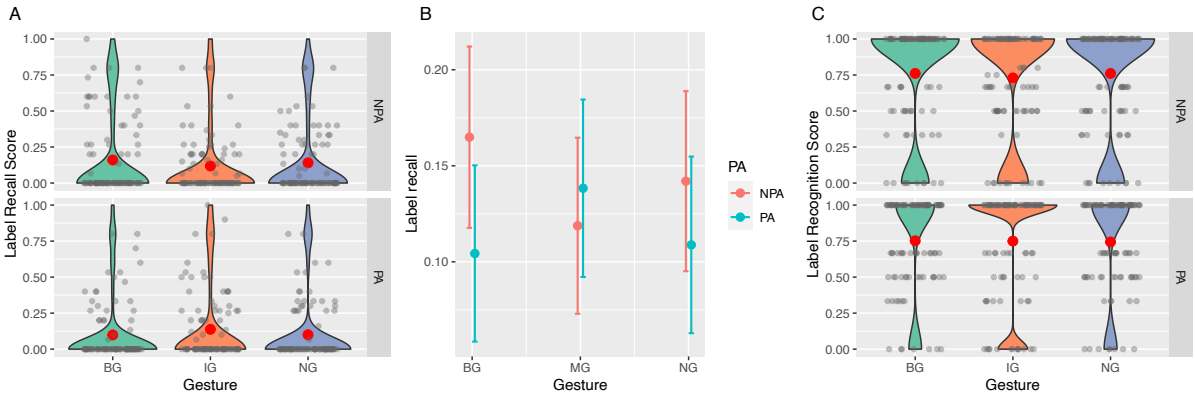


Figure 3.

Effects of prosody and gestures on memory of label. A) Distribution of label recall score. B) Predicted value of label recall. C) Distribution of label recognition accuracy. Each grey dot represents the average score of one participant in the specific condition, the red dot represents the group average. Same conventions apply to all following violin plots.

Memory of the information We did not find any significant effect of conditions for memory of local or global information, either automatically or manually coded. The pattern is the same both for the sum coding analysis and for the dummy coding analysis. Again, participants' memory of the information is generally quite poor. Participants on average remember 1 out of 6 pieces of local information (automatic: M = 1.04, SD = 1.63; manual: M = 1.02, SD = 1.37), and around 16% of global information (automatic: M = 0.17, SD = 0.13; manual: M = 1.82, SD = 2.25, with 12 being full score). See Figure 4. for distribution of data.

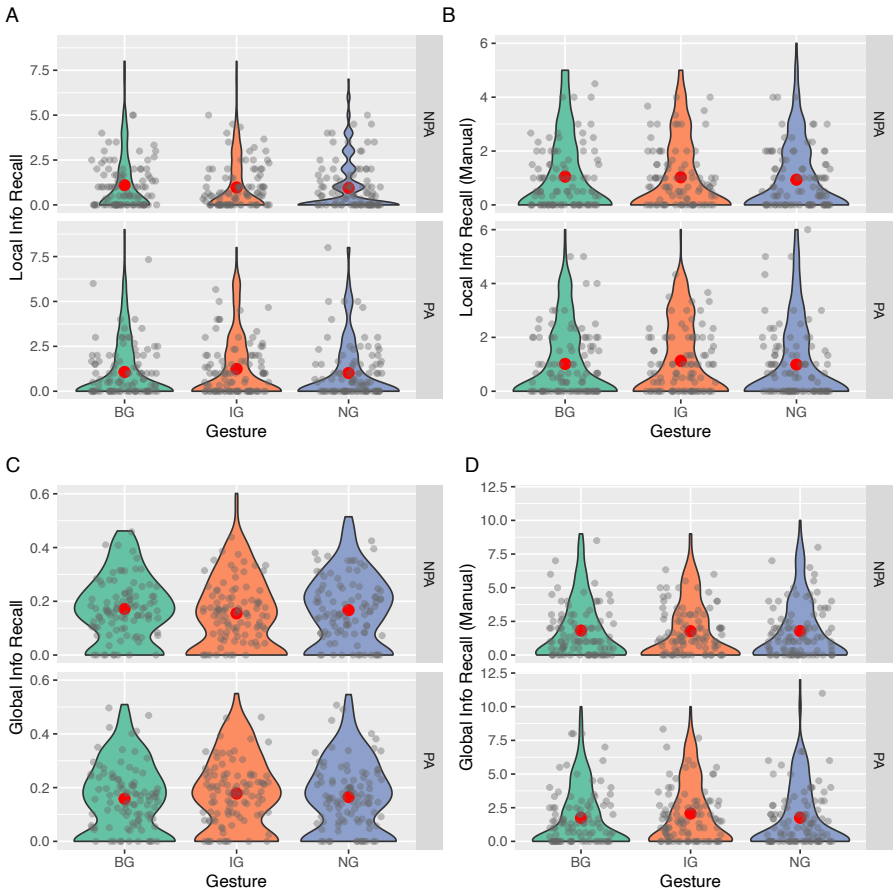


Figure 4.

Effects of prosody and gestures on memory of information of a concept. A) Distribution of local information recall score, measured automatically (hit rate of keyword-pairs). B) Distribution of local information recall score, measured manually (manual scoring of hit rate of keyword-pairs). C) Distribution of global information score, measured automatically (cosine similarity between participants' responses and original passage), D) Distribution of global information score, measured manually (manual counting of first and second sentence information in participants' responses).

Linking label with information We did not find any significant effect of conditions for linking the label with information (M = 0.62, SD = 0.49). In the dummy coding analysis with beat gesture as referent, we found a marginally significant

interaction between prosody and iconic gestures (B = 0.56, SE = 0.33, p = 0.088): iconic gestures tend to result in more accurate link between label and information when the information is also accentuated, in contrast to beat gestures (see Figure 5).

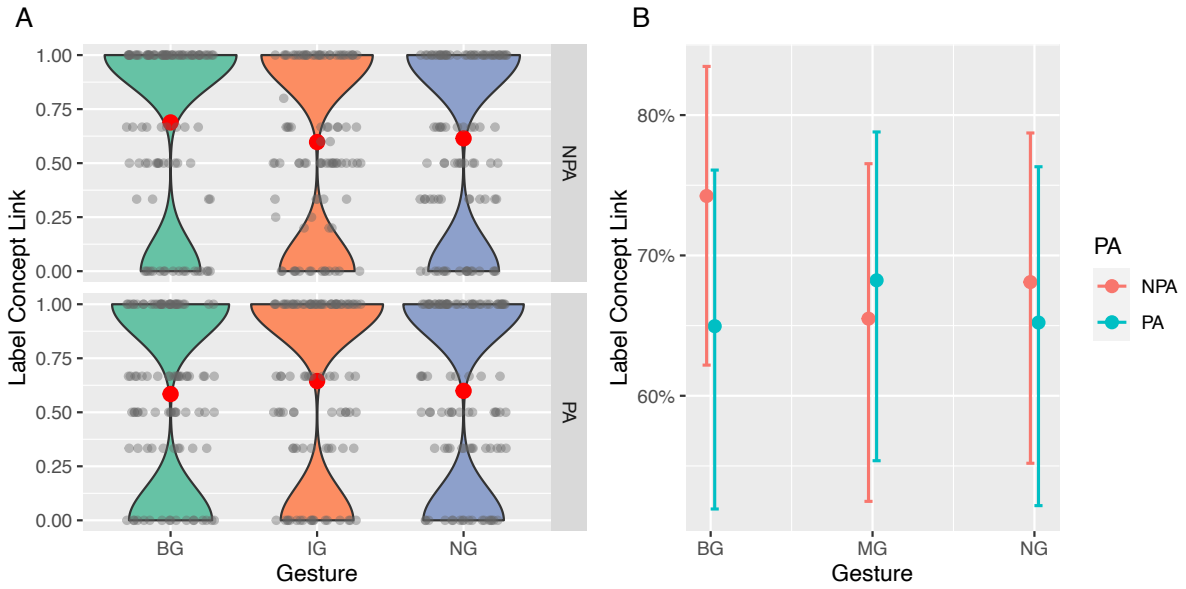


Figure 5.

Effect of prosody and gestures on participants' accuracy in linking the label with the information. A) Distribution of participants' accuracy of label information link, measured manually. B) Predicted accuracy of label information link.

6.2.3 Discussion

Experiment 1 found better recall memory of the label of the concepts and (marginally) better accuracy in associating the information with its label when the videos contains both iconic gestures and prosodic accentuation. However, this effect is only found for the analysis using beat gestures as reference level. Participants also performed poorly in the label recall task and information recall task in general, possibly due to the difficulty of the tasks.

Participants' memory of the label, both in terms of the spelling and in terms of the association with the information, was better when the stimuli contain both iconic gestures and prosodic accentuations. This is broadly in line with our EEG findings in

Chapter 4 and 5, that co-occurrence between prosodic accentuations and meaningful gestures (containing iconic ones) further reduced N400 amplitude, indicating easier processing. It is possible that prosodic accentuations further highlight the presence of iconic gestures, so that the semantic/motor information conveyed by these gestures is comprehended better, leaving a deeper memory trace.

However, we did not find other effects that we predicted (e.g. main effect of iconic gesture or prosodic accentuations). This is likely to be due to the flooring effect in the task. Participants learnt 15 new objects with a total of 195 sentences in our experiment, which may be too challenging to encode. Moreover, the nature of the recall task is that participants must retrieve the memory based only on the label, which poses additional challenge. The overall flooring effect may hide away any effect of multimodal cues and their interactions.

Therefore, to further investigate the effect of multimodal cues in concept learning, we carried out a second experiment in which we reduced the difficulty of the tasks by reduced the number of items (from 15 to 12) and by replacing the information recall task with forced-choice questions about information conveyed by keyword pairs.

6.3 Experiment 2

6.3.1 Methods

6.3.1.1 Participants

Based on power analysis conducted for Exp.1, 100 native English speakers (female = 51, mean age = 39.43, SD = 15.08, range = 18-81) recruited from Prolific (www.prolific.co) were paid £3 to participate the experiment in Gorilla

(www.gorilla.sc). Same selection criteria was used as Exp.1. All procedure was approved by the UCL ethics committee.

6.3.1.2 Materials

We used the same experimental materials as in Exp1. Due to the overall flooring effect in Exp.1, Three items were removed to reduce the difficulty of the task (hulusi, kiwano, rambutan. Hulusi was removed due to the overall lower memory performance in Exp.1, kiwano and rambutan were removed due to the confusability with other fruits). To further reduce the difficulty of the task, we also replaced the information recall task with a forced-choice information task. In this task, one question was asked per keyword-pair (e.g. “The strigil has a straight handle on the side.” Question: “What is on the side of the strigil?”) and three options were given (e.g. “A. A straight handle B. A round handle C. A strap”). The order of objects was randomized across the information memory task, and the order of questions and options are randomized within objects and questions respectively. We dropped the measure of the global memory, because questions about the second sentences in passages would result in a total of 144 questions (rather than 72 in the current design), which might make the task too long. We also dropped any measure of label-concept link, as the new task prohibits such measures.

6.3.1.3 Procedure

The training session is identical with Experiment 1, with the exception that each participant was presented with 12 instead of 15 videos in total. After the presentation of all videos, participants’ memory was tested via three separate tasks. As in experiment 1, participants first completed the label recall and a label recognition task. Then, instead of completing an information recall task, participants were presented with the forced-choice information task described above, in which

participants answered multiple choice question about the information in the videos. Finally, participants were asked to indicate whether they knew any of the items prior to the experiment by ticking the labels (see figure 2 for illustration of the procedure). The known words were removed from all following analysis. On average participants took around 30 minutes to complete the experiment.

6.3.1.4 Quantification of performance

Memory of the label (label recall task & label recognition task)

We quantified participants' memory of the label based on responses from the label recall and recognition tasks using identical method with Exp.1. To calculate the score of label recall task, we calculated the Levinstein distance between the object names and participants' responses, and transformed it into a score ranged from 0 and 1 (higher score represents shorter distance and more accurate recall). Participants' accuracy in label recognition task was also calculated. Additionally, we calculated d' value of the label recognition task as exclusion criteria (as in Exp.1).

Memory of the information (forced choice information task)

Participants' performance in the information memory task was measured by their accuracy per question.

6.3.1.5 Statistical analysis

The statistical analysis was identical to Exp.1. We performed linear mixed effect analysis (LMER) for continuous measures (namely label recall score) and generalized mixed effect analysis (GLMER) for the binary measures (namely label recognition score and forced-choice information task responses) using the lme4 package in R. Prosody status (PA, NPA), gesture status (IG, BG, NG) and their interactions were added as fixed variables, and the order of training was included as a control variable. The categorical variables were sum coded in the main analysis.

Again, we attempted dummy coding categorical variables to investigate difference between all levels. Participant and object were added as random intercept. We excluded from the analysis trials where participants indicated previous knowledge of an object (n=197). We further excluded two participants whose d' values from the label recognition task were 2 SD below the threshold and one participant whose performance in the information memory task was below chance level (0.33).

6.3.2 Results

Memory for the label

For the performance in the label recall task, we found a significant negative main effect of beat gestures (B=-0.03, SE = 0.01, p = 0.027), as well as order (B=0.02, SE=0.003, p<.001). Words with beat gestures are remembered less accurately and words toward the end of the list are remembered more accurately. In the analysis where variables are dummy coded, the negative effect of beat gesture is only marginally significant (B=-0.07, SE = 0.04, p = 0.091, NG as reference), and the effect of order is still significant (B=0.018, SE = 0.003, p<.001). No significant effect of conditions was found for label recognition performance in either sum coding or dummy coding analysis. See Figure 6 for distribution of recall and recognition

scores.

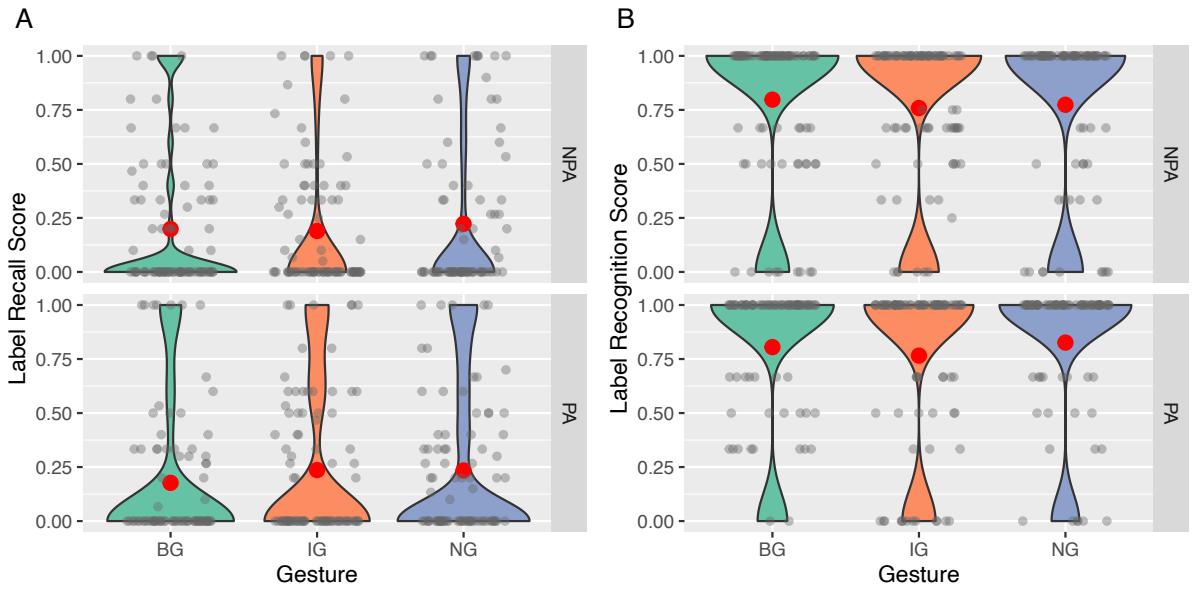


Figure 6.

Effects of prosody and gestures on memory of label. A) Distribution of label recall score. B) Distribution of label recognition accuracy.

Memory for the information We found a significant positive main effect for iconic gestures (B=0.12, SE = 0.04, p = .005) and order (B=-0.026, SE = 0.01, p=.007) in the main analysis sum coding all variables. Information accompanied with iconic gestures were remembered better. In the dummy coding analysis, this effect of iconic gesture is only significant when using PA and NG as baseline (B=0.25, SE = 0.12, p = .039), so that the impact of iconic gestures on information memory is only significant when there are prosodic accentuation and compared with no gesture condition. The effect of order remains significant (B=-0.032, SE = 0.009, p=.001). See Figure 7 for illustrations.

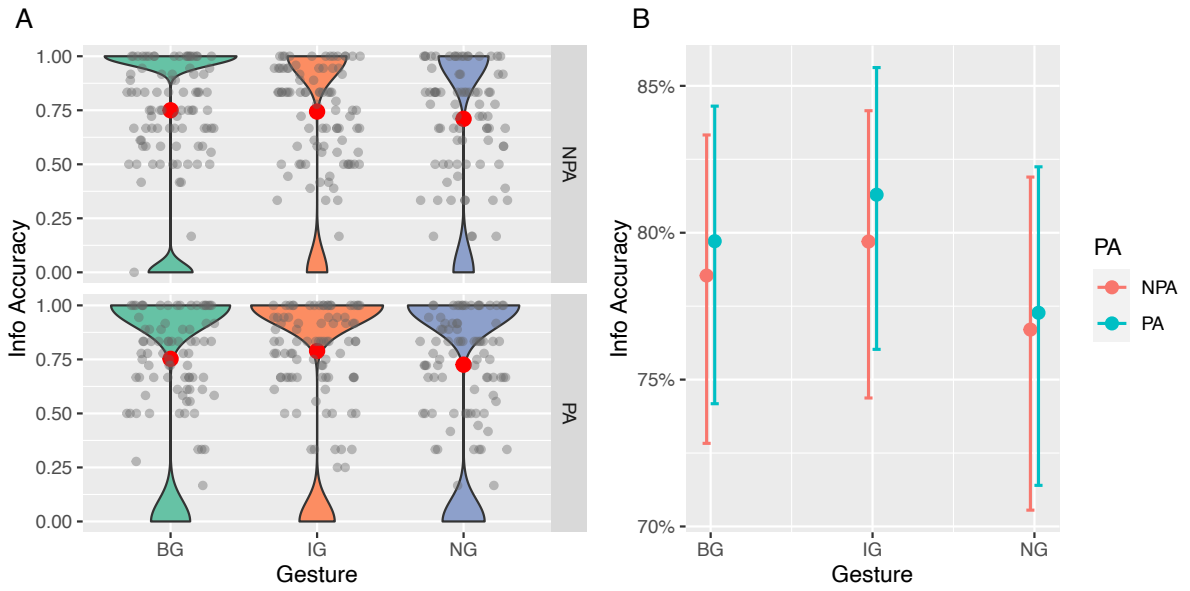


Figure 7.

Effect of gestures and prosody on memory of the information. A). Distribution of information question accuracy. B). Predicted accuracy of information questions for each condition.

6.4 General discussion

Our study presents the first investigation of whether and how gestures and prosody jointly affect the encoding of label and semantic information for new concepts. Across two experiments, we presented participants with videos where an actor introduces new objects with/without gestures (iconic or beat) and prosodic accentuation. We then measured participants' memory for the label of the concept, the content of the concept and the ability to correctly link the label with its content. For the label of new objects, Exp.1 (dummy coded) reported better (recall) memory for items containing both iconic gestures and prosodic accentuation, whereas for the videos with beat gestures, the recall rate was higher without prosodic accentuation. Instead, Exp.2 reported a negative main effect of beat gesture, indicating worse memory for labels when videos contained beat gestures. With regard to the content

information of the concept, Exp.1 did not find any modulation of memory either by prosody or gestures (in information recall task), with a general low performance across all participants. Exp.2 replaced the information recall task with a forced choice information task, where a positive main effect of iconic gestures was found, indicating that iconic gestures improved the memory of information presented with them. Finally, Exp.1 (dummy coded) reported a tendency for more accurate link between the label and content information when iconic gestures co-occur with prosodic accentuation, different from beat gesture. However, Exp.2 did not assess this link due to task constraint. Overall, we reported potential effect of iconic gestures (for remembering content information), beat gestures (for recalling label), as well as interaction between gestures and prosodic accentuation (for recalling label and linking label with the correct information). However, the robustness of these effects remains unclear. The above effects are either not replicated across two experiments (such as the effect of beat gestures and the interaction between iconic gestures and prosody in label recall task), or we have not accessed their replicabilities due to changes of tasks (the effect of iconic gestures on forced-choice accuracy was only assessed in Exp.2 but not Exp.1; while the interaction between iconic gesture and prosody on label-concept link was only assessed in Exp.1 but not Exp.2).

We found that iconic gestures improve the memory of the content of new concepts (Exp.2). This is in line with previous studies reporting similar facilitatory effect of iconic gestures (e.g. Cohen & Otterbein, 1992; Feyereisen, 2006; Church et al., 2007). The co-speech iconic gestures depict the shape or motion properties of a new item, which may result in deeper memory trace (Church et al., 2007; Macedonia, Muller & Friederici, 2001). However, we only found this effect when asking participants specific questions regarding information modulated by gestures

and giving them options to choose from (Exp.2), but not when asking them to write down what they remember of each item (automatic or manual marking). This is different from previous literatures reporting an effect of iconic gestures in information recall tasks (e.g. Cohen & Otterbein, 1992; Feyereisen, 2006; Church et al., 2007). The lack of gesture effect in information recall is potentially related with the flooring effect caused by the difficulty of our material, as participants attempted to learn 15 new objects (Exp.1) with unfamiliar properties, which may be more challenging (than e.g. remembering 2 stories, Cohen & Otterbein, 1992). Moreover, the total number of sentences to be remembered is also larger than previous studies (Exp.1 contains 195 = 15 objects * 13 sentences, compared with e.g. Feyereisen, 2006, which contains 52 unconnected sentences). Alternatively, or additionally, as we presented participants with longer passages, it is possible that the contextual linguistic information is richer and therefore potentially hide away any effect of multimodal cues, as was reported by Cohen and Otterbein (1992) who found that iconic gestures only enhanced recall in scrambled unconnected sentences but not entire passage (but see Dargue & Sweller, 2018, which reported that iconic gestures improved memory of ~2 mins passages).

In terms of the memory of the labels, Exp.1 found that iconic gestures produced better memory (when with prosodic accentuation) for recall of the label of the new concepts, while Exp.2 found that beat gestures induce worse memory of labels in general. Apart from the number of objects being presented (Exp.1 = 15; Exp.2 = 12), the two tasks are otherwise identical. Therefore, we refrain from further interpretation of the results due to the lack of consistency across the two studies.

Finally, we found in Exp.1 that iconic gestures with prosodic accentuation induce better performance than those without in linking the label with the content of a

concept, different from beat gestures. To note, this effect should be viewed tentatively, as it is only marginally significant in Exp.1 and is not verified in Exp.2 (due to task constraints: it is impossible to ask participants questions about an object without naming it). It is possible that presence of prosodic accentuation enhances attention to the iconic gestures, thus enlarging any facilitatory effects of iconic gestures on learning. This explanation is in line with the EEG findings in Chapter 4 and 5, such that words with meaningful gestures and higher pitch showed even larger N400 reduction. However, note that the current behavioural experiment is not directly comparable with the EEG studies reported in the previous chapters: while the EEG studies identified the interaction between prosody and gestures in the online processing of each word, the current behavioural study assessed learning performance in offline measures. Alternatively, as gestures tend to be produced with accentuation (e.g. Krahmer & Swerts, 2007; Brentari, Marotta, Margherita, & Ott, 2013; Esteve-Gibert & Prieto, 2013), iconic gestures without accentuation may trigger an incongruent effect, thus distracting participants from the information being presented, making it harder to correctly associate the information with the label. This pattern is different with beat gesture, as no such interaction was observed. One possibility is that when beat gestures are present, participants may treat them as the only index of prominence, thus overriding any effect of prosodic accentuation. Morett and Fraundorf (2019) found that when participants are presented with videos where beat gestures are manipulated (sometimes present and sometimes absent), prosodic accentuation did not modulate participants' memory; instead, when participants are only presented with the beat gesture absent videos, prosodic accentuation enhances memory of accompanied information. The authors inferred that participants treat beat gestures as the marker of prominence when they (sometimes) appear in the stimuli,

thus the information with no beat gestures but only prosodic accentuations is deemed unimportant. As our participants are presented with a mixture of iconic, beat and no gestures, they may also disregard the prosodic cue due to similar mechanism.

Multimodal cues and learning: material sensitivity and robustness

Compared with previous studies reporting clear facilitatory effect of each multimodal cue (iconic gestures: e.g. Feyereisen, 2006; beat gestures: e.g. Morett & Fraundorf, 2019; prosodic accentuation: e.g. Kushch, Igualada & Prieto, 2017), the pattern in our study is much less consistent. We did not find any replicable effects of beat gestures or prosodic accentuations on memory, and the facilitatory effect of iconic gestures is only found in questionnaires but not information recall task.

One potential reason for the general lack of robustness may be insufficient power. We conducted power analysis prior to the experiment, and assumed a small to moderate effect size d=0.3 for all target variables. However, the standardized regression coefficient for all variables rarely exceeded 0.1 (maximum value being iconic gestures in forced-choice information tasks in Exp.2, which is 0.12). This indicated that the effect of multimodal cues (if any) would be much smaller than assumed, and therefore we may not have sufficient power to identify the effect of each cue reliably. Interestingly, according to a recent meta-analysis (Dargue, Sweller & Jones, 2019), the unbiased effect size of gestures is around .61, which is much higher than observed in the current study. Note that this effect size of gestures is calculated based on studies using comprehension tasks in general, which may not be complelely equivalent with the learning tasks in our experiments. However, the comprehension tasks employed in these studies typically assessed the offline recall or response accuracy of the presented materials, which is the same with our

experiments. The lower effect size in our study may be related with the general difficulty of the task, as the lower performance in general offered smaller variance for factors to account for. It is possible that the materials are simply too hard to remember even with the help of multimodal cues. Another possibility is that previous literature may suffer from “drawer effect”, such that not significant results are not published, which may result in larger estimated effect size based on published materials only.

Another factor that may contribute to the overall smaller effect in our study is the properties of the material. To enhance the naturalness of the stimuli, we asked a native English-speaking student instead of a professional actor to narrate our stimuli. A recent study suggested that typical iconic gestures provide larger facilitatory effect on narrative comprehension compared with atypical ones (Dargue & Sweller, 2018a). Therefore, it is possible that the gestures our actor performed are less comprehensible, therefore being less informative for the learners. Similarly, the prosodic accentuation manipulations are also subtle: to ensure naturalness of the stimuli, the actor was instructed to produce the clips naturally (without exaggerated accentuation/de-accentuation). Further, to make sure any effect of prosody manipulations is due to prosodic accentuation instead of general changes of prosody, the actor was instructed to make the PA/NPA conditions as similar as possible, with changes only in the keyword-pairs. These instructions might make the prosody manipulation less auditorily salient, thus showing no effect overall. However, it is arguable that gestures and prosody produced in daily communications are likely to be “imperfect”, unlike the carefully manipulated ones performed by professional actors. Therefore, the ecological validity of the impact of multimodal cues on memory deserves deeper investigation.

To sum up, our studies suggested that the learning of new concepts may be modulated by multimodal information, namely iconic gestures, potentially in interaction with prosodic accentuations. However, due to the lack of consistency in the results, the effect size of multimodal cues and how properties of the materials affect the impact of multimodal cues needs further exploration.

Reference

1. Austin, E. E., & Sweller, N. (2014). Presentation and production: The role of gesture in spatial communication. *Journal of experimental child psychology*, *122*, 92-103.

2. Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412.

3. Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R package version*, *2*(1), 74.

4. Brentari, D., Marotta, G., Margherita, I., & Ott, A. (2013). The interaction of pitch accent and gesture production in Italian and English. *Studi e saggi linguistici*, *51*(1), 83-101.

5. Church, R. B., Ayman-Nolley, S., & Mahootian, S. (2004). The role of gesture in bilingual education: Does gesture enhance learning?. *International Journal of Bilingual Education and Bilingualism*, *7*(4), 303-319.

6. Church, R. B., Garber, P., & Rogalski, K. (2007). The role of gesture in memory and social communication. *Gesture*, *7*(2), 137-158.

7. Cohen, R. L., & Otterbein, N. (1992). The mnemonic effect of speech gestures: Pantomimic and non-pantomimic gestures compared. *European Journal of Cognitive Psychology*, *4*(2), 113-139.

8. Dargue, N., & Sweller, N. (2018,a). Not all gestures are created equal: The effects of typical and atypical iconic gestures on narrative comprehension. *Journal of Nonverbal Behavior*, *42*(3), 327-345.

9. Dargue, N., & Sweller, N. (2018,b). Donald Duck's garden: the effects of observing iconic reinforcing and contradictory gestures on narrative comprehension. *Journal of experimental child psychology*, *175*, 96-107.

10. Dargue, N., Sweller, N., & Jones, M. P. (2019). When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. *Psychological Bulletin*, *145*(8), 765.
11. Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements.
12. Feyereisen, P. (2006). Further investigation on the mnemonic effect of gestures: Their meaning matters. *European Journal of Cognitive Psychology*, *18*(2), 185-205.
13. Filippi, P., Gingras, B., & FITCH, W. T. (2014). The effect of pitch enhancement on spoken language acquisition. In *Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10)* (pp. 437-438).
14. Filippi, P., Laaha, S., & Fitch, W. T. (2017). Utterance-final position and pitch marking aid word learning in school-age children. *Royal Society open science*, *4*(8), 161035.
15. Harriman, J., & Buxton, H. (1979). The influence of prosody on the recall of monaurally presented sentences. *Brain and Language*, *8*(1), 62-68.
16. Huang, X., Kim, N., & Christianson, K. (2019). Gesture and vocabulary learning in a second language. *Language Learning*, *69*(1), 177-197.
17. Igualada, A., Esteve-Gibert, N., & Prieto, P. (2017). Beat gestures improve word recall in 3-to 5-year-old children. *Journal of Experimental Child Psychology*, *156*, 99-112.
18. Jwo, J. S., & Cheng, Y. C. (2010). Pseudo software: A mediating instrument for modeling software requirements. *Journal of Systems and Software*, *83*(4), 599-608.
19. Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and cognitive processes*, *24*(2), 313-334.
20. 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.
21. 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.
22. Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmertest'. *R package version*, *2*(0), 734.

23. Llanes-Coromina, J., Vilà-Giménez, I., Kushch, O., Borràs-Comes, J., & Prieto, P. (2018). Beat gestures help preschoolers recall and comprehend discourse information. *Journal of Experimental Child Psychology*, *172*, 168-188.

24. Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human brain mapping*, *32*(6), 982-998.

25. Macoun, A., & Sweller, N. (2016). Listening and watching: The effects of observing gesture on preschoolers' narrative comprehension. *Cognitive Development*, *40*, 68-81.

26. Männel, C., & Friederici, A. D. (2013). Accentuate or repeat? Brain signatures of developmental periods in infant word recognition. *Cortex*, *49*(10), 2788-2798.

27. Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092.

28. Ping, R. M., & Goldin-Meadow, S. (2008). Hands in the air: using ungrounded iconic gestures to teach children conservation of quantity. *Developmental psychology*, *44*(5), 1277.

29. Riseborough, M. G. (1981). Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication. *Journal of Nonverbal Behavior*, *5*(3), 172-183.

30. Rohrer, P. L., Delais-Roussarie, E., & Prieto, P. (2020). Beat gestures for comprehension and recall: Differential effects of language learners and native listeners. *Frontiers in Psychology*, 2836.

31. So, W. C., Sim Chen-Hui, C., & Low Wei-Shan, J. (2012). Mnemonic effect of iconic gesture and beat gesture in adults and children: Is meaning in gesture important for memory recall?. *Language and Cognitive Processes*, *27*(5), 665-681.

32. Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, *28*(2), 187-204.

Chapter 7

7 General discussion

7.1 Characteristics of multimodal comprehension

In daily face-to-face communications, multimodal cues, such as prosody, gestures and mouth movements, always accompany speech. While previous studies suggested that each of these cues modulate language comprehension (e.g. Prosody: Cole, 2015; Gestures: Hostetter, 2011; Mouth: Peelle & Sommers, 2015), most of the investigations focused on individual cues out of their naturalistic context, which usually contains other cues.

This PhD thesis investigated how multimodal cues individually and jointly modulate language comprehension in more naturalistic settings, where more than one cue co-occur. We first constructed a corpus of mouth informativeness for English words (Chapter 3) in order to quantify how helpful mouth movements can be in processing words. This study documented the identifiability of mouth patterns in more naturalistic setting (where informativeness was measured based on full words not single phonemes) and served as a building block for the following studies, which used this score to capture mouth informativeness per word.

We then conducted EEG studies (Chapter 4), where native English-speaking participants watched videos of a person speaking with naturally co-occurring multimodal cues. We found replicable patterns that multimodal cues each modulated the EEG component – N400 – sensitive to linguistic predictability: meaningful gesture (iconic and concrete deictic) and pitch prosody were associated with smaller N400, especially for less predictable words, indexing easier comprehension;

whereas beat gestures enhanced N400 especially for higher surprisal words. Further, multimodal cues interacted with each other, with the combination of higher pitch prosody and meaningful gestures, as well as the combination of informative mouth movements and gestures (meaningful or beat) inducing even larger N400 reduction. Therefore, multimodal cues modulated language comprehension individually, but more crucially, interactively, depending on the linguistic context and other cues present.

We then investigated whether and how multimodal cues jointly modulate L2 comprehension and compared it against native comprehenders (Chapter 5). We found that highly proficient L2 comprehenders benefitted from multimodal cues, as meaningful gestures and informative mouth movements reduced N400 especially for higher surprisal words, while higher pitch prosody reduced N400 especially for low surprisal words. L2 comprehenders were also sensitive to the interaction between multimodal cues: higher pitch enhanced the facilitatory effect (indexed by N400 reduction) of meaningful gesture (especially for high surprisal words) but decreased the same effect for mouth movement. Co-occurrence between mouth informativeness and meaningful gestures induced less negative N400 while co-occurrence between mouth informativeness and beat gestures induced more negative N400. Compared with L1 users, L2 comprehenders typically showed smaller facilitatory effect (N400 reduction) for multimodal cues and their interactions; however, they actually benefit more from meaningful gestures (especially when co-occurring with prosodic accentuation) and informative mouth movement (especially when co-occurring with meaningful gestures) when linguistic predictability is lower.

Finally, in order to further tease apart the interaction between gestures and prosody, and to explore how these cues may contribute to learning in addition to

processing, we manipulated their presence in an online experiment measuring whether multimodal cues jointly modulate memory of new concept (Chapter 6). We found some evidence that iconic gestures improved the memory of the information accompanied with it, and improved the correct association between a concept and its label when co-occurring with prosodic accentuation (different from beat gestures).

Overall, our results support the argument that multimodal information is central to language comprehension. We found that comprehenders (both native and non-native) make use of naturally occurring multimodal cues in online language comprehension, indexed by both behavioural memory measurements and electrophysiological markers of comprehension. Importantly, our experiment did not induce any explicit (e.g. direct instruction) or implicit (e.g. hide other cues to enhance saliency of the target cue) requirements for comprehenders to pay attention to these cues. Therefore, our results indicate that such integration of multimodal cues with speech and with each other is automatic. These findings have high ecological validity because in the real-world, as in our experiment, the goal of the comprehenders is to understand the meaning of the speech given all multimodal information. Based on the fact that daily face-to-face communication is invariably accompanied by multimodal information, and our findings that comprehenders naturally take these multimodal cues into account when comprehending speech, we argue that multimodal information, just like linguistic information, is also central to language comprehension in the real-world.

Our results also suggest that the modulation of multimodal cues on comprehension is dynamic. Comprehenders adjust the weight on multimodal cues based on different factors. Linguistic predictability of information affects how much comprehenders rely on each cue, such that the facilitatory effect of multimodal cues

is typically larger when the linguistic information is less predictable and therefore more difficult to process. The presence of other cues also impacts how speaker process co-occurring cues: for example, higher pitch prosody enlarges the effect of meaningful gestures, potentially by highlighting the presence of gestural information. Finally, comprehenders may adjust which cue to place more weight on based on the availability of cognitive resources: L2 comprehenders, whose cognitive resources are usually more limited (as processing a non-native language is more cognitively demanding), assign more weight to the directly meaningful multimodal information (such as meaningful gesture, providing semantic information, and mouth movements, providing sensory information) when linguistic information is less predictable based on prior linguistic context.

Therefore, multimodal language comprehension in the real-world should not be viewed as a static process where the only task of the comprehender is to decode the linguistic signal; but rather a dynamic process where the comprehender actively integrates and balances between different multimodal signals (both linguistic and non-linguistic), in order to construct meaning efficiently under the constraint of limited cognitive resources.

7.2 Efficiency as the drive for dynamic multimodal communication

7.2.1 Efficiency principle in multimodal communication

One potential framework within which to account for why comprehenders always use multimodal cues and dynamically adjust their weight is in the context of efficiency in communication. Some scholars argue that human language is optimised for efficiency (e.g. Gibson et al., 2019). Communication, under information theory, can be abstracted as a process (See Figure 1) where the speaker (“sender”) encodes a message (“source information”) into a signal, which is transmitted to the

comprehender (usually in noisy environment, with “noise” referring to the potential loss of parts of the signal, due to e.g. literal “noise” in conversation). The comprehender (“receiver”) then decodes the signal to recover the message, deriving the destination information (which may not always be identical with the source).

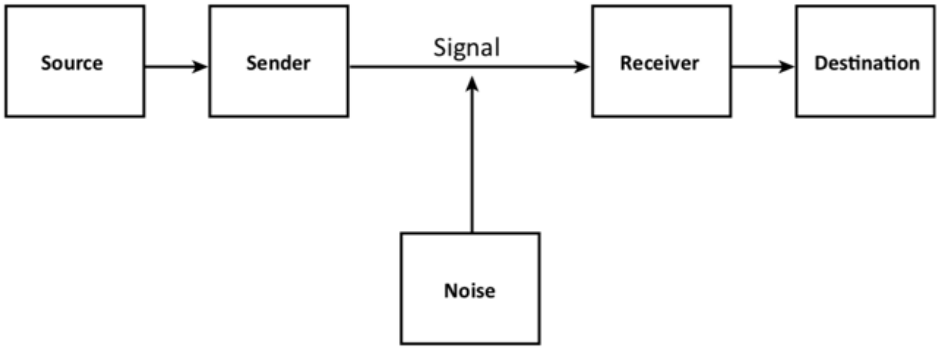


Figure.1

Abstraction of communication under information theory.

The efficiency principle is that the amount of effort when sending a signal (most commonly operationalized as the length of the signal) should be proportional to the amount of information being conveyed (usually operationalized as surprisal, measuring the unpredictability of the information given prior context). More specifically, efficient communication is given by: (a) the tendency to reduce effort in sending the signal whenever possible and (b) the tendency to increase the amount of signal to ensure successful communication. Trade-offs between reducing effort and increasing likelihood of successful communication have been argued to govern both the use of language in communication (e.g., Mahowald, Fedorenko, Piantadosi & Gibson, 2013) and the design of language as a system (e.g. Piandosi et al., 2021). For an example, researchers showed in behavioural experiments that across individual utterences, participants tended to choose the shorter form of a word (e.g. chimp v.s. chimpanzee) when the context highly predicts the word, compared with a

more neutral context (Mahowald, Fedorenko, Piantadosi & Gibson, 2013). Therefore, it was hypothesized that the signal encoded by the sender is optimised to be proportional to the amount of information conveyed: given the requirement of successful communication (which can be viewed as an acceptable similarity between the source information and the destination information), the more predictable source information can be recovered with shorter signal and hence tend to be coded with less effort, whereas the less predictable information needs longer signal and more effort in order to be successfully recovered. Moreover, at the system level, it was found that the length of words across different languages is proportional to the amount of information they carry, measured by averaged surprisal across all contexts, so that higher surprisal words tend to be longer and lower surprisal words tend to be shorter (Piantadosi, Tily, & Gibson, 2021). This suggest that, in accordance with the efficiency principle, speakers across different languages may prefer the shorter form for words that are generally more predictable and vice versa. This systematic preference is then lexicalized via historic changes across different languages, suggesting the generalisability of the efficiency principle. To note, while the efficiency principle offers plausible explanation towards different language phenomenon (e.g. lexicon, Kemp & Reiger, 2012; Gibson et al., 2017; grammar, Jaeger & Levy, 2006; Maurits Navarro & Perfors, 2010), most of these studies investigated whether the properties of language as a system is efficient, but not the performance of individuals during communication (despite the logical assumption that any system level effects are consolidated from the preference towards more efficient forms across individual utterances), and certainly not multimodal communication.

The efficiency principle may also underlie multimodal communication. In a multimodal context, the “effort” in sending a piece of information is not only restricted to the temporal dimension (e.g. durations of signals/words), but additionally, the tendency to send information over multiple channels (e.g. not just linguistic signal, but additional gestures, enhanced prosodic features and mouth movements). Therefore, the efficiency principle in multimodal communication predicts that speakers are more likely to distribute a piece of information over signals encoded via multiple channels (e.g. utter a word with a meaningful gesture) when linguistic surprisal is high. While very limited number of studies investigated explicitly whether the production of multimodal signals are in accordance with the efficiency principle, one exception is Gryzb, Frank and Vigliocco (2022), which found evidences of multimodal efficiency. They reported that the likelihood of iconic gestures produced by a speaker is proportional to the surprisal of a word, so that words with higher surprisal are more likely to be produced with iconic gestures than lower surprisal words. In terms of prosody, some studies suggested that prosodic accentuation marks the predictability of words, so that less predictable words are more likely to be accompanied with prosodic accentuations (e.g. Watson, Arnold & Tanenhaus, 2008; Brandt, Mobius & Andreeva, 2021). Although this phenomenon has been mainly discussed in terms of prosodic accentuation as a pragmatic mark of new information, in light of the efficiency principle, it is possible that words that are less predictable requires more effort in production (i.e. with an accentuation), while words that are more predictable allows for more reduction of efforts (i.e. without any accentuations). So far, to our knowledge, there has been no studies explicitly investigating whether mouth exaggerations that may accompany speech are also predicted by word surprisal. However, given that mouth movements are likely to correlate with prosodic

accentuations, with accentuated words being produced with larger range of mouth movements, it is possible that the efforts in producing mouth movements also follow the principle of efficiency. Therefore, the efficiency principle may be seen as a general principle applying to multimodal communication, with more predictable information enabling more reduction and less effort (i.e. not just shorter pronunciations, but also absence of gestures, flatter/softer prosody and smaller range of mouth movements), and less predictable information requiring more effort (i.e. longer pronunciations, but also presence of gestures, higher pitch and bigger mouth movements) to ensure successful communication.

The efficiency principle applies to both speakers and comprehenders as modifications occur to reduce effort (for the speaker to produce and for the comprehender to process) but also to ensure success in communication (for the speaker and the listener). Comprehenders encounter more difficulty when the information is less predictable. As multimodal cues provide (partially correlated) signals in addition to language, the signal becomes more robust and therefore should allow higher success rate. Therefore, in general, multimodal communication may be a highly (potentially optimally) efficient system as a whole.

Speakers and comprehenders share very similar reality, thus making the distribution of predictabilities similar across both sides. A less predictable information is then likely to be produced with more multimodal signals from the speakers' side, which is likely to be more helpful to the comprehender (as the information is likely to be less predictable for the comprehender as well). On the other hand, when a piece of information is highly predictable, it is likely to carry less multimodal signals, which saves energy both for the speaker in producing it and for the comprehender in processing it. Multimodal communication may therefore be more efficient than

unimodal communication, as it offers more channels across which to distribute the information to be transmitted, while unimodal communication can only modify the length of the signal.

7.2.2 Constraints on multimodal efficiency

While the general principle of efficiency may well apply for multimodal communication, there are processing constraints. For example, when comprehenders encounter less predictable information alongside multimodal signals, their cognitive load may potentially be even larger as they have to process signals across multiple channels simultaneously. However, multimodal comprehension may not be more difficult to process (if not easier than unimodal comprehension): instead of providing independent information (as in e.g. dual task experiment), multimodal channels convey generally partially correlated signals encoding the same source information, and therefore they can predict each other. Along similar lines, Holler and Levinson (2019) proposed that the information convey by the different modalities in multimodal communication is processed as a package, similarly to visual gestalts so that the processing of additional multimodal signals is not necessarily computationally expensive.

Additionally, comprehenders may mitigate the challenge of simultaneous processing more than one signals by placing more weight on the more informative channel. Fourtassi and Frank (2020) found that when presented with auditory and visual information simultaneously, comprehenders systematically prefer the channel that is noise free. In the context of multimodal communication, comprehenders may similarly employ a trade-off strategy, assigning more cognitive resources to the more informative or more reliable channel while potentially ignoring other information sources. For example, multimodal cues have larger effect when linguistic information

is less predictable, suggesting that comprehenders may prefer multimodal signals when linguistic processing is relatively hard. For another instance, Skipper, Goldin Meadow, Nusbaum, & Small (2007) reported that Broca's areas exert less influence on other areas when iconic gestures are present, indicating that multimodal signals, namely iconic gestures, may be able to by-pass the processing of linguistic signal. Similar trade-off may occur between multimodal signals, such as in the case of L2 comprehenders relying more on meaningful cues (i.e. mouth movements and meaningful gestures, when linguistic surprisal is high) but not attentional cues (i.e. prosodic accentuations and beat gestures). Curiously, while we observed in L1 trade-offs between linguistic and multimodal cues (e.g. comprehenders benefit more from meaningful gestures when linguistic information is less predictable), we did not observe such trade-off between multimodal cues in L1 comprehenders in our current studies, as combinations of multimodal cues usually provide larger facilitatory effects. This may potentially be because multimodal comprehension in L1 is comparatively easy and therefore leaves less need to pick between which multimodal signals to rely on. It would be potentially interesting to investigate multimodal comprehension in special population or adverse conditions, which will pose more challenge to multimodal comprehension and activate a stronger trade-off mechanism.

7.3 Neurobiological model of multimodal comprehension

Traditional neurobiological models are typically built on imaging studies presenting participants with linguistic stimuli. For example, Fedorenko and Thompson-Schill (2014) argued that the processing of language activates two distinct brain networks, one core language network (mainly consisted of left lateral frontal and temporal regions) and one domain-general cognitive control network

(bilateral frontal and temporal regions). The separation is based on the stability of network across time and tasks: while the core language network is reliably activated by language tasks, the domain general cognitive control network is activated by non-linguistic tasks as well. However, one crucial problem to such classification is that multimodal cues are typically absent in the linguistic tasks. Therefore, at the face of it, any brain regions or networks sensitive to the processing of multimodal cues seems peripheral at most (if recognized at all), as they are only activated in the "special" multimodal processing tasks, but lacks stability across all "language tasks".

However, one can argue that multimodal processing is not a special case, but instead, the default mode of language comprehension. Language has evolved, is learnt and is most often used in multimodal context, rather than unimodal context (e.g. auditory only, such as talking over a phone; or visual only, such as in reading). Not only the majority of the early evolution of language (conceivably) happened in multimodal context, it is hypothesized that language itself may have originated from gestural communicative systems (e.g. Holler and Levinson, 2014; Kendon, 2004; Tomasello, 2008; Vigliocco, Perniss & Vinson, 2014). Moreover, language is also acquired in a multimodal context, with different cues (e.g. hand gestures: Iverson & Goldin-Meadow, 2005; prosody: Fernald et al., 1989; mouth movements: Lewkowicz & Hansen-Tift, 2012) playing important roles in how young children learn language. Finally, language in the real world is most often used in the multimodal context (in the form of face-to-face conversations or video meetings). The long evolutionary trajectory, the developmental process and the common usage in daily life provided extensive experience of multimodal communication. This could shape the human brain to automatically draw on resources from multimodal information during language processing. As has been shown in our studies, multimodal cues actively

modulate the effect of linguistic predictability in online comprehension, hence the brain network processing linguistic information must constantly exchange information with the (potential) networks processing multimodal cues, so that cues have larger effect when words are less predictable. Moreover, multimodal cues constantly interact with each other, therefore the potentially different networks processing each cue need to communicate with each other via some forms of connections for such dynamic modulation to happen. Such effect of multimodal cues on language comprehension is arguably stable, as we found similar effects across different materials (Chapter 4, Exp 1 & 2) and across different populations (Chapter 5, L1 & L2). Therefore, if we take multimodal communication as the default mode (and the unimodal communication as a special case where the multimodal network is "turned off"), then the networks associated with the processing of multimodal cues should at least be viewed as closely linked with the language network (not dissimilar with the domain-general cognitive control network), or even arguably, a part of the core language network, if stability is taken as the criteria for defining the "core".

Some more recent frameworks can better capture the processing mechanism of multimodal communication. For an example, Holler & Levinson (2019) hypothesized that comprehenders automatically make use of the statistical correlations of multimodal information, so that different sources of information are bonded and processed as a gestalt, enabling faster multimodal processing. This theory is supported by previous findings that speech accompanied with multimodal information shows faster responses (e.g. Holler, Kendrick & Levinson, 2018), and is also in line with our findings that multimodal cues reduced the N400 amplitudes, indexing easier processing. Some neurobiological models further attempted to capture the underlying neurobiological structure enabling multimodal processing. 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 (Skipper, 2015). 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 (Skipper, Goldin-Meadow, Nusbaum & Small, 2007, 2009). 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 (e.g., Skipper, van Wassenhove, Nusbaum, and Small, 2007). Our finding of the interactions between cues is compatible with this this view. However, to note, although our findings are generally in line with these theories, in that multimodal cues consistently and dynamically modulate comprehension, these theories remains largely underspecified. For example, it is difficult to predict or account for the existence and direction of the specific effects reported in our study (e.g. positive interaction between mouth informativeness and meaningful gestures). Therefore, future theoretical works should further specify the underlying mechanism of multimodal comprehension, potentially based on the constraints provided by our findings.

7.4 Future directions of multimodal language comprehension studies

As one of the first studies investigating multimodal comprehension in a naturalistic context, our study suggests that multimodal cues are integral parts of language comprehension and that the joint impact of cues should not be ignored. A lot of empirical and theoretical works remain to be done in order to fully characterize multimodal comprehension in the real world. From the experimental perspective, our study suggests that multimodal cues jointly impact language comprehension. However, we do not know whether such effect varies with different speakers and

comprehenders. Our stimuli are performed by an actor/actress, and thus might be more informative in general. Could comprehenders adjust the relative weight on each cue based on the style of the speaker (e.g. whether they tend to perform more informative gestures or more meaningless gestures, such as grooming; speakers' gesture styles have been found to affect comprehenders' reliance on them, see Obermeier, Kelly & Gunter, 2015)? Similarly, the difference between L1 and L2 comprehenders reported in our studies points to potential individual differences. Apart from language experience, could other factors such as vocabulary size, working memory size or the tendency to attend to individual cues (e.g. mouth v.s. hand) affect how each person process multimodal information?

A critical aspect of future work needs to address the localization of networks responsible for the processing of multimodal communications. Previous imaging works typically focused only on single multimodal cues, which may not capture the pattern of multimodal communication. Are similar regions/network activated per each cue in more natrualistic context with presence of other cues? Will the activation of these networks modulate the processing of linguistic information, such that the linguistic network shows reduced activation or connections with other areas when multimodal information is present (as in e.g. Skipper, Goldin Meadow, Nusbaum, & Small, 2007, that iconic gestures reduced the connectivity between Broca's areas and other areas)? Will the activation of the region/network associated with one single cue modify the activation of other multimodal cues or their connections (e.g. the positive interaction between prosody and meaningful gestures might predict enhanced activation/connection of gesture networks with the presence of prosodic accentuations)?

More generally, the impact of multimodal cues and their interactions found in our studies indicates that multimodal cues and their combinations should not be excluded in the experimental setting. To preserve naturalness in experimental design, future studies can choose to manipulate multimodal cues (as in Chapter 6), quantify multimodal cues (as in Chapter 4 and 5) or keep them stable in the material, and be aware of how specific designs can potentially affect the ecological validity of the results.

Reference

1. Brandt, E., Möbius, B., & Andreeva, B. (2021). Dynamic Formant Trajectories in German Read Speech: Impact of Predictability and Prominence. *Front. Commun. 6: 643528. doi: 10.3389/fcomm*.

2. Cole, J. (2015). Prosody in context: a review. *Language, Cognition and Neuroscience*, *30*(1-2), 1-31.

3. Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in cognitive sciences*, *18*(3), 120-126.

4. Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of child language*, *16*(3), 477-501.

5. Fourtassi, A., & Frank, M. C. (2020). How optimal is word recognition under multimodal uncertainty?. *Cognition*, *199*, 104092.

6. Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... & Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785-10790.

7. Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*, *23*(5), 389-407.

8. Gryzb, B., Frank, S., & Vigliocco, G. (2022). Communicative efficiency in multimodal language

9. Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639-652.

10. Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic bulletin & review*, *25*(5), 1900-1908.

11. Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological bulletin*, *137*(2), 297.

12. Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological science*, *16*(5), 367-371.

13. Jaeger, T., & Levy, R. (2006). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, *19*.

14. Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049-1054.

15. Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

16. Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130302.

17. Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, *109*(5), 1431-1436.

18. Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313-318.

19. Maurits, L., Navarro, D., & Perfors, A. (2010). Why are some word orders more common than others? A uniform information density account. *Advances in neural information processing systems*, *23*.

20. Obermeier, C., Kelly, S. D., & Gunter, T. C. (2015). A speaker's gesture style can affect language comprehension: ERP evidence from gesture-speech integration. Social cognitive and affective neuroscience, 10(9), 1236-1243.

21. Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181.

22. Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526-3529.

23. Skipper, J. I. (2015). The NOLB model: A model of the natural organization of language and the brain.

24. Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2007). Speech-associated gestures, Broca's area, and the human mirror system. *Brain and language*, *101*(3), 260-277.

25. Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2007). Speech-associated gestures, Broca's area, and the human mirror system. *Brain and language*, *101*(3), 260-277.

26. Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2009). Gestures orchestrate brain networks for language understanding. *Current Biology*, *19*(8), 661-667.

27. Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, *17*(10), 2387-2399.

28. Tomasello, M. (2008). Why don't apes point?. *Trends In Linguistics Studies And Monographs*, *197*, 375.

29. Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130292.

30. Watson, D. G., Arnold, J. E., & Tanenhaus, M. K. (2008). Tic Tac TOE: Effects of predictability and importance on acoustic prominence in language production. *Cognition*, *106*(3), 1548-1557.