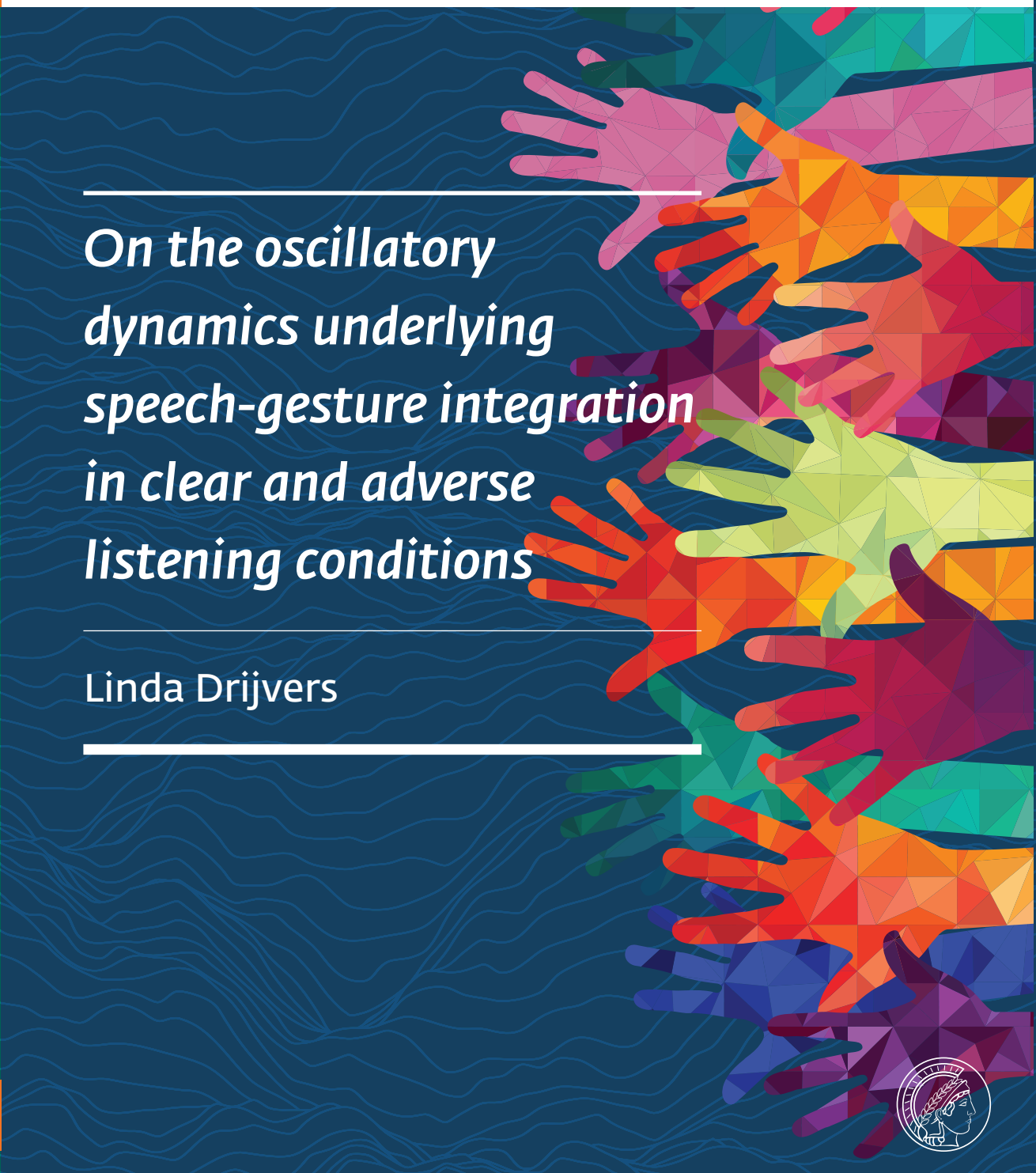
MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

WWW.MPI.NL



On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions

Linda Drijvers

On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions

Linda Drijvers

The work described in this thesis was carried out at the Centre for Language Studies, the Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen and the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, with financial support from the Netherlands Organization for Scientific Research (NWO Gravitation Grant, 024.001.006) awarded to the Language in Interaction consortium.

ISBN/EAN: 978-94-92910-01-1

Design & layout
Wouter Spaak (cover & inside)

Print
Ipskamp Printing

On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions

Proefschrift

ter verkrijging van de graad van doctor aan de
Radboud Universiteit Nijmegen, op gezag van de rector magnificus
prof. dr. J.H.J.M. van Krieken, volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 13 mei 2019 om 14:30 uur
precies

door

Linda Drijvers
geboren op 12 december 1990
te 's-Hertogenbosch

Promotoren

Prof. dr. H.A. Özyürek

Prof. dr. O. Jensen
University of Birmingham, Verenigd Koninkrijk

Manuscriptcommissie

Prof. dr. J.M. McQueen

Prof. dr. M. Gullberg
Lunds universitet, Zweden

Prof. dr. J. Obleser
Universität zu Lübeck, Duitsland

To my family

Contents

Chapter 1

Introduction

Imagine you enter a crowded cafe to have a drink with a friend. As you enter the cafe through the noise-filled seating area where people are amiably chatting, you can already see your friend standing at the counter. When your friend would now try to shout over the sound of grinding coffee machines to ask you whether you want something to drink, you would probably not understand her due to all of the surrounding noise. This might even be more difficult when you are not a native speaker of a language. However, when she would ask the same question while she mimics a glass that is being brought up to the mouth, you probably *would* understand what she means.

Integrating redundant and complementary audiovisual cues is a key strategy to reduce uncertainty under adverse listening conditions, and has been thoroughly studied in the field of non-semantic audiovisual integration (e.g., Rohe & Noppeney, 2016, 2018; Saldern & Noppeney, 2013). In this thesis, I aim to investigate the behavioral and neural integration of audiovisual information at a semantic level, by investigating how iconic gestures, such as that drinking-gesture, enhance language comprehension under adverse listening conditions.

1.1. Studying language comprehension in a multimodal context

To successfully communicate in a face-to-face situation as the one I sketched above, our brain needs to weigh and integrate auditory information (e.g., 'Would you like something to drink?') and sensory, visual information (e.g., a drinking-gesture). However, the integration of these two channels might be challenging under adverse listening conditions, and the weighing of these inputs might depend on multiple contextual factors, such as for example the amount of noise in that cafe, or whether you are a native or non-native language user.

The multimodal nature of these contexts is a feat that is regularly overlooked in current research on language comprehension (see for a similar argument: Perniss, 2018). Instead, the focus of previous research on how listeners understand language has been mostly unimodal and speech-oriented. This seems counter-intuitive, as we rarely encounter language in a purely unimodal context, or even without some form of adverse listening condition. I believe we need to move forward and study how language works as a *multimodal* system in a context in

which adverse listening conditions, such as that noisy cafe or being a non-native listener, are the norm, rather than the exception. Importantly, studying different types of adverse listening conditions also provide more real-life experimental manipulations to study language comprehension in a multimodal context.

So how does a listener's brain make sense of these auditory and visual inputs during multimodal language comprehension, especially when comprehension is challenging? To understand how the brain achieves combining and weighing these inputs, we should consider that especially under adverse listening conditions, visual contextual cues can aid a listener in comprehension. These visual contextual cues could for example consist of visible speech, consisting of lip movements, tongue movement and teeth (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollock, 1954), but also iconic co-speech gestures, which can be described as co-speech hand gestures that depict object attributes, motion, actions and space (McNeill, 1992).

It is important to note that there are different types of relations between these visual contextual cues and the speech signal. For example, the relation between speech and visible speech consists of a form-to-form mapping between syllables and visible speech on a phonological level. Visible speech information can provide temporal details about the speech signal (e.g., on the amplitude envelope) and information on the spatial location of a speaker's articulators (e.g., place and manner of articulation), which can be specifically useful when perceiving speech in adverse listening conditions. The phonological information conveyed by visible speech has been shown to enhance comprehension of speech in adverse listening conditions (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollock, 1954; Campbell, 2008).

On the other hand, iconic gestures are related to speech on a semantic level, due to the similarities to the objects they represent (McNeill, 1992). A very clear cut example of such an iconic gesture is the drinking-gesture that I described in the cafe-example above. Here, the gesture which actually mimics a glass that is being brought up to the mouth to depict the action of 'drinking'. Most iconic gestures however are more ambiguous in the absence of speech. If for example your friend would have not made a drinking gesture, but would have mimicked

a tea bag that is being dipped into a glass, you could have thought that she asked you whether you wanted some tea. Alternatively, that dipping-gesture could have equally well meant yoyo-ing or dipping a cookie rather than tea, or even have been completely ambiguous without the accompanying speech. Speech and gestures are thought to be processed as an integrated system, in the sense that they bi-directionally interact during semantic processing (Kelly, Özyürek, & Maris, 2010). This means that although the gesture can enhance speech comprehension, speech will also enable the listener to make sense of the gestural information. Moreover, previous work has demonstrated that the semantic information conveyed by gestures enhances comprehension of speech in adverse listening conditions (Holle, Obleser, Rueschemeyer, & Gunter, 2010).

So far it remains unknown how visible speech and gestures enhance language comprehension *in a joint context*. Studies that investigated visible speech enhancement have for example used stimuli that only showed the lips or lower half of the face (e.g., Callan et al., 2003; Ross et al., 2007) to eliminate influences from the rest of the face or body. Similarly, studies on how gestures enhance language comprehension have deliberately blocked out the actor's face or mouth, or have only shown the torso of the speaker, to eliminate influences from visible speech (e.g., Holle et al., 2010; Obermeier, Dolk, & Gunter, 2012; Obermeier, Holle, & Gunter, 2011). However, both types of visual articulators help to understand speech comprehension in adverse listening conditions, and should be studied in a joint context to understand how visual contextual cues can enhance language comprehension in adverse listening conditions.

The drinking-gesture in the noisy cafe is an example of how a gesture can enhance comprehension in an adverse listening condition that is induced by an *external* factor. However, this prompts the question whether and how this gestural enhancement would work differently when this adverse listening condition would be induced by an *internal* factor, such as when you are a non-native listener of a language (Lecumberri, Cooke, & Cutler, 2010; Peelle, 2018, see Mattys, Davis, Bradlow, & Scott, 2012, for similar classification). An important difference between the two types of adverse listening conditions is that when you are a non-native listener, your prior knowledge of the language is somewhat diminished,

whereas your prior knowledge of a language is intact when you are in a noisy cafe. This means the origin or cause of the disruption during comprehension differs, and the deficit is language-independent in the context of noise, but language-dependent when you are a non-native listener (Krizman, Bradlow, Lam, & Kraus, 2016). In more detail, one could thus ask whether the weighing and integration of the auditory and visual inputs is similar or different in these two adverse listening conditions, as well as how the two types may interact during language comprehension.

To investigate this, I aim to uncover how gestures, on top of visible speech information, enhance language comprehension in adverse listening conditions induced by both external factors, such as noise, and internal factors, such as being a non-native listener. Second, I will study modulations of neuronal oscillations as a means to investigate the dynamic mechanism by which the brain performs this multimodal integration of speech and gestures, and engages brain regions that are relevant for this process over time. Studying these modulations in clear and adverse listening conditions also allows testing the generality of this mechanism on a neural level.

1.2. Neuronal oscillations as a means to study speech-gesture integration

Neuronal oscillations are thought to subserve cognitive processing and the integration of information from different modalities (Kayser & Logothetis, 2009; Schepers, Schneider, Hipp, Engel, & Senkowski, 2013; Schroeder et al., 2008; Senkowski, Saint-Amour, Hofle, & Foxe, 2011; Varela, Lachaux, Rodriguez, & Martinerie, 2001). Specifically, patterns of rhythmic activity change in amplitude to engage or inhibit brain regions that are relevant for a certain process. For example, low-frequency oscillations in the 'alpha' (8-12 Hz) and beta (13-30 Hz) frequency bands, typically decrease when a certain brain area is engaged (Jensen & Mazaheri, 2010; Klimesch, Sauseng, & Hanslmayr, 2007; Payne & Sekuler, 2014; Pfurtscheller & Lopes da Silva, 1999). Such oscillatory modulations have been studied in unimodal adverse listening conditions, such as the comprehension of auditory degraded speech (e.g., Obleser, Wöstmann, Hellbernd, Wilsch, & Maess,

2012), as well as non-semantic audiovisual integration of ambiguous stimuli (Hipp, Engel, & Siegel, 2011), but not in the context of meaningful, semantic audiovisual integration, such as for example during speech-gesture integration.

Studying the oscillatory mechanisms that underlie speech-gesture integration can be particularly relevant when considering that speech-gesture integration is a higher-order cognitive process that engages many different processes and brain regions over time, such as brain regions that are involved in processing language, vision, audition, and motor actions. However, existing neuroimaging work on this has mostly focused on identifying brain regions involved in speech-gesture integration (e.g., LIFG, pSTS/MTG, visual and motor regions, see: Dick, Mok, Raja Beharelle, Goldin-Meadow, & Small, 2014; Green et al., 2009; Straube, Green, Jansen, Chatterjee, & Kircher, 2010; Straube, Green, Sass, & Kircher, 2014; Straube, Green, Weis, & Kircher, 2012; Willems, Özyürek, & Hagoort, 2007, 2009). It remains unknown how this distributed network of brain regions is engaged in this process *over time*.

The current thesis aims to investigate whether and how *spatiotemporal* neuronal oscillations reflect the process of online speech-gesture integration. Specifically, the studies described in this thesis are grounded in a theoretical framework that assumes a mechanistic role of brain oscillations in different frequency bands in enabling integration from different modalities and engaging brain areas that contribute to this multimodal integration process (Jensen & Mazaheri, 2010; Varela et al., 2001). I will test different cases of speech-gesture integration in clear and adverse listening conditions to test the generality of this mechanism. These adverse listening conditions will be externally induced by speech degradation, and internally induced by non-nativeness.

This thesis will report the results of eight experimental studies that were designed to the questions laid out above. Although the main focus of the thesis is on whether and how modulations of neuronal oscillations reflect the multimodal integration of speech and gestures, behavioral groundwork is needed to investigate the degree to which gestures can enhance language comprehension in adverse listening conditions. For example, it first needs to be understood whether and to what extent gestures, on top of information from visible speech, enhance speech

comprehension in both externally and internally induced listening conditions, and whether this is dependent on different noise levels. This is necessary as most behavioral and neural work has not included the two visual articulators in a joint context (e.g., Holle et al., 2010; Obermeier et al., 2011; 2012). Moreover, it needs to be understood whether and how the semantic integration of speech and gestures differs when externally and internally induced adverse listening conditions interact during comprehension, and whether and how listeners allocate visual attention to these inputs during comprehension. I will thus report on eight experimental studies using behavioral methods, EEG, eye-tracking and MEG to highlight different aspects of the cognitive processing related to how these gestures enhance language comprehension in clear and different types of adverse listening conditions.

1.3. Previous research

To set the background for this thesis, I will first introduce different gesture types and their relation to the speech signal. Then, I will discuss existing behavioral and neuroimaging literature on the role of co-speech gestures during language comprehension. Subsequently, I will review studies that investigated how iconic gestures can aid speech comprehension in externally induced adverse listening conditions (e.g., by speech degradation) and internally induced adverse listening conditions (e.g., by being a non-native listener), as well as studies that investigated the effects of these adverse listening conditions in unimodal contexts. This will be followed by an outline of the studies described in this thesis, as well as an introduction to the different methods used in this thesis. Note that all chapters were written to be understandable when read outside of the context of this thesis (i.e., in journal publications). This inevitably means that some literature and content will be repeated in (several) of the following chapters.

1.3.1. Gesture types and their relation to the speech signal

Co-speech gestures can be described as spontaneous hand movements that co-occur with relevant speech segments (i.e., words, phrases, etc.). These co-speech gestures can be classified into different categories, which all differ in the forms and

functions they have in relation to speech (McNeill, 1992). For example, *deictics,* or pointing gestures, can be used to refer to objects, events or locations and to refer to concrete objects, but also abstract ideas (McNeill, 1992). *Beats* are co-speech gestures that do not have any semantic relationship to speech, but which can be described as bi-phasic, rhythmic flicks of the hand that are synchronous with the rhythmic structure of the speech signal (McNeill, 1992). *Metaphoric gestures* can be described as meaningful gestures that illustrate some abstract concept. Closely related and of main interest to this thesis, are *iconic gestures* (henceforth: gestures), which are gestures that illustrate concrete object attributes, actions and space (Goldin-Meadow, 2005; McNeill, 1992). Iconic gestures have a semantic relationship to speech as they can have similarities to the objects, events and spatial relations that they represent. Even though some iconic gestures can have a very clear and conventional meaning (e.g., a drinking gesture, in which the gesture mimics a glass that is being brought up to the mouth to depict 'drinking'), most iconic gestures are ambiguous in the absence of speech. A fitting example of such ambiguity is for example a 'sweeping'-gesture (see Figure 1). This gesture can, in the absence of speech, easily be understood as 'mopping' or 'rowing', and thus needs speech to be disambiguated.

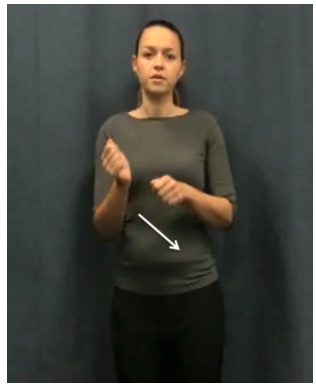


Figure 1: Example of an iconic gesture that is ambiguous without speech. This example could for instance depict 'rowing', 'mopping' or 'sweeping', or apparently even 'punting', and needs speech to be disambiguated.

1.3.2. Iconic gestures and their role in speech comprehension: behavioral and neural evidence

Gestures are systematically related to the speech they co-occur with. This relation can be described as both a temporal relation and a semantic relation. For instance, the onset and stroke of a gesture (i.e., the meaningful part of the gesture), will often commence before the onset of the speech signal, causing the meaningful part of speech and gesture to temporally align (Butterworth & Beattie, 1978; McNeill, 1992). The semantic relation between a gesture and speech can exist in the sense that they convey redundant or complimentary information, such as in the cafe-example that I mentioned at the start of this introduction.

The semantic information that is conveyed by gestures is picked up by listeners. For example, Graham & Argyle (1975) demonstrated that participants were more able to correctly draw two-dimensional shapes when someone described the shape using speech and gestures than by using speech only. Similarly, Beattie & Shovelton (1999) demonstrated that listeners can more accurately describe the size and position of an object when the speaker used gestures in their description than when speakers did not. Other work has argued that the integration of speech and gestures is mutual and obligatory in language comprehension (Kelly, Ozyürek, & Maris, 2010), but the degree of semantic overlap between speech and gestures impacts how a listener integrates these two types of information. In their study, Kelly and colleagues presented participants with action primes and bimodal speech and gesture targets. Primes and gestures were more quickly and accurately related when speech and gesture were congruent than incongruent, and the strength of this incongruency affected the strength of this effect.

Neuroimaging studies have provided more evidence for the fact that the semantic information that is conveyed by gestures is integrated with the speech signal. Most of this evidence comes from studies using electroencephalography (EEG) to investigate event-related potentials (ERPs) related to semantic processing, such as the N400 component. ERPs can be described as averaged deflections in electric potential of the EEG signals. EEG signals are measured and recorded from electrodes that are placed on a participant's scalp. The N400 is a negative-going ERP component that peaks around 400 ms after the onset of a stimulus, and is

considered to reflect the ease of semantic unification or integration of a word in its context (Kutas & Federmeier, 2000, 2014)[1]. More specifically, the amplitude of the N400 typically increases when the meaning of a word violates a certain context or semantic expectancy, and can be triggered by subtle differences in the semantic fit between the meaning of a word and its context. This context can be a single word, a sentence, or even a discourse (e.g., Kutas & Hillyard, 1984; Van Berkum, Zwitserlood, Hagoort, & Brown, 2003). N400 effects have also been observed in response to extralinguistic information, such as world knowledge violations (e.g., "Dutch trains are white", Hagoort & Berkum, 2007), and pictures (Federmeier & Kutas, 2001, 2002).

In the speech-gesture integration literature, the N400 component has been studied as an index of the semantic processing of a gesture. For example, the N400 component has been often studied in violation paradigms, where a more negative N400 amplitude has been observed when speech was presented with a mismatching, incongruent gesture than when a gesture was presented with a matching, congruent[2] gesture (e.g., Cornejo et al., 2009; Habets et al., 2011; Kelly, Kravitz, & Hopkins, 2004; Kelly, Ward, Creigh, & Bartolotti, 2007; Kelly & Lee, 2012; Ozyürek et al., 2007; Sheehan, Namy, & Mills, 2007; Wu & Coulson, 2007).

Another way to probe the semantic integration of speech and gestures is by using disambiguation paradigms. For example, Holle & Gunter (2007) conducted an EEG study in which they presented participants with videos of an actor who uttered sentences together with gestures that could disambiguate the meaning of an ambiguous word later in the sentence. For example, the experimental sentences contained an unbalanced homonym in the first part of the sentence (e.g., 'She controlled the ball') which was then disambiguated in the subsequent clause (e.g. 'which during the game'/'which during the dance'). When the actor uttered the homonym, he would simultaneously produce an iconic gesture that

[1] Note that integration refers to when different sources of information converge on a common memory representation, and unification refers to the process in which a semantic representation is constructed that is not necessarily readily available in memory (Hagoort et al., 2009). Note that throughout this thesis, these terms are often used interchangeably.

[2] Note that matching/congruent gesture and mismatching/incongruent gesture will be used interchangeably throughout this thesis.

either depicted the dominant ('game') or the subordinate meaning ('dance') of the word. The N400 in response to the target word in the subsequent clause was smaller after a congruent gesture and larger after an incongruent gesture. This thus suggests that listeners use the semantic information from a gesture to disambiguate upcoming speech.

Whereas studies using EEG have focused on the temporal characteristics of speech-gesture integration, studies using functional magnetic resonance imaging (fMRI) have attempted to focus on the brain regions associated with speech-gesture integration. fMRI measures the blood-oxygen-level-dependent (BOLD) response in response to speech-gesture integration. The BOLD response is thought to relate to an increase in neural firing in a given brain region, which is then followed by an increase in the flow of deoxygenated blood to that region. This response is, in contrast to electrophysiological methods, an indirect response of neural firing, but is well-suited to characterize the spatial neural infrastructure that underlies a certain process. Using this technique, previous studies have suggested that the integration of meaningful gestures and speech recruits motor regions, visual regions, and the extended language network (posterior superior temporal sulcus (pSTS), middle temporal gyrus (MTG) and left inferior frontal gyrus (LIFG)) in this process (e.g., Dick et al., 2014; Green et al., 2009; Holle, Gunter, Ruschemeyer, Hennenlotter, & Iacoboni, 2008; Holle et al., 2010; Willems et al., 2007, 2009).

However, the role of the pSTS/MTG and LIFG in speech-gesture integration has been disputed. For example, Willems, Özyürek, & Hagoort (2007, 2009) have argued that LIFG constructs new and unified representations of speech and gestures and is thus involved in semantically integrating gestures and speech (e.g., speech and iconic gestures), whereas pSTS/MTG responds to the matching of speech and gestures who have a relatively stable common object representation (e.g., speech and pantomimes), and is thus more related to the mapping and coupling of low-level audiovisual information. In contrast to this, other studies argued that pSTS/MTG is the primary integration region for iconic gestures and speech and that pSTS/MTG is sensitive to semantic aspects of speech-gesture integration, whereas LIFG responds to a modulation or revision of the integrated speech-gesture information (Green et al., 2009; Holle et al., 2008, 2010). Others have suggested

that both LIFG and pSTS/MTG are involved in semantically integrating the two inputs (Dick et al., 2014). Recent work using continuous theta burst stimulation (cTBS) and repetitive transcranial magnetic stimulation (rTMS) has attempted to investigate whether LIFG and pSTS/MTG are causally involved in this integration process. This work suggested that both regions possibly work together during speech-gesture integration, with LIFG regulating strategic semantic access on temporal storage areas, whereas pSTS/MTG is involved in accessing supramodal representations (Zhao, Riggs, Schindler, & Holle, 2018; but see Drijvers & Trujillo, 2018 for a critical commentary).

In summary, what becomes clear is that although there is ample work on the *temporal* and *spatial* correlates of speech-gesture integration, it remains unknown what the *spatiotemporal* correlates of the neural integration of speech and gestures are. Such an approach would aid in disentangling how different brain areas involved in speech-gesture integration engage in this process over time. Second, it remains unclear whether this integration would work similarly when listeners are challenged by external and internal adverse listening conditions during speech-gesture integration. For example, the weighing and integration of audiovisual inputs might differ when comprehension is challenging. These two types of adverse listening conditions thus form a manipulation of speech-gesture integration to understand the cognitive and neural processes that underlie this multimodal integration.

1.3.3. Iconic gestures and their role in speech comprehension in adverse listening conditions: behavioral and neural evidence

While the previously discussed studies have demonstrated that listeners process and integrate the semantic information that is conveyed by gestures in clear speech, few studies have also touched upon speech-gesture integration in adverse listening conditions. These adverse listening conditions can exist because comprehension is challenged by external factors, such as noise (Peelle, 2018), but also internal factors, such as being a non-native listener (Lecumberri et al., 2010). Below I will review relevant literature in both types of adverse listening conditions, in a multimodal and unimodal context.

1.3.3.1. Adverse listening conditions induced by external factors

When zooming into the existing literature considering these external and internal factors that induce adverse listening conditions, only a few studies have focused on the effect of external factors impacting speech-gesture integration. Behavioral work has demonstrated that gestures occur more frequently in externally induced adverse listening conditions, (Hoskin & Herman, 2001; Kendon, 2004), and that listeners use gestural information for comprehension when speech quality is suffering (Rogers, 1978). Early work has demonstrated that gestures facilitate word-recall, especially when speech is less intelligible (Riseborough, 1981).

Similar results were found in neuroimaging studies on speech-gesture integration in externally induced adverse listening conditions. For example, Obermeier et al. (2011) demonstrated that when participants are not asked to explicitly attend to gestures and there is no temporal overlap between a word and a gesture, speech-gesture integration does not occur. However, when a listener observes speech and gestures in multi-talker babble noise, an external factor that influences comprehension, the gestural information is integrated with the speech signal to disambiguate the meaning of an utterance. Speech-gesture integration can thus be modulated by specific characteristics of a communicative situation.

Other work by Holle et al. (2010) studied speech-gesture integration in multi-talker babble noise to zoom in on the spatial brain bases of how iconic gestures and speech are integrated during comprehension. They focused on two key properties of multimodal integration to identify regions involved in this process, namely bimodal enhancement and inverse effectiveness. Bimodal enhancement refers to the fact that when auditory and visual stimuli are presented in combination, the neural response to these stimuli surpasses the sum of the unimodal responses to these stimuli in isolation (Meredith & Stein, 1983). Bimodal enhancement is thought to be strongest for unimodally least effective stimuli (i.e., inverse effectiveness). Applied to investigating speech-gesture integration under adverse listening conditions, this could mean that when the signal-to-noise ratio of speech (SNR) is low, gestural enhancement could be greatest. Brain areas that are relevant for speech-gesture integration (as identified by bimodal enhancement) can then be probed for sites where gesture facilitates speech comprehension by

enhancing the neural response (as indicated by inverse effectiveness). To achieve this, participants were presented with videos of an actor with a masked face who uttered short sentences (e.g., 'And now I grate the cheese") accompanied by an iconic gesture or not, while speech was presented in a good (+2 dB) or moderate (-6 dB) SNR using multi-talker babble noise. Bilateral STS and STG, but not LIFG, were found to be sensitive to bimodal enhancement, especially when speech quality was low. This confirmed the inverse effectiveness principle, by showing that the neural response in these areas was enhanced when a gesture enhanced comprehension of speech under adverse listening conditions. However, although this phenomenon thus has been investigated in both the spatial and temporal domain, the spatiotemporal neural dynamics that support speech-gesture integration in externally adverse listening conditions remain unknown. Moreover, the stimuli used in this experiment blocked out all facial features that normally co-occur with these gestures.

In the unimodal domain, similar brain areas have been observed in studies that investigated the spatial neural correlates of degraded speech comprehension. Intelligibility-specific responses to speech have been identified in STS (Scott, 2000; see for reviews: McGettigan et al., 2012; Peelle, 2018), and frontal, motor and premotor regions are thought to contribute to the comprehension of degraded speech (Adank & Devlin, 2010; Davis & Johnsrude, 2003; Eisner, McGettigan, Faulkner, Rosen, & Scott, 2010; Obleser, Wise, Alex Dresner, & Scott, 2007). Other work that investigated oscillatory modulations during unimodal auditory degraded speech comprehension have reported enhanced parietal alpha power when speech is degraded (e.g., Becker, Pefkou, Michel, & Hervais-Adelman, 2013; Drijvers, Mulder, & Ernestus, 2016; Obleser & Weisz, 2012; Strauß, Wostmann, & Obleser, 2014; Weisz, Hartmann, Müller, Lorenz, & Obleser, 2011; Wostmann, Herrmann, Wilsch, & Obleser, 2015). This enhanced alpha power is thought to reflect an increased auditory cognitive load when the language processing system is inhibited or challenged due to acoustic degradation. It however remains unknown whether similar results would be obtained in a multimodal semantic context, where visual semantic information might lower this imposed load and aid comprehension.

The role of semantic context in the comprehension of unimodal degraded speech has been mostly studied in an auditory sentential context. These studies have for example revealed that native listeners demonstrate a reduced N400 response in reaction to incongruent items presented in degraded speech as compared to clear speech. This N400 response has even been found to be completely absent when the speech signal is too severely degraded (Aydelott, Dick, & Mills, 2006; Boulenger, Hoen, Jacquier, & Meunier, 2011; McGettigan et al., 2012; Obleser & Kotz, 2011). Strauß, Kotz, & Obleser (2013) proposed that degraded speech might narrow the expectancies a listener has about the speech signal. This would mean that when the sensory input is diminished, the neural system might rely more on signal-driven expectancies than on contextual information. Yet, this response might be modulated when a listener is presented with a visual semantic context instead an auditory semantic context. Here, one could expect that a visual semantic context (e.g., a gesture), which is unaffected by auditory degradation, might allow for listeners to elicit predictions about the degraded word, which could result in enhanced comprehension.

1.3.3.2. Adverse listening conditions induced by internal factors

Similar to research on how external factors impact speech-gesture integration, there is little research that investigated how internal factors impact speech-gesture integration, or how internal and external factors interact during comprehension. Such an internal factor can, for example, be introduced by having a hearing impairment, or when someone is a non-native listener of a language. In such situations, it could be imaginable that the language processing system takes into account as much contextual information as possible to aid comprehension.

This is especially relevant for hearing-impaired listeners, who are challenged by adverse listening conditions on a constant basis. These listeners might have changed their processing strategies to optimally take into account information from additional sources next to speech, such as gesture. Obermeier et al. (2012) investigated this assumption by presenting participants with a disambiguation paradigm that followed the design of Holle & Gunter (2007). Here, participants were presented with videos in which an actress uttered sentences that were

accompanied by gestures in either clear speech and multi-babble noise (Experiment 1, normal hearing listeners) or clear speech (Experiment 2, hearing-impaired listeners vs. normal hearing listeners). In these sentences, a homonym would occur that could be disambiguated by a gesture that was presented earlier in a sentence. This gesture was either related to the subordinate or the dominant meaning of the homonym, and was used to investigate the disambiguation success of the sentence. In Experiment 1, normal-hearing listeners only incorporated the gestural information when speech was presented in noise, but not when speech was clear. The results of Experiment 2 demonstrated that hearing-impaired listeners, but not normal-hearing listeners, indeed used these gestures to disambiguate the target words in clear speech. These results suggested that both external (i.e., multi-talker babble noise) and internal (i.e., hearing impairment) impact speech-gesture integration. However, how externally and internally induced adverse listening conditions impact speech-gesture integration in a joint context remains unknown.

Another example of an internal factor that introduces an adverse listening condition is being a non-native listener of a language. An important difference between these two internal factors is that whereas a hearing impairment introduces a permanent internal factor impacting language comprehension, being a non-native listener does not. For hearing-impaired listeners, Obermeier et al. (2012) argue that internally driven factors that induce adverse listening conditions lead to the permanent use of gestural information during comprehension. For non-native listeners however, this potential strategy change of taking into account additional (visual) information sources may be more dynamically adapting to the listening situation and native listener status, and therefore have less of a permanent character.

For example, in the multimodal domain, studies have suggested that in clear speech, non-native listeners might benefit more from the semantic information that is conveyed by gestures than native listeners, as it may compensate for their poorer proficiency of the language (Dahl & Ludvigsen, 2014; Sueyoshi & Hardison, 2005). Iconic gestures have also been suggested to improve language learning for non-native listeners, with words being remembered better when instructed or taught with an accompanying gesture (Kelly, McDevitt, & Esch, 2009). Other

studies on tone and pitch perception have demonstrated that both facial (i.e., visible speech) and gestural input can enhance non-native tone perception, especially in noise (Hannah et al., 2017).

In unimodal contexts however, previous work on semantic information uptake in sentence contexts has suggested that non-native listeners only can utilize auditory semantic information to resolve information loss at the phoneme level when the speech signal is of sufficient quality (Bradlow & Alexander, 2007; Bradlow & Bent, 2002; Gat & Keith, 1978; Mayo, Florentine, & Buus, 1997). This has however not been studied in a multimodal semantic context.

Taken together, these results from the unimodal and multimodal domain thus sketch two different hypotheses on how non-native listeners, who are thus challenged by internally driven adverse listening condition during language comprehension, might benefit from the semantic information of gestures in externally driven adverse listening conditions. Non-native listeners might not be able to benefit from the semantic information conveyed by gestures when sound quality is insufficient (following Bradlow & Alexander, 2007; Bradlow & Bent, 2002; Gat & Keith, 1978; Mayo, Florentine, & Buus, 1997) or might benefit more to compensate for their poorer proficiency (Dahl & Ludvigsen, 2014; Sueyoshi & Hardison, 2005). Moreover, both of these hypotheses point back to the possible differences that might exist between the two types of adverse listening conditions. In an internally induced adverse listening condition, such as when you are a non-native listener, your prior knowledge of a language is diminished, but in an externally induced adverse listening condition, your prior knowledge is intact. This difference might affect the weighing and integration of the auditory and visual inputs when processing language depending on the type of adverse listening condition.

1.4. Outline of this thesis

The main goal of this thesis is to uncover the cognitive and neural mechanisms underlying gestural enhancement of language comprehension in externally (i.e., speech degradation) and internally (e.g., non-nativeness) induced adverse listening conditions. Specifically, I aim to investigate this by studying whether

and how oscillatory brain dynamics underlie gestural enhancement of degraded speech comprehension in both native and non-native listeners. Using oscillations as a means to study speech-gesture integration will reveal how areas involved in language comprehension connect, temporally and spatially, to other domains of cognition, such as vision, motor, action and attention. I therefore aim to uncover how the brain performs this multimodal integration of speech and gestures and how the brain engages areas that are relevant for this process over time. Speech-gesture integration will be probed by using paradigms that focus on gestural enhancement of speech, as well as violation paradigms. The main hypotheses of this thesis are that gestures enhance speech comprehension in both externally and internally induced adverse listening conditions and that low-frequency decreases of oscillatory power in the alpha and beta band reflect the engagement of task-relevant brain regions in this process (following Jensen & Mazaheri, 2010; Klimesch, Sauseng, & Hanslmayr, 2007; Payne & Sekuler, 2014; Pfurtscheller & Lopes da Silva, 1999). This mechanism is expected to be general for all listening conditions. In this thesis, I will present results from eight studies, which were all designed to examine these questions. **Chapters 2-4 & 9** will use behavioral (**chapter 2**) and neural measures (**chapters 3, 4 & 9**) to investigate speech-gesture integration in externally induced adverse listening conditions. **Chapters 5-8** will use behavioral (**chapter 5**), neural (**chapters 6 & 7**), and eye-tracking measures (**chapter 8**) to investigate speech-gesture integration in internally and externally induced adverse listening conditions.

As a first step, answering these questions requires a thorough behavioral investigation of whether and to what extent iconic gestures contribute to information from visible speech to enhance degraded speech comprehension at different levels of noise-vocoding. As face-to-face communication involves both visible speech and gestures, and gestures always occur on top of visible speech, **chapter 2** will investigate how these visual articulators, in a joint and natural multimodal context, interact with each other during clear and degraded speech comprehension, and whether this is dependent on the severity of speech degradation. Currently, it is unknown how these gestures enhance speech comprehension in a context of visible speech information.

I used the results of this experiment as input for an MEG experiment described in **chapter 3**, where I aimed to investigate how the brain supports this gestural enhancement of degraded speech comprehension over time. Here, I studied modulations of neuronal oscillations in different frequency bands and examined whether these modulations could predict how much a listener could benefit from gestures during degraded speech comprehension.

In **chapter 4**, I used the same MEG dataset as reported in **chapter 3** again in a violation paradigm to probe speech-gesture integration in a different manner. This allowed testing the generality of the oscillatory mechanisms that were observed in **chapter 3**.

The next step was then to compare the observed results in externally induced adverse listening conditions to adverse listening conditions induced by internal factors, such as when you are a non-native listener. This is highly relevant as non-native listeners might process or integrate visual semantic information with the speech signal differently than native listeners, and this directly tests whether the mechanisms that were observed in externally induced listening situations are domain-general or specific to only externally induced adverse listening conditions. In **chapters 5-8** I thus switch the focus from studying only externally induced adverse listening conditions to studying both externally and internally induced adverse listening conditions in a joint context. In **chapter 5**, a similar experiment as described in **chapter 2** is conducted with highly-proficient non-native listeners of Dutch, and the results are directly compared to the results of **chapter 2**.

In **chapter 6** I used EEG to investigate the neural correlates that underlie speech-gesture integration in clear and degraded speech, and whether this differed for native and non-native listeners. Here, I studied modulations of the N400 component to assess differences in semantic unification operations.

In **chapter 7**, we followed up on these results by using MEG to study modulations of neuronal oscillations during gestural enhancement of degraded speech comprehension in non-native listeners. This allowed us to study the generality of the oscillatory mechanisms that were observed in **chapter 3** and **4,** by making a direct comparison of the observed modulations to the modulations

observed in native listeners. Additionally, I again aimed to investigate whether these oscillatory modulations were predictive of how much a listener benefits from gestural information during language comprehension in adverse listening conditions.

In **chapter 8** I used eye-tracking to investigate how native and non-native listeners allocate visual attention to gestural information during clear and degraded speech comprehension, and whether gaze allocation to gestures could predict degraded speech comprehension. Using eye-tracking complements the research done in the previous chapters, as it provides a direct online measure of overt visual attention.

The final experimental work described in **chapter 9** is exploratory work that tests a novel method that can be used to study speech-gesture integration. Here, I used rapid invisible frequency tagging (see paragraph 1.5.7.) to study the integration of gestural information in clear and degraded speech. Frequency-tagging periodically modulates a stimulus at a specific frequency, for example by modulating the luminance of a visual stimulus or the amplitude of an auditory stimulus. When such an auditory input (f1) and a visual input (f2) interact, this could result in so-called "intermodulation frequencies" (f2+f1/f2-f1) (Regan, He, & Regan, 1995). Intermodulation frequencies are thought to reflect neural activity that non-linearly combines the two inputs (Zemon & Ratliff, 1984). I aim to provide a proof-of-principle of the use of this method to study the integration of audiovisual inputs in a semantic context.

In **chapter 10** I aim to conclude this thesis with a broader discussion and summary of all findings within a wider context, by comparing speech-gesture integration in both types of adverse listening conditions as well as discuss implications for current debates on the (neural) integration of speech and gestures. Finally, I will propose future investigations to further uncover how a listener behaviorally and neurally combines auditory and visual semantic information in adverse listening conditions to enhance comprehension.

1.5. Methods used in this thesis

As has become clear from the discussion of the existing literature and the outline above, different methods are needed to answer how gestures enhance language comprehension in adverse listening conditions, as well as to identify how different brain areas that are relevant for speech-gesture integration are engaged in this process over time. I will use behavioral methods (**chapter 2** and **5**), EEG (**chapter 6**), MEG (**chapters 3, 4, 7, 9**) and eye-tracking (**chapter 8**) to study this phenomenon. Before I will give a background on these methods, I will first introduce the participant groups, the stimuli types, and tasks that are used throughout the thesis. Note that more detailed information on the methods can be found in the methods sections of **chapters 2-9**.

1.5.1. Participants

As mentioned above, in **chapters 2-4** and **9** only native listeners of Dutch were recruited as participants, as only externally induced adverse listening conditions were studied. Note that **chapters 3, 4** and **7** report the data of the same native Dutch participants. In **chapters 5-8**, where both externally and internally adverse listening conditions were studied, both native and highly-proficient non-native listeners were recruited. All of these non-native listeners were highly-proficient German speakers of Dutch. Highly-proficient listeners were recruited because the non-native listeners had to have sufficient vocabulary knowledge to understand the verbs used in the videos (see below). Low proficient participants would not be able to recognize all the verbs in the videos, and would possibly only be able to pick up on the semantic information that is conveyed by the gestures. This would not be sufficient to study speech-gesture integration, or gestural enhancement of (degraded) speech comprehension. Note that in **chapters 5-8** we used different groups of non-native listeners per study.

The Dutch proficiency of all non-native listeners was always assessed by the Dutch version of the Lexical Test for Advanced learners of English (LexTALE), a vocabulary test using non-speeded visual lexical decision (Lemhöfer & Broersma, 2012). In all relevant chapters, only non-native listeners with a proficiency level of

60% or higher were allowed to participate in the experiment. This level corresponds to a B2 level or higher (i.e., upper intermediate). Although the knowledge of the verbs that we used in the stimuli was thoroughly pre-tested for non-native listeners, we always also included an adapted version of the LexTALE after all experiments reported in **chapters 5 - 8** to assess each non-native listener's knowledge of the specific verbs that we used in the respective experiment. Similarly, the iconicity of all gestures was thoroughly pretested to ensure the non-native listeners were able to recognize the gestures we used in the videos (see **chapter 5** for more details on these pre-tests).

1.5.2. Stimulus materials: videos

In all chapters, we used the same video materials as stimuli. These videos were recorded as 240 short video clips of a female, native Dutch non-professional actress who uttered Dutch action verbs. All verbs that were used in these videos were highly frequent action verbs, which ensured that they were easily coupled with iconic gestures. In 80 of the 240 videos, the actress uttered these action verbs without a gesture. In the other 160 videos, the actress either uttered the verb accompanied by a matching, congruent gesture (80) or a mismatching, incongruent gesture (80). The mismatching gestures were created by providing the actress with two verbs, and instructing the actress to pronounce one verb, and perform the other verb in her gestures. This was done because the face and lip movements of the actress were included in the videos, and we did not want to create a visible speech mismatch. The actress was not instructed on how to perform the gestures. All gestures were performed by the actress on the fly, while she was standing in front of a neutrally colored background, wearing neutrally colored clothes (see Figure 2).

All videos were on average 2000 ms long, and the preparation of the gesture (i.e., the first frame that showed movement of the hand), started at 120 ms. The stroke, (i.e., the meaningful part of the gesture), started on average at 550 ms, followed by speech onset at on average 680 ms, gesture retraction at on average 1380 ms, and gesture offset at on average 1780 ms. The temporal lag between the stroke (550 ms) and the gesture (680 ms) has been found to be ideal for information from

the two inputs to be integrated during comprehension, and allowed for mutual enhancement of the two inputs for comprehension (as demonstrated by Habets et al., 2011). All videos were extensively pre-tested to ensure they were potentially ambiguous in the absence of speech, to ensure they were equally iconic, and recognizable for both native and non-native speakers. Details on these pretests can be found in **chapter 2** and **chapter 5**, and a full list of all the verbs that were used is included in Appendix I.

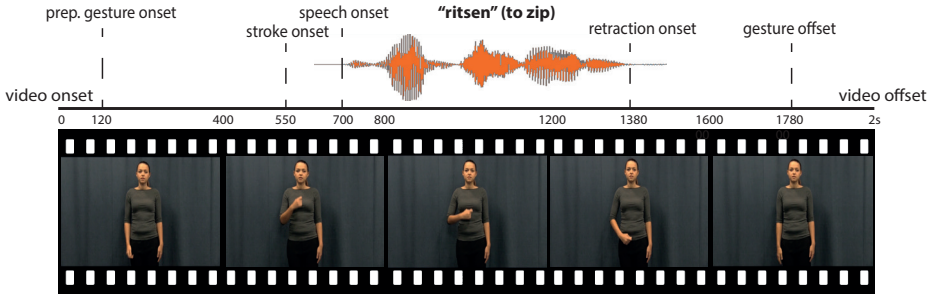


Figure 2: Timeline of the videos used in this thesis. Under each dashed line the average timing of the specific event mentioned above the dashed line is depicted.

1.5.3. Stimulus materials: speech degradation

The sound in the videos was manipulated by using noise-vocoding to induce externally induced adverse listening conditions. This is a form of speech degradation that degrades the spectral detail that is present in a speech signal, while preserving much of the slowly varying temporal cues. Noise-vocoded speech is created by dividing the speech signal into separate frequency bands, and using the amplitude envelope from each band to modulate bands of noise that are centered over the same frequency bands. These bands of noise are then recombined to create a noise-vocoded version of the sound (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). The number of bands that are used in creating the noise-vocoded signal affects the intelligibility of the signal: the more bands are used in the vocoder, the more intelligible the sound (Davis & Johnsrude, 2003). I pretested what the optimal level of speech degradation was for native and non-native listeners to benefit most from the semantic information that is conveyed

by the gesture. Details on these pretests can be found in **chapter 2** and **chapter 5**.

1.5.4. Behavioral methods

1.5.4.1. Types of paradigms to target speech-gesture integration

In all experiments, I used similar behavioral methods to probe speech-gesture integration. Specifically, I either used an 'enhancement paradigm' (**chapters 2, 3, 5, 7 & 8**), where I focused on the comparisons between conditions that contained a gesture or no gesture, or a 'violation paradigm' (**chapters 4, 6 & 9**), where I focused on the comparisons between conditions that contained a gesture that semantically matched or mismatched the speech signal. Both paradigms target speech-gesture integration, and using both paradigms allows to test whether effects are general for speech-gesture integration or specific for a certain type of paradigm.

1.5.4.2. Types of tasks to test speech-gesture integration

All experiments used a form of a recall task to see whether the participants had behaviorally resolved the speech and gestures. In **chapter 2** and **chapter 5** we used an open-set identification task, where participants had to type what verb they thought the actress tried to convey. However, we used a 4-alternative forced choice identification task in the other chapters to avoid motion artifacts in our M/EEG data. Moreover, using the 4-alternative forced choice identification task allowed including semantic and phonological competitors to test whether participants would focus only on gestural information or speech information when speech was degraded. This could be informative about the processing strategies of listeners during comprehension. We have included all possible items and distractors in Appendix I.

1.5.5. Magneto- and Electroencephalography

In **chapters 3, 4, 6, 7 and 9**, M/EEG were used as a method to record brain activity during the online comprehension of speech and gestures. MEG and EEG record brain activity with a temporal resolution in the order of milliseconds (Vrba & Robinson, 2001). This can be measured either by placing electrodes on the scalp

(EEG), or by placing sensors close by the scalp that are sensitive to magnetic fields (MEG). The excellent temporal resolution of M/EEG makes these methods ideal candidates to study the detailed time course of speech-gesture integration.

Both MEG and EEG are used to measure neuronal firing, but the M/EEG signal does not directly reflect spiking activity, but a smoothened version of the local field potential (Buzsáki, Anastassiou, & Koch, 2012). In more detail, M/EEG can be used to measure the brain activity resulting from the synchronization of activity from columns of pyramidal neurons. The electrochemical gradient that exists across the cell membrane of these neurons can produce a difference in voltage between the inside and outside of a cell, which is called a membrane potential. When this membrane potential depolarizes, an action potential is generated at the presynaptic neuron, which results in the release of neurotransmitters at the synaptic cleft. These released neurotransmitters couple to receptors on the dendrites of a postsynaptic neuron, which leads to a change in the membrane potential, and, in turn, a postsynaptic potential.

When this postsynaptic potential occurs at apical dendrites, current flows to the soma of a neuron. This is known as the intracellular or primary current, which, in turn, can generate an extracellular, or secondary current which is deflected by the surrounding tissue. Each of these currents produces a magnetic field perpendicular to its current. EEG measures the electrical potential arising from the secondary current, whereas MEG measures both the magnetic field generated by both primary and secondary current (see Figure 3).

Thousands of these neurons need to receive synchronized synaptic input for the signal to be measurable by MEG and EEG. MEG is mostly sensitive to currents in the walls of sulci of the brain, due to the spatial alignment and tangential orientation to the cortical surface, causing the primary and secondary current to not cancel out. EEG however is sensitive to sources that are oriented both radially and tangentially to the scalp.

1.5.5.1.Measuring evoked activity - event-related potentials

In **chapter 6**, ERPs derived from the EEG signal are used as a means to investigate

speech-gesture integration. ERPs result from the analysis of evoked activity that is phase- and time-locked across trials. As stated above, ERPs can be described as electrical potential changes that are time-locked to a certain external event. These electrical potential changes can either occur as 'peaks' or 'troughs' relative to a certain stimulus at a certain point in time. Some of these characteristic deflections of the signal have been called 'ERP components'. In **chapter 6**, modulations of the N400 component, a component that is sensitive to semantic unification operations, are studied to assess differences in speech-gesture integration in externally and internally induced adverse listening conditions.

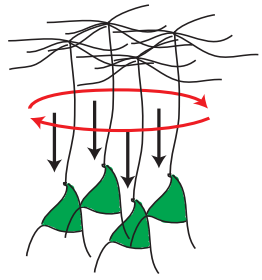


Figure 3. Synchronized post-synaptic potentials of aligned pyramidal neurons will generate a measurable electric current (black, primary current) and magnetic field (red line, secondary current).

1.5.5.2. Measuring induced activity - neuronal oscillations

In **chapters 3, 4, 7 and 9**, neuronal oscillations derived from the MEG signal are used as a means to investigate speech-gesture integration. Oscillations reflect patterns of rhythmic brain activity that are not phase-locked across trials. Oscillations consist of a certain frequency (measured in Hertz, which denotes the number of cycles per second), amplitude (the amount of energy, or power (in amplitude squared)), and phase, which is the location in an oscillatory cycle (in degrees) relative to the origin of the oscillation.

In **chapters 3, 4, 7 and 9**, I used MEG to investigate power modulations in the oscillatory signal to quantify synchronous activity of neurons within task-relevant brain regions. These power modulations were calculated using time-frequency analysis. To achieve this, I applied a Fourier transform to segments of the data using a moving time-window. Oscillatory power and phase can then be estimated for each of these data segments, and averaged across trials. In this thesis I solely

focused on analyzing the power of oscillations. These power modulations are not necessarily phase-locked to the stimulus.

Throughout this thesis, these power modulations will also be referred to as event-related desynchronization, which is a relative decrease in power, and event-related synchronization, which is a relative increase in power (see Figure 4). Low-frequency (i.e., alpha (8-12 Hz) and beta (13-30)) desynchronization has often been associated with the engagement of task-relevant brain regions, whereas low-frequency synchronization is thought to reflect functional inhibition or disengagement (Jensen & Mazaheri, 2010; Klimesch et al., 2007; Payne & Sekuler, 2014; Pfurtscheller & Lopes da Silva, 1999). High-frequency (i.e., gamma (30-100+ Hz) synchronization has been proposed to reflect enhanced neuronal computation (Fries, Nikolić, & Singer, 2007; Jokisch & Jensen, 2007). To extract these power modulations, I used time-frequency analysis to estimate power for each data segment and averaged these values across trials.

I localized the observed effects in the MEG sensor data to investigate the potential spatial sources of these effects. To do this, one must consider that the observed MEG signal can be seen as a superposition of magnetic fields caused by multiple sources across the brain. Mathematical models can be used to estimate the potential loci of the observed MEG sensor data. To achieve this, these models need to be coupled to the anatomical brain structures of a participant, which were collected by making structural MRI scans of all the participants. In **chapters 3, 4, 7 and 9**, we used a beamforming algorithm known as 'Dynamic Imaging of Coherent Sources' (DICS) to reconstruct the sources of our MEG data (Gross et al., 2001). This algorithm takes into account how much the data from each sensor relates to the data from other sensors and assumes that there is no correlation between the time courses of source activity. It also takes into account a model of the head that contains the conductivity values of the different tissue types in the brain, in the form of a three-dimensional grid. For each of those grid points, a spatial filter is computed to estimate the activity at that grid point location. The observed activity at those locations forms a weighted and linear combination of all sensor signals, which follows a unit-gain constraint to reduce the variance of the activity at each location (Gross et al., 2001).

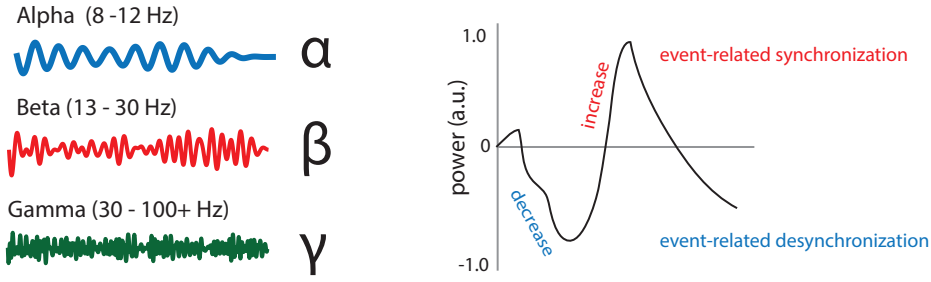


Figure 4. Left: overview of the oscillatory frequency bands studied in this thesis. Throughout this thesis, alpha and beta oscillations are also called 'low frequencies', and gamma oscillations 'high frequencies'. Right: Illustration of power increase / event-related synchronization and power decrease / event-related desynchronization.

1.5.6. Eye-tracking

The data of **chapters 3, 4, 6** and **7** hinted that when listeners observed gestures in both externally and internally induced adverse listening conditions, more visual attention might be allocated to these gestures to aid comprehension. However, although previous literature suggested that listeners rarely gaze at gestures when speech is clear (e.g., Gullberg & Holmqvist, 2002, 2006; Gullberg & Kita, 2009), it is unclear whether listeners allocate overt visual attention to gestures to aid comprehension under adverse listening conditions. We explored this question by using eye-tracking, to see whether visual attention to gestures could predict language comprehension under adverse listening conditions. Eye-tracking provides an excellent method to investigate online how externally and internally induced adverse listening conditions impact visual information uptake during language comprehension, as well as how attention allocation is reflected in gaze behavior (McQueen & Huettig, 2012; Van Engen & McLaughlin, 2018). Next to the behavioral and neuroimaging methods used in this thesis to investigate the effects of adverse listening conditions on comprehension, eye-tracking can thus provide an objective physiological measure of the online processing of speech and gestures.

In **chapter 8**, we presented participants with continuous videos of a speaker to and measured gaze allocation to gestures in both externally and internally induced

adverse listening conditions. Here, we used novel analysis methods to investigate the specific time course of gaze allocation under these conditions, instead of more common methods that divide the eye-tracking signal in predefined time bins. This allowed us to zoom in to the exact timing of our effects, and relate those effects to the video structure and data from previous chapters.

1.5.7. Speech-gesture integration studied by rapid invisible frequency tagging

In **chapter 9**, we used rapid invisible frequency tagging (RIFT) to study the interaction between auditory and visual information during speech-gesture integration. Frequency tagging is a method where stimuli are periodically modulated at a specific frequency (for example, at 6 Hz, which introduces a flickering stimulus). This rhythmic sensory stimulation can be introduced in an auditory manner, by using amplitude modulation, or in a visual manner, which induces visual 'flickering'. This 'flickering' causes a brain response called the steady-state visually evoked response (SSVEPs, for EEG, or fields, SSVEFs, for MEG, or in the case of auditory amplitude modulation: auditory steady state response (ASSR)) with high power at the frequency-tagged stimulus (Vialatte, Maurice, Dauwels, & Cichocki, 2010). The amplitude of these responses tends to be higher for attended than for unattended stimuli (Gulbinaite, van Viegen, Wieling, Cohen, & VanRullen, 2017; Müller et al., 2006; Muller et al., 1998; Vialatte et al., 2010) and has been used to study audiovisual integration of non-semantic stimuli (e.g., tones and gratings, Giani et al., 2012).

Using frequency-tagging to study audiovisual integration allows tagging an auditory and visual stimulus at two different frequencies (f_1 and f_2) to study how these two inputs interact in the brain. Here, the idea is that when the auditory and visual signals interact, this would result in a non-linear interaction of the two signals. This then would be reflected in power at the intermodulation frequencies of f_1 and f_2 (f_2-f_1 or f_2+f_1) (Regan & Regan, 1989). Using this method thus allows to test for non-linear neural convergence of two inputs (Norcia, Appelbaum, Ales, Cottereau, & Rossion, 2015; Regan & Regan, 1988; Vialatte et al., 2010; Zemon & Ratliff, 1984).

Recent work has failed to find evidence for the existence of non-linear interactions across modalities (Giani et al., 2012; but see Regan, He, & Regan, 1995). However, previous work has only used low-frequency tagging to uncover these intermodulation frequencies. This could be problematic when considering that the frequency-tagging in itself is likely to entrain spontaneous neural oscillations at lower frequencies (e.g., alpha/beta oscillations) (Keitel, Quigley, & Ruhnau, 2014; Spaak, de Lange, & Jensen, 2014). Previous work has provided proof-of-principle for using frequency tagging with complex input at high frequencies ('rapid invisible frequency tagging'), to study the propagation of information from early visual to higher order downstream brain areas. In **chapter 9** we use rapid invisible frequency tagging to investigate the integration of speech and gestures. Second, following Herring (2017), we aim to provide proof-of-principle that rapid invisible frequency tagging can be used as a tool to study the interaction between auditory and visual inputs in complex dynamic settings, such as multimodal language comprehension.

Chapter 2

The joint contribution of iconic gestures and visible speech to degraded speech comprehension in native listeners

2.1. Abstract

This study investigated whether and to what extent iconic co-speech gestures contribute to information from visible speech to enhance degraded speech comprehension at different levels of noise-vocoding. Previously, the contributions of these two visual articulators to speech comprehension have only been studied separately. Twenty participants watched videos of an actress uttering an action verb and completed an open-set identification task. The videos were presented in three speech (2-band; 6-band noise-vocoding; clear), three multimodal (Speech+Lips blurred; Speech+VisibleSpeech; Speech+VisibleSpeech+Gesture) and two visual only conditions (VisibleSpeech; VisibleSpeech+Gesture). Accuracy levels were higher when both visual articulators were present compared to one or none. The enhancement effects of a) visible speech, b) gestural information on top of visible speech and c) both visible speech and iconic gestures were larger in 6-band than 2-band noise-vocoding or visual only conditions. Gestural enhancement in 2-band noise-vocoding did not differ from gestural enhancement in visual only conditions. When perceiving degraded speech in a visual context, listeners benefit more from having both visual articulators present compared to one. This benefit was larger at 6-band than 2-band noise-vocoding, where listeners can benefit from both phonological cues from visible speech, and semantic cues from iconic gestures to disambiguate speech.

This chapter is based on Drijvers, L., & Ozyurek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research,* 60, 212-222. doi:10.1044/2016_JSLHR-H-16-0101.

2.2. Introduction

Natural, face-to-face communication often involves an audiovisual binding that integrates information from multiple inputs such as speech, visible speech, and iconic co-speech gestures. Notably, the relationship between these two visual articulators and the speech signal seems to differ: Iconic gestures, which can be described as hand movements that illustrate object attributes, actions and space (e.g., Clark, 1996; Goldin-Meadow, 2005; McNeill, 1992), are related to speech on a semantic level, due to the similarities to the objects, events and spatial relations they represent. In contrast, the relation between visible speech, consisting of lip movements, tongue movements and teeth, and speech consists of a form-to-form mapping between syllables and visible speech on a phonological level. Previous research has argued that both iconic gestures and visible speech can enhance speech comprehension, especially in adverse listening conditions, such as degraded speech (Holle, Obleser, Rueschemeyer, & Gunter, 2010; Obermeier, Dolk, & Gunter, 2012; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollock, 1954). However, the contribution of iconic gestures and visual speech to audiovisual enhancement of speech in adverse listening conditions has been mostly studied separately. Since natural, face-to-face communication involves gestures and visual speech as possible visual articulators, this raises the question of whether, to what extent and how the co-occurrence of these two visual articulators influences speech comprehension in adverse listening conditions. To this end, the current study aims to investigate the contribution of both types of visual information to degraded speech comprehension in a joint context.

Iconic gestures are frequently prevalent in natural, face-to-face communication and have both a temporal and semantic relation with the speech they occur with, causing them to be hard to disambiguate without speech. It has been theorized that iconic gestures are an integral part of language (Kendon, 2004; McNeill, 1992): Speech and iconic gestures are integrated continuously during comprehension, and target linguistic processing on semantic, syntactic and pragmatic levels (Holle et al., 2012; Kelly, Ozyürek, & Maris, 2010; McNeill, 1992, see for a review and meta-analysis: Hostetter, 2011). Previous research has shown that semantic information from iconic gestures is indeed processed by listeners

and that iconic gestures can impact language comprehension, at behavioral and neural levels (e.g. Beattie & Shovelton, 1999; 2002; Holle & Gunter, 2007; Holler et al., 2014; Holler, Kelly, Hagoort, & Ozyurek, 2010; Holler, Shovelton, & Beattie, 2009; Kelly, Barr, Church, & Lynch, 1999; Kelly, Healey, Özyürek, & Holler, 2015; Obermeier, Holle, & Gunter, 2011 see for a review, Özyürek, 2014). For example, in an EEG study, Holle & Gunter (2007) showed participants videos of an actor who uttered a sentence while gesturing. Here, the experimental sentences contained an unbalanced homonym in the first part of the sentence (e.g. 'She controlled the ball'). This homonym was disambiguated in the subsequent clause (e.g. 'which during the game'/'which during the dance'). When the actor uttered the homonym, he would simultaneously produce an iconic gesture that either depicted the dominant ('game') or the subordinate meaning ('dance') of the homonym. When the gesture was congruent, they found a smaller N400 as compared to an incongruent gesture. This suggests that listeners use the semantic information from gestures to disambiguate speech.

So far, it has been argued that in adverse listening conditions, gestures occur more frequently (Hoskin & Herman, 2001; Kendon, 2004) and that listeners take gestures more into account than in clear speech (Rogers, 1978). This was also found by Obermeier et al., (2011), who used a similar paradigm as Holle & Gunter (2007), to reveal that when there was no temporal overlap of a word and a gesture and participants were not explicitly asked to attend to the gestures, speech-gesture integration did not occur. However, in a subsequent study where the same stimuli were presented in multi-talker babble noise, listeners did incorporate the gestural information with the speech signal to disambiguate the meaning of the sentence. This effect was also found for hearing-impaired individuals (Obermeier et al., 2012). These results underline that speech-gesture integration can be modulated by specific characteristics of the communicative situation.

Another fMRI study by Holle, Obleser, Rueschemeyer & Gunter (2010) investigated the integration of iconic gestures and speech by manipulating the signal-to-noise ratio (SNR) of the speech to target areas that were sensitive to bimodal enhancement and inverse effectiveness (i.e. greater bimodal enhancement for unimodally least effective stimuli, i.e. the noisiest speech level). Participants

watched videos of an actor with a covered face, who uttered short sentences (e.g. 'And now I grate the cheese') with or without an accompanying iconic co-speech gesture. These videos were presented with speech in a good SNR (+2 dB) or in a moderate SNR (-6 dB), using multitalker babble tracks. Their results revealed that the superior temporal sulcus and superior temporal gyrus in both hemispheres were sensitive to bimodal enhancement, and the neural enhancement for bimodal enhancement was even larger when participants were processing the speech and gestures in the degraded speech conditions. On both a neural as a behavioral level (i.e. response accuracy), this study showed that attending to a gesture under adverse listening conditions can significantly enhance speech comprehension and help in the disambiguation of a speech signal that is difficult to interpret. This gestural enhancement had already been described by Rogers (1978), who manipulated noise levels to show that gestures could only benefit speech comprehension when sufficient noise was added to the speech signal.

As Holle and colleagues (2010) note however, their study (and other studies, see e.g. Obermeier et al., 2011; Obermeier et al., 2012) have only focused on one visual articulator in speech-related audiovisual integration, namely iconic gestures. Other visual articulators, such as lip movements, were deliberately excluded from the stimuli that were used by blocking the actor's face with a black mask. Yet, these lip movements are inherently part of natural, face-to-face communication: Lip movements can provide temporal information about the speech signal (e.g. on the amplitude envelope) and information on the spatial location of a speaker's articulators (e.g. place and manner of articulation), which can be specifically useful when perceiving speech in adverse listening conditions. Additionally, lip movements can convey phonological information, because of the form-form relationship between lip movements and syllables or segments that are present in the speech stream (for a recent review see Peelle & Sommers, 2015).

The enhancement effect visible speech (consisting of lip movements, tongue movements and information from teeth) has on speech in clear and adverse listening conditions, has been reported by several studies (e.g. Erber, 1969, 1971; Ma, Zhou, Ross, Foxe, & Parra, 2009; Ross et al., 2007; Schwartz et al., 2004; Sumby & Pollock, 1954). Recognizing speech in noise is easier when a visual cue is present

than when auditory information is presented alone, and has shown to improve recognition accuracy (Tye-Murray, Sommers, & Spehar, 2007). Previously, studies have argued that this beneficial effect increases as the SNR decreases (Sumby & Pollack 1954; Erber, 1969; 1975, Callan et al., 2003). However, more recent studies have reported that visual enhancement of speech by lip movements seems to be largest at "intermediate" SNR's where the auditory input is at a level between "perfectly audible" and "completely unintelligible" (Ross et al., 2007; 2008, Ma et al., 2009). This has also been reported by Holle et al., (2010), for gestural enhancement of speech in noise. Nevertheless, most studies on lip movements as a visual enhancement of speech have used stimuli that only showed the lips or lower half of the face (e.g. Callan et al., 2003; Ross et al., 2007; Schwartz, Berthommier, & Savariaux, 2004) to eliminate influences from the rest of the face or body. This is similar to studies in the domain of gestural enhancement of speech in noise, where most studies block the face of the speaker, the mouth, or just show the torso of the speaker, to eliminate influences from visible speech (e.g. Obermeier et al., 2011; Obermeier et al., 2012; Holle et al., 2010).

Although there has not been a study that investigated the contribution of visible speech and iconic gestures on speech comprehension in adverse listening conditions, a few studies used both visual articulators in their stimuli. In a fMRI study, Skipper, Goldin-Meadow, Nusbaum, & Small (2009) showed that when clear speech was accompanied by meaningful gestures, there was strong functional connectivity between motor planning and production areas and areas that are thought to mediate semantic aspects of language comprehension. This suggests that the motor system works together with language areas to determine the meaning of those gestures. When just facial information (incl. visual speech) was present, there were strong connectivity patterns between motor planning and production areas and areas that are thought to be involved in phonological processing of speech. These results suggest that information from visible speech is integrated with phonological information, whereas meaningful gestures target semantic aspects of language comprehension. However, it remains unknown how these two articulators interact when both are able to enhance language comprehension in adverse listening conditions.

Two other studies by Kelly et al., (2008) and Hirata & Kelly (2010) examined the effects of lip movements and iconic gestures on auditory learning of second language speech sounds (i.e. prosody and segmental phonology of Japanese). They hypothesized that having both modalities present would benefit learning the most, but found that only lip movements resulted in greater learning. They explain their results by stating that hand gestures might not be suited to learn lower-level acoustic information, such as phoneme contrasts. Again, this study underlines the different relations of visible speech and iconic gestures to speech: visible speech can convey phonological information that can be mapped to the speech signal, whereas gestural information conveys semantic information. It remains unknown how these visual articulators interact when both can enhance language comprehension, such as when speech is degraded.

2.2.1. The present study

The current study aims to investigate the enhancement effect of iconic gestures and visible speech on degraded speech comprehension, by studying these visual articulators in a joint context. Specifically, we ask what gestural information adds on top of the enhancement of visible speech on degraded speech comprehension, and we test the hypothesis whether the occurrence of two visual articulators (i.e. Speech+VisibleSpeech+Gesture) enhances degraded speech comprehension more than having only visible speech (i.e. Speech + VisibleSpeech) present, or having no visual articulators present (i.e. Speech+Lips blurred). As iconic gestures convey semantic cues that could add to degraded speech comprehension and visible speech conveys phonological cues that could add to degraded speech comprehension, we expect iconic gestures to have an additional enhancement effect on top of the enhancement effect from visible speech.

We hypothesize that the enhancement from visible speech compared to speech alone (i.e. VisualSpeech enhancement: Speech+VisibleSpeech compared to Speech+Lips blurred) will be larger at an intermediate level of degradation compared to a severe level of degradation, allowing a listener to map the phonological information from visible speech to the speech signal. Additionally, we expect the enhancement from iconic gestures on top of visible speech (i.e.

Gestural enhancement: Speech+VisibleSpeech+Gesture - Speech+VisibleSpeech) to be largest at an intermediate level of degradation compared to a severe level of degradation, which would indicate that a listener can benefit more from the semantic information from iconic gestures when there are more clear auditory cues to map this information too. Lastly, we predict that the enhancement of both articulators combined (i.e. Double enhancement: Speech+VisibleSpeech+Gesture compared to Speech+Lips blurred), to be largest at an intermediate level of degradation compared to severe degradation. Since iconic gestures occur on top of information from visible speech, we expect that that should only be possible when enough auditory cues are available to the listeners. This way, listeners can benefit from both phonological information that is conveyed by visible speech, and from semantic information that is conveyed by iconic gestures.

Based on previous results on gestural enhancement of degraded speech comprehension (Holle et al., 2010, with no information from visible speech present) and enhancement of visible speech (e.g. Ross et al., 2007, with no information from iconic gestures present), we hypothesize that for double enhancement from both iconic gestures and visible speech we find a similar moderate range for optimal integration where our language system is weighted to an equal reliance on auditory inputs (speech) and visual inputs (iconic gestures and visible speech).

2.3. Methods

2.3.1. Participants

Twenty right-handed native speakers of Dutch (11 females, M_{age} = 23;2 years, *SD* = 4.84) participated in this experiment. All participants reported no neurological or language-related disorders, no hearing impairments, and had normal or corrected-to-normal vision. None of the participants participated in the pre-test (described below). All participants gave informed written consent before the start of the experiment and received a financial compensation for participation.

2.3.2. Stimulus materials

In the main experiment, we presented participants with 220 short video clips of

a female, native Dutch actress uttering a Dutch action verb. The auditory and visual stimuli consisted of the Dutch high frequent action verbs, to make sure that the verbs could easily be coupled with iconic gestures. All video materials were recorded with a JVC GY-HM100 camcorder. Each recording of an action verb resulted in a video length of 2 seconds with an average speech onset of 680ms after video onset. All videos displayed the female actress from head to knees, appearing in the middle of the screen and wearing neutrally colored clothes (grey and black), in front of a unicolored and neutral background. Upon onset of the recording, the actress' starting position was the same for all videos. She was standing straight, facing the camera, with her arms hanging casually on each side of the body. During recording, she was instructed to utter the action verb while making a hand gesture that she found representative for the verb, without receiving feedback from the experimenter. The gestures she made were not instructed by the experimenter but were created by the actress on the fly. If the actress would have received explicit instructions per gesture, the gestures would have looked unnatural or choreographed, and the conscious effort to make a certain gesture could have drawn the attention to the participants explicitly to the gestures. All gestures that accompanied the action verbs were iconic movements for the actions that the verbs depicted (e.g. a drinking gesture resembling a cup that is raised towards the mouth for the verb 'to drink'). The preparation of all gestures started 120 ms after video onset, and the stroke (the meaning bearing part) of the gestures always coincided with the spoken verb.

The auditory sound files were intensity-scaled to 70 dB and de-noised in *Praat* (Boersma & Weenink, 2015). All sound files were re-combined with their corresponding video files in Adobe Premiere Pro. From each video's clear audio file, we created noise-vocoded degraded versions, using a custom-made script in *Praat*. Noise-vocoding effectively manipulates the spectral or temporal detail while preserving the amplitude envelope of the speech signal (Shannon et al., 1995). This way, the speech signal remains intelligible to a certain extent, depending on the number of vocoding bands, with more bands resulting in a more intelligible speech signal. We bandpass filtered each sound file between 50 Hz and 8000 Hz, and divided the signal into logarithmically spaced frequency bands between 50

and 8000 Hz. This resulted in cutoff frequencies at 50 Hz, 632.5 Hz and 8000 Hz for 2-band noise-vocoding and 50 Hz, 116.5 Hz, 271.4 Hz, 632.5 Hz, 1473.6 Hz, 3433.5 Hz and 8000 Hz for 6-band noise-vocoding. We used the frequencies to filter white noise in order to obtain six noise bands. We extracted the amplitude envelope of each band by using half-wave rectification. We then multiplied the amplitude envelope with the noise bands and recombined the bands to form the distorted signal.

In addition to clear speech, we included 2-band noise-vocoding and 6-band noise-vocoding in our experiment. In total, eleven conditions were created for the experiment (see Figure 5 for an overview). First, nine conditions were created in a 3 (Speech+Lips blurred, Speech+VisibleSpeech, Speech+VisibleSpeech+Gesture) by 3 (2-band noise-vocoding ('severe' degradation), 6-band noise-vocoding ('moderate' degradation), clear speech) design. Second, we added two extra conditions without sound (VisibleSpeech only, which is similar to lip reading, and VisibleSpeech+Gesture) to test how much information participants can resolve from visual input by itself. These conditions did not contain an audio file, so participants only could utilize the visual input. The final experimental set contained 220 videos with 220 distinct verbs that were divided over these eleven conditions (20 per condition) to test the different contributions of visible speech and gestures to clear speech comprehension and in these two degraded listening conditions.

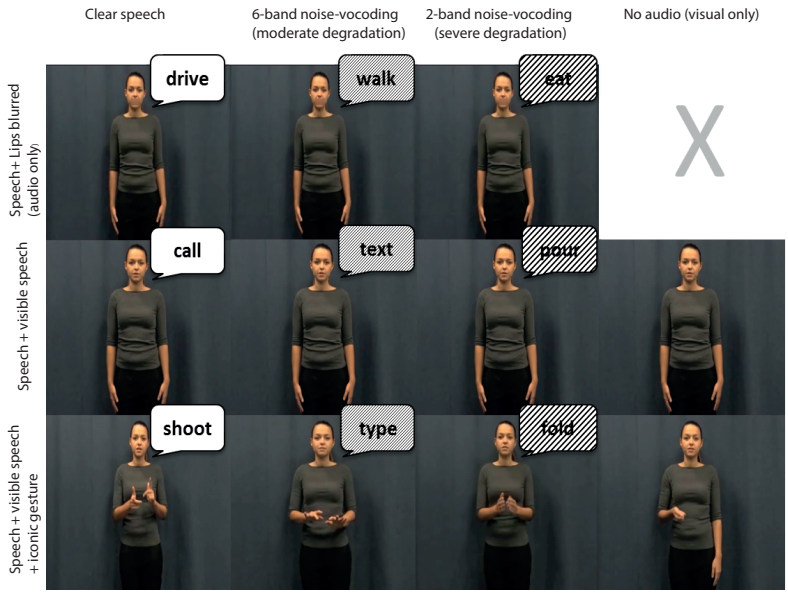


Figure 5: Overview of the design and conditions used in the experiment.

2.3.3. Pre-test

To ensure that the verbs that we chose could be disambiguated by the iconic gestures that we recorded we conducted a pre-test to examine whether the gestures that the actress made in the video indeed depicted the verbs we matched them with in our audio files. This pre-test included videos that were used in all the experiments in the current thesis, and therefore included gestures that mismatched a speech signal, and gestures that matched a speech signal. In this experiment, twenty native Dutch speakers (10 female, M_{age} = 22;2, SD = 3,3) with no motor, neurological, visual, hearing or language impairments and who did not participate in the main experiment, were presented with 170 video stimuli that contained a gesture (only four conditions (speech+visible speech+gesture in clear, 2-band noise-vocoding and 6-band noise-vocoding, and the visual only lips + gesture condition) contained a gesture in this experiment), but without the audio file that contained the verb. All stimuli were presented on a computer screen using Presentation software (Neurobehavioral Systems, Inc.), and presented in a different, randomized order per participant. First, participants were presented with a fixation cross for 1000

ms, after which the video stimulus started playing. After video offset, participants were asked to type down the verbs they associated the movement in the video with. After they filled out the verbs, we showed them the verb we originally matched it with in our auditory stimuli, and asked the participants to indicate on a 7-point-scale (ranging from "does not fit the movement at all" to "fits the movement really well") how iconic they found the movement in the video of the verb that was presented on the screen. This way, we could ensure that in the main experiment, the spoken verbs matched the gesture and participants could use the information from the gestures to disambiguate speech. If the gestures were not a good match with the verb, this gestural information would not enhance speech comprehension. All participants completed the task in approximately 35 minutes and could take self-paced breaks twice during the experiment (after the 55th item, and the 110th item).

The typed answers on the first question of this pre-test ('Which verb do you associate with this video?') were used to determine which verbs had to be renamed to a possibly more occurring synonym, or which verbs were not recognizable and had to be discarded. We coded the answers either as 'correct', when the correct verb or a synonym was given, or as 'incorrect', when the input consisted of an unrelated verb. The results revealed a mean recognition rate of 59% over all gesture videos. The percentage reported here indicates that the gestures are potentially ambiguous in the absence of speech, which is similar to how they are perceived in everyday communication (Krauss et al., 1991). Although this seems like a low overall consistency between participants, one must note that co-speech gestures, such as the iconic co-speech gestures used in these videos, normally occur in the presence of speech, and a higher overall percentage would have indicated that the gestures in our video were more like pantomimes, which are often understood and produced without speech. Since our study aims to understand the possible effects of iconic co-speech gestures on degraded speech comprehension, we did not use pantomimes. Four items that scored especially high (mean recognition rate > 95%) or especially low (mean recognition rate < 15 %) were therefore removed from the final experimental set.

The second question in this pretest targeted the question whether the video

depicted the verb we matched it with in our auditory stimuli. Out of all videos, there were six videos that did not score above a mean rating of '5' on our 7-point scale (ranging from "does not fit the movement at all" (1) to "fits the movement really well" (7), indicating that '5' corresponds to "fits the movement"). These videos had a mean score of 4.79, 4.05, 4.15, 4.94, 4.89 and 4.94) and were not used in this experiment. The mean score on 'iconicity' over the other videos was 6.1 (SD = 0.64). Interestingly, participants indicated after the experiment that when they saw the corresponding verb, they often found that verb (which was often a synonym of their own answer) fitting for the gesture in the video as well, even though it did not always correspond to their own answer. This shows that the mean recognition rate might be negatively biased: even though participants may have filled in a different verb in the first task, they still highly agreed that the gesture in the video corresponded to the verb (as indicated by the score on the second task).

The pre-test resulted in 10 items to be removed from the first set of verbs, resulting in 160 items that contained a gesture. Of these items, 80 items contained speech that matched the semantic meaning of the gesture, and 80 items contained speech that did not match the semantic meaning of the gesture. For the current experiment, we included the 80 verbs that contained a matching speech signal. The other items were used in other experiments described in **chapter 4**, **chapter 6** and **chapter 9**. We added 140 items without a gesture to this set (corresponding to the remaining 7 conditions that did not include a gesture), resulting in 220 videos that were included in the main experiment. Note that all used verbs in all experiments are included in Appendix I.

2.3.4. Procedure

In our main experiment, participants were tested in a dimly-lit soundproof booth, and seated in front of a computer with headphones on. Before the experiment started, the experimenter gave a short verbal instruction that prepared the participant for the different videos that were going to be presented. All stimuli were presented full screen on a 1650x1080 monitor using Presentation software (Neurobehavioral Systems, Inc.), at a 70 cm distance in front of the participant.

A trial started with a fixation cross of 1000, after which the stimulus was played. Then, in an open-set identification task, participants were asked to type which verb they thought the actress tried to convey. After the participants typed in their answers, a new trial began after 500 ms. An answer was coded as 'correct' when a participant wrote down the correct verb, or minor spelling mistakes were made. Synonyms or category-related verbs (e.g. 'to bake' for 'to cook') were counted as incorrect.

All participants were presented with a different pseudo-randomization of the stimuli, with the constraint that a specific condition could not be presented more than twice in a row. The stimuli were presented in blocks of 55 trials, and participants could take a self-paced break in between blocks. All participants completed the tasks within 45 minutes.

2.4. Results

As a first step, we employed a 3 x 3 repeated measures analysis of variance with the factors Visual Articulator (Speech+Lips blurred; Speech+VisibleSpeech; Speech+VisibleSpeech+Gesture) and Noise-Vocoding Level (2-band noise-vocoding; 6-band noise-vocoding; clear speech) to subject the percentage of correct answers to. Note that we excluded the Visual Only conditions from this analysis (where we only tested VisibleSpeech and VisibleSpeech+Gesture, and not VisibleSpeech+Lips blurred, as this would result in a silent movie with no movement), since this would make our analysis unbalanced. As hypothesized, we found a significant main effect of Noise-Vocoding ($F(2,38) = 1569.78$, $p < .001$, $\eta^2 = .96$) indicating that the more the speech signal was noise-vocoded, the less correct answers were given by the participants. We also found a main effect of VisualArticulator ($F(2,38) = 504.284$, $p < .001$, $\eta^2 = .98$) indicating that the more visual articulators were added to the signal, the more correct answers were given. In addition, we found a significant interaction between Noise-Vocoding level and VisualArticulator ($F(4,76) = 194.11$, $p < .001$, $\eta^2 = .91$), which seemed to be driven by the relatively higher amount of correct responses in the 6-band noise-vocoding condition compared to the other speech conditions (see Figure 6 for the percentages of correct responses per condition).

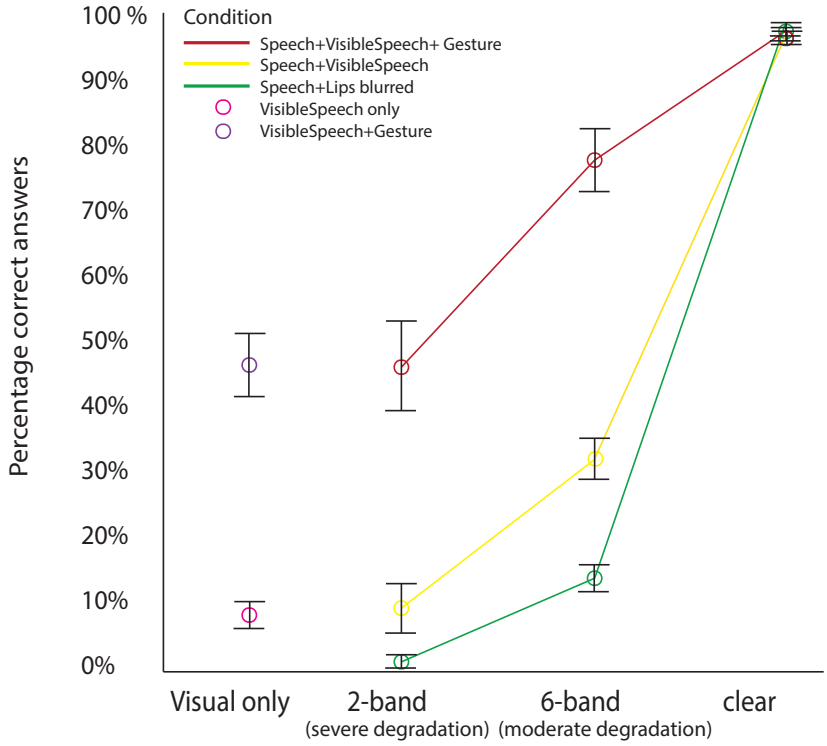


Figure 6: Percentage of correctly identified verbs (% correct) per condition. Error bars represent SD.

To further investigate this interaction, we compared the differences between and within the different noise-vocoding levels and visual articulators in a separate analysis. This analysis allowed us to compare the enhancement driven by different visual articulators as well as compare those enhancement effects between noise-vocoding levels. In comparing the enhancement from the different visual articulators, we recognized that calculating the absolute gain in terms of difference scores is limited in appropriately characterizing the maximum gain per condition. This is because there is an inverse relationship that exists between the performance in the Speech+Lips blurred and Speech+VisibleSpeech conditions and the maximum benefit that is derived when calculating the enhancement of the different visual articulators (see Grant & Walden, 1996). For example, we found a 2.75% recognition rate for Speech+Lips blurred in 2-band noise-vocoding as compared to 11.75% in 6-band noise-vocoding. The maximum gain possible

on the basis of pure difference scores would therefore be 97.75% for 2-band noise-vocoding, and 88.25% for 6-band noise-vocoding, which would be hard to compare, since the maximal gain that is possible in 2-band noise-vocoding is larger than in 6-band noise-vocoding.

Therefore, to avoid possible floor effects and in keeping with previous studies, such as Sumby & Pollack (1954), we controlled for this by defining three difference scores ((A-B/100-B) (i.e., enhancement types)) for a) VisibleSpeech enhancement: Speech+VisibleSpeech - Speech+Lips blurred; b) Gestural enhancement: Speech+VisibleSpeech+Gesture - Speech+VisibleSpeech; and c) Double enhancement: Speech+VisibleSpeech+Gesture - Speech+Lips blurred, (see Ross et al., 2007 for a discussion of other calculation methods) divided by the maximal possible enhancement (for VisibleSpeech enhancement: 100 -Speech+Lips blurred; for Gestural enhancement: 100 - Speech+VisibleSpeech; for Double enhancement: 100 - Speech+Lips blurred). We subjected these outcomes to a repeated measures ANOVA with the factors Noise-Vocoding (2-band, 6-band, clear) and EnhancementType (VisibleSpeech enhancement, Gestural enhancement, Double enhancement). Our analysis revealed a main effect of Noise-Vocoding ($F(2,38) = 320.23$, $p < .001$, partial $\eta^2 = .94$), indicating that the more degraded the signal was, the less enhancement was present. Moreover, we found a main effect of EnhancementType ($F(1.06, 20.19) = 276.74$, $p < .001$, partial $\eta^2 = .94$, Greenhouse-Geisser corrected), indicating that the more visual information was present, the more participants answered correctly. Importantly, we found a significant interaction between EnhancementType and Noise-Vocoding ($F(1.97, 37.37) = 102.65$, $p < .001$, partial $\eta^2 = .84$, Greenhouse-Geisser corrected). Pairwise comparisons (all Bonferroni corrected) showed a significant difference between Gestural enhancement and VisibleSpeech enhancement in both the 2-band noise-vocoding condition ($t(19) = 9.41$, $p_{bon} < .001$ and the 6-band noise-vocoding condition ($t(19) = 12.94$, $p_{bon} < 0.001$) Furthermore, the difference between Gestural enhancement and VisibleSpeech enhancement was larger for 6-band noise-vocoding than 2-band noise-vocoding ($F(1,19) = 64.48$, $p_{bon} < .001$, partial $\eta^2 = .77$). Finally, Double enhancement was larger at 6-band noise-vocoding than in 2-band noise-vocoding ($t(19) = -10.035$, $p_{bon} < .001$) (see

Figure 7). Pairwise comparisons showed a significant difference in VisibleSpeech enhancement and Double enhancement in both 2-band noise-vocoding ($t(19)$ = 12.47, p_{bon} < . 001) and 6-band noise-vocoding ($t(19)$ = 20.79, p_{bon} < .001). This difference between VisibleSpeech enhancement and Double enhancement was larger in 6-band noise-vocoding than in 2-band noise-vocoding ($F(1,19)$ = 163.20, p_{bon} < .001, partial η^2 = .90). Additionally, pairwise comparisons showed a significant difference in Gestural enhancement and Double enhancement in both 2-band noise-vocoding ($t(19)$ = 3.36, p_{bon} <0.01) and 6-band noise-vocoding ($t(19)$ = 7.79, p_{bon} < .001), which was again largest in 6-band noise-vocoding ($F(1,19)$ = 30.44, p_{bon} < .001, partial η^2 = .62).

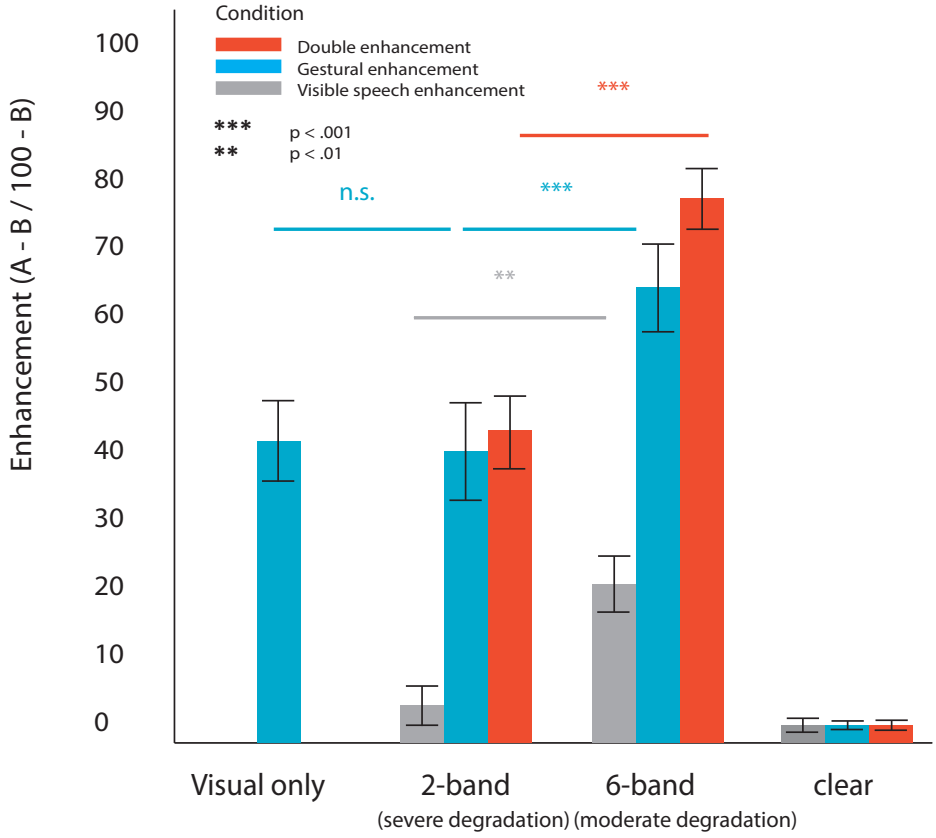


Figure 7: Enhancement effect (A – B/100 – B) corrected for floor effects. Error bars represent SD; n.s. = not significant.

Initially, we did not include the two Visual Only conditions (VisibleSpeech only, VisibleSpeech+Gesture) in our main analysis, because they would create an unbalanced design for analyzing all conditions together. However, these conditions were still of interest to determine how much information participants could obtain from visual input alone without speech being present. Therefore, we first tested the difference between the two separate Visual only conditions by means of a paired samples t-test. We found a significant difference between VisibleSpeech only and VisibleSpeech+Gesture ($t(19) = 15.12$, $p < 0.001$), indicating that response accuracy was higher for trials containing both visible speech and gestures, compared to videos that just contained visible speech (see Figure 6). Subsequently, we compared this difference between VisibleSpeech+Gesture and VisibleSpeech Only (i.e. gestural enhancement, computed as the difference between (VisibleSpeech+Gesture – VisibleSpeech only/100 - VisibleSpeech Only)) to the Gestural enhancement in the context of speech (Speech+VisibleSpeech+Gesture -Speech+VisibleSpeech/100 - Speech+VisibleSpeech) both in the 6-band and 2-band noise-vocoding conditions (see Figure 7). Our analysis revealed a significant difference between Gestural enhancement in the Visual Only conditions and Gestural enhancement in 6-band noise-vocoding, ($t(19) = -3.23$, $p_{bon} < 0.05$), but not compared to 2-band noise-vocoding condition ($t(19) = 1.1$, $p_{bon} > .1$). These results confirmed that Gestural enhancement in 6-band noise-vocoding was significantly greater compared to 2-band noise-vocoding and compared to Gestural enhancement in the Visual only conditions. However, Gestural enhancement in the Visual Only conditions was not larger than Gestural enhancement in 2-band noise-vocoding, indicating that if there are no longer reliable auditory cues available (as in 2-band noise-vocoding), comprehension might be comparable to when there is no auditory input at all (as in Visual Only conditions).

We have explored the error types per Visual Articulator, per Noise-Vocoding level. However, since the percentage of error type in some conditions was very low, we did not subject these error types to a statistical analysis. To test for possible confounding effects of fatigue or learning, we also compared the amount of correct answers per block. We found no difference between the different blocks in the experiment in correct answers ($p > .1$).

2.5. Discussion

The first aim of our study was to reveal whether and to what extent iconic gestures can contribute on top of information from visible speech to enhance degraded speech comprehension, and whether double enhancement from both visual articulators is more beneficial for comprehension than having just visible speech present as a visual articulator, or having no visual articulators present. Whereas previous studies have approached the contribution of these two visual articulators only separately, we investigated the enhancement effects of iconic gestures and visible speech in a joint context. Since iconic gestures can provide information on a semantic level and visible speech can provide information on a phonological level, we expected an additive effect of gestures on top of the enhancement of visible speech during degraded speech comprehension. Our data indeed showed that while perceiving degraded speech in a visual context, listeners benefit most from having both visible speech and iconic gestures present, as compared to having just visible speech present, or having only auditory information present. Here, gestures provide an additional benefit on top of the enhancement of visible speech.

Our second aim was to demarcate the noise conditions under which this double enhancement from both visible speech and iconic gestures in the context of visible speech add the most to degraded speech comprehension. Our data suggests that at a moderate level of noise-vocoding (6-band), there is an optimal range for maximal multimodal integration where listeners can benefit most from the visual information. The enhancement effects of VisibleSpeech enhancement, Gestural enhancement and Double enhancement were significantly larger in 6-band noise-vocoding than in 2-band noise-vocoding or in the Visual Only conditions. However, we did not find a difference in Gestural enhancement between 2-band noise-vocoding and Visual Only conditions. Taken together, our results showed that at this optimal enhancement level of 6-band noise-vocoding, auditory cues were still moderately reliable and listeners were able to combine and integrate information from both visible speech and iconic co-speech gestures to aid in comprehension, resulting in an additive effect of double, multimodal enhancement from visible speech and iconic gestures. Here, semantic information from iconic gestures adds

to the mapping between the speech signal and phonological information that is derived from lip movements in visible speech. Below we will discuss these results in more detail.

In line with previous research, we found a significant benefit of adding information from visible speech to the speech signal (VisibleSpeech enhancement), in response to stimuli from both noise-vocoding levels (e.g. Sumby & Pollack, 1954). This benefit from solely visible speech was significantly larger at a moderate level of noise-vocoding (6-band) than at a severe level of noise-vocoding (2-band). Previously, it has been suggested that the benefit from visible speech continues to increase as the information that is available from auditory inputs decreases (Sumby & Pollack 1954; Erber, 1969; 1975; Meredith & Stein, 1983), as would be predicted by the principle of inverse effectiveness. However, recent studies have argued that there are minimal levels of auditory information necessary before recognition accuracy can be most enhanced by congruent visible input (Ross et al., 2007). Our data concurs with this latter idea, by finding an optimal range for multimodal integration and enhancement, where auditory cues are moderately reliable, and enhancement from visible speech has its maximal effect.

Importantly, the current results provide novel evidence by showing that iconic gestures can enhance this benefit from visible speech even more: We found a significant difference between Gestural enhancement (Speech+VisibleSpeech+Gesture – Speech+VisibleSpeech) and VisibleSpeech enhancement (Speech+VisibleSpeech – Speech+Lips blurred) at both noise-vocoding levels. In addition, we found significant differences between Double enhancement and Gestural enhancement, as well as significant differences between Double and VisibleSpeech enhancement at both noise-vocoding levels. Our results therefore suggest that although both visual modalities enhance degraded speech comprehension, having both iconic gestures and visible speech present (Double enhancement) in the input enhances speech comprehension most. This is in line with previous literature on the benefits of gestures in language processing and theories of communication that postulate that multimodal information combines with speech to aid language comprehension (McNeill 1992; Clark 1996; Goldin-Meadow 2003, see for a review Kelly, Manning, & Rodak, 2008). Interestingly, the

enhancement of both visual articulators (Double enhancement) was significantly larger than VisibleSpeech enhancement at both noise-vocoding levels. This suggests, in line with previous research, that gestures are actively processed and integrated with the speech signal (Kelly et al., 2010; Kendon, 2004), even under conditions where visible speech is visible (also see Holler et al., 2014).

It is important to note that this double enhancement of both iconic gestures and visible speech is in itself still a product of integrating the auditory (speech) and visual input (iconic gestures and visual speech), and not a result of our participants focusing solely on the visual input. The gain in recognition accuracy in our Visual only (VisibleSpeech+Gesture – Visible Speech only) conditions was significantly smaller than the gain we found in the moderate noise (6-band noise-vocoding) condition. The fact that we did not find a similar difference in enhancement between the Visual only conditions and the severe degradation (2-band noise-vocoding) condition suggests that in 2-band noise-vocoding, visible speech cannot be reliably matched to phonological information in the speech signal, and listeners might have focused more on semantic information from gestures to map to the speech signal for disambiguation. As a result, listeners seem to lose the additive effect of double enhancement from visible speech and gestures for speech comprehension in 2-band noise-vocoding because there are not enough reliable auditory cues present in the speech signal to map visible speech too. Consequently, in 2-band noise-vocoding and Visual Only conditions, Gestural enhancement is solely consisting of what can be picked up semantically from the gesture, in addition to information from visible speech. Taken together, we therefore suggest that listeners are only able to benefit from double enhancement from both gestures and visible speech when auditory information is still moderately reliable, to facilitate a binding that integrates information from visible speech, gestures and speech into one coherent percept that exceeds a certain reliability threshold, forming an optimal range where maximal multimodal integration and enhancement can occur.

In earlier work on the contribution of visible speech and hand gestures to learning non-native speech sounds, Kelly et al. (2008) argued that lip and mouth movements help in auditory encoding of speech, whereas hand gestures only can

help to understand the meaning of words in the speech stream when the auditory signal is correctly encoded. Based on their results, Kelly et al. (2008) argue that the benefits of multimodal input target different stages of linguistic processing. Here, mouth movements seem to aid during phonological stages, whereas hand gestures aid during semantic stages, which, according to the authors, fits with McNeill's (1992) interpretation of speech and gesture forming an integrated system during language comprehension.

The results from the present study indeed concur with the idea that speech and gesture form an integrated system and that the benefits of multimodal input target different stages of linguistic processing. Indeed, visible speech possibly plays a significant role during auditory encoding of speech, but according to our current results, iconic gestures not only benefit comprehension when auditory information can be correctly encoded and understood, but also benefit comprehension under adverse listening conditions (cf. Kelly et al., 2008). Even in 2-band noise-vocoding, when auditory cues are no longer reliable and correct encoding of the auditory input is difficult, gestures significantly enhance comprehension. Instead, our data suggests that when encoding of auditory information is difficult or when auditory cues are largely unreliable, listeners are mostly driven by the semantic information from gestures to guide comprehension, which can be beneficial to disambiguate the auditory cues. However, when auditory cues are moderately reliable and there are enough auditory cues available to map the phonological information of visible speech to, listeners can benefit from a 'double' multimodal enhancement from the two visual articulators, integrating both the phonological information from visible speech and semantic information from gestures with the speech signal. This, in turn, results in an additive effect of the semantic information provided by iconic gestures on top of the phonological information from visible speech. However, in 2-band noise vocoding where phonological information from visible speech can no longer be reliably matched to the speech signal, listeners lose this additive double enhancement effect of visible speech and iconic gestures, and mostly utilize the semantic information from gestures (i.e. Gestural enhancement) to resolve the form of the speech signal. Based on these results, we suggest that at least in adverse listening conditions where auditory cues are no longer reliable, language

processing might be more driven by semantic information that is abstracted from iconic co-speech gestures.

Our findings suggest that the use of iconic gestures can play a pivotal role in natural face-to-face communication: gestural information can help to access the meaning of a word to resolve the form of the speech signal when a listening situation is challenging, such as in noise. One limitation of our work can be that our actress uttered the stimuli in a setting with optimal listening conditions, without any noise. We edited her auditory input after recording, to test the effect of different noise-vocoding bands. In this regard, it is important to note that in a natural adverse listening condition, our speaker would have probably adjusted her articulatory movements to optimally communicate her message. This effect has been previously described as the Lombard effect, which refers to the tendency of speakers to increase their vocal effort when speaking in noise to enhance the audibility of their voice (which is not limited to loudness, but also to the length of phonemes and syllables, speech rate and pitch, amongst others) (Lombard, 1911). Alternatively, this could also have an effect on the production of iconic co-speech gestures as well: for example producing a larger iconic gesture in an adverse listening condition could have resulted in a larger co-speech gesture than in clear speech. Future research could test this possibility by recording stimuli in an adverse listening condition and present these videos to participants, to increase ecological validity. A second limitation of our study can be that our participants were only presented with single action verbs. Future research could investigate whether presenting these verbs in a sentence context might have an influence on how much a listener depends on different visual articulators. In addition, future endeavors could consider that natural face-to-face communication does not only consists of a binding of speech and visual information from gestures and visible speech. Instead, research can tap into the influence of other nonverbal behavior (such as head and brow movements, see e.g., Krahmer & Swerts, 2007) and their co-occurrence with visible speech and gesture to fully understand the optimal conditions for visual enhancement of speech in adverse listening conditions. This, in turn, can further elucidate the results from the current study, but also inform debates on audiovisual training for both clinical populations and

educational instruction. Finally, replicating the effects found in this study with hearing-impaired populations will provide a better diagnosis of their speech comprehension in ecologically valid contexts (i.e., in a multimodal context). This in turn could inform debates on audiovisual training for both clinical populations and educational instruction.

2.6. Acknowledgements

This research was supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from Netherlands Organization for Scientific Research. We thank two anonymous reviewers for their helpful comments and suggestions that helped to improve the paper. We are very grateful to Nick Wood, for helping us in editing the video stimuli and to Gina Ginos, for being the actress in the videos.

Chapter 3

Alpha, beta and gamma oscillations predict gestural enhancement of degraded speech comprehension

3.1. Abstract

During face-to-face communication, listeners integrate speech with gestures. The semantic information conveyed by iconic gestures (e.g., a drinking gesture) can aid speech comprehension in adverse listening conditions. In the current magnetoencephalography (MEG) study, we investigated the spatiotemporal neural oscillatory activity associated with gestural enhancement of degraded speech comprehension. Participants watched videos of an actress uttering clear or degraded speech, accompanied by a gesture or not and completed a 4-alternative forced choice identification task after watching every video. When gestures semantically disambiguated degraded speech comprehension, an alpha and beta power suppression and a gamma power increase revealed engagement and active processing in the hand-area of the motor cortex, the extended language network (LIFG/pSTS/STG/MTG), medial temporal lobe and occipital regions. These observed low- and high-frequency oscillatory modulations in these areas support general unification, integration and lexical access processes during online language comprehension, as well as simulation of and increased visual attention to manual gestures over time. All individual oscillatory power modulations associated with gestural enhancement of degraded speech comprehension predicted a listener's correct disambiguation of the degraded verb after watching the videos. Our results thus go beyond the previously proposed role of oscillatory dynamics in unimodal degraded speech comprehension and provide first evidence for the role of low and high-frequency oscillations in predicting the integration of auditory and visual information at a semantic level.

This chapter is based on Drijvers, L., Ozyurek, A., & Jensen, O. (2018). Hearing and seeing meaning in noise: Alpha, beta and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human Brain Mapping*, 39(5), 2075-2087. doi:10.1002/hbm.23987

3.2. Introduction

Successful face-to-face communication, especially under adverse listening conditions, needs a weighing and integration of linguistic (e.g., speech) and sensory information (e.g., a co-speech gesture). In order to understand how the brain adapts to such audiovisual contexts, a functional network approach is needed in which patterns of ongoing neural activity are considered to allocate computational resources by engaging and disengaging task-relevant brain areas (Jensen and Mazaheri, 2010; Pfurtscheller and Lopes da Silva, 1999; Siegel et al., 2012). Suppression of alpha and beta oscillations is often related to the engagement of task-relevant brain areas, whereas an increase reflects functional inhibition or disengagement (Jensen and Mazaheri, 2010; Klimesch et al., 2007; Pfurtscheller and Lopes da Silva, 1999). Increases in gamma activity have been proposed to reflect enhanced neuronal computation (Fries et al., 2007; Jensen et al., 2007). Previously, oscillatory dynamics in these frequency bands have been studied during auditory comprehension of degraded speech, but it is unknown whether similar mechanisms apply to degraded speech comprehension in the context of meaningful visual input, such as hand gestures. Based on previous research that demonstrated that the magnitude of low and high-frequency activity can predict the degree of audiovisual integration (Hipp et al., 2011), we here investigate whether such oscillatory mechanisms also apply to more realistic settings and audiovisual integration at the semantic level, such as gestural enhancement of degraded speech comprehension.

Listeners routinely process speech and meaningful co-speech gestures. Behavioral and neuroimaging studies on gesture processing have shown that iconic gestures (e.g., a hand mimicking a drinking action) enhance degraded speech comprehension and are integrated with speech (Beattie and Shovelton, 1999; Drijvers and Ozyürek, 2017; Holle et al., 2010; Obermeier et al., 2012; Josse et al., 2012; Özyürek, 2014). fMRI studies have demonstrated that speech-gesture integration involves the LIFG, STS, middle temporal gyrus (MTG), motor and visual cortex (Dick et al., 2014; Green et al., 2009; Straube et al., 2012; Willems et al., 2007; Willems et al., 2009). However, the spatiotemporal neural dynamics of this integration remain unknown.

Studies on unimodal auditory degraded speech comprehension have demonstrated that parietal alpha power is enhanced when speech is degraded (Becker et al., 2013; Drijvers et al., 2016; Obleser and Weisz, 2012; Weisz et al., 2011; Wostmann et al., 2015). These results were interpreted as reflecting increased auditory cognitive load when the language processing system is inhibited due to degradation. Previous research on gesture processing has reported low-frequency (2–7 Hz) modulations to emblems (e.g., thumbs-up gesture occurring without speech) and beat gestures (non-semantic rhythmic hand flicks) (Biau and Soto-Faraco, 2015; He et al., 2015), but the spatiotemporal neural dynamics supporting gestural enhancement of speech remain unknown. By using the good temporal and spatial resolution of MEG we can quantify the spatiotemporal oscillatory dynamics supporting audiovisual integration at a semantic level.

In the current study, we presented participants with videos that either contained clear or degraded speech, accompanied by a gesture or not. Our central hypothesis is that gestures enhance degraded speech comprehension and that comprehension relies on an extended network including the motor cortex, visual cortex, and language network to perform this multimodal integration. Here, brain oscillations are assumed to have a mechanistic role in enabling integration of information from different modalities and engaging areas that contribute to this process. We predict that when integration demands increase, we will observe an alpha (8-12 Hz) power suppression in visual cortex, reflecting more visual attention to gestures, and an alpha and beta (15-20 Hz) power decrease in the language network, reflecting the engagement of the language network and a higher semantic unification load (Wang et al., 2012). Secondly, we expect an alpha and beta power suppression in the motor cortex, reflecting engagement of the motor system during gestural observation (Caetano et al., 2007; Kilner et al., 2009; Koelewijn et al., 2008). Lastly, we predict an increase in gamma power in the language network, reflecting the facilitated integration of speech and gesture into a unified representation (Hannemann et al., 2007; Schneider et al., 2008; Wang et al., 2012; Willems et al., 2007; Willems et al., 2009).

3.3. Methods

3.3.1. Participants

Thirty-two Dutch native students of Radboud University (mean age = 23,2, SD = 3,46, 14 males) were paid to participate in this experiment. All participants were right-handed and reported corrected-to-normal or normal vision. None of the participants had language, motor or neurological impairment and all reported normal hearing. The data of three participants (two females) was excluded because of technical failure (1), severe eye-movement artifacts (> 60% of trials) (1) and excessive head motion artifacts (> 1cm) (1). The final dataset therefore included the data of twenty-nine participants. All participants gave written consent before they participated in the experiment.

3.3.2. Stimulus materials

Participants were presented with 240 short video clips of a female actress who uttered a Dutch action verb, which would be accompanied by a matching iconic gesture, a mismatching iconic gesture or no gesture. In the current chapter, we report the results of a subset of this experiment, containing 160 video clips that either contained a matching iconic gesture or no gesture, to zoom in on the effect of 'gestural enhancement'. These video clips were originally used and pre-tested as part of a previous behavioral experiment in Drijvers & Ozyurek (2017), reported in **Chapter 2**.

The action verbs that were used were all highly frequent Dutch action verbs so that they could easily be coupled to iconic gestures. All videos were recorded with a JVC HY-HM100 camcorder and had an average length of 2000 ms (SD = 21.3 ms.). The actress in the video was wearing neutrally colored clothes and was visible from the knees up, including the face. In the videos where she made an iconic gesture, the preparation of the gesture (i.e., the first video frame that shows movement of the hand) started 120 ms. (SD = 0 ms.) after onset of the video, the stroke (i.e., the meaningful part of the gesture) started on average at 550 ms. (SD = 74.4 ms.), gesture retraction started at 1380 ms (SD = 109.6 ms.) and gesture

offset at 1780 ms. (SD = 150.1 ms.). Speech onset started on average at 680 ms. (SD = 112.54 ms.) after video onset, In previous studies this temporal lag was found to be ideal for information from the two channels to be integrated during online comprehension (Habets et al., 2011). In 80 of the 160 videos, the actress produced an iconic gesture. All gestures were iconic movements that matched the action verb (see below). In the remaining 80 videos the actress uttered the action verbs with her arms hanging casually on each side of the body.

It is important to note here that all of the iconic gestures were not prescripted by us but were renditions by our actress, who spontaneously executed the gestures while uttering the verbs one by one. As such, these gestures resembled those in natural speech production, as they were meant to be understood in the context of speech, but not as pantomimes which can be fully understood without speech. We investigated the recognizability of all our iconic gestures outside a context of speech by presenting participants with all video clips without any audio, and asked them to name a verb that depicted the video (as part of Drijvers & Ozyurek, (2017)). We coded answers as 'correct' when a correct answer or a synonym was given in relation to the verb each iconic gesture was produced with, and as 'incorrect' when the verb was unrelated. The videos had a mean recognition rate of 59% (SD = ~16%;) which indicates that the gestures were potentially ambiguous in the absence of speech, as they are in the case of naturally occurring co-speech gestures (Krauss et al., 1991). This ensured that our iconic gestures could not be understood fully without speech (e.g., a 'mopping' gesture, which could mean either 'rowing', 'mopping', 'sweeping' or 'cleaning', and thus needs the speech to be understood) and that our participants could not disambiguate the degraded speech fully by just simply looking at the gesture and labelling it. Instead, participants needed to integrate speech and gestures for successful comprehension. For further details on the pretesting of our videos, please see Drijvers & Ozyurek (2017).

We extracted the audio from the video files, intensity-scaled the speech to 70 dB and de-noised the speech in *Praat* (Boersma and Weenink, 2015). All sound files were then recombined with their corresponding video files. The speech in the videos was presented either clear or degraded (Shannon et al., 1995). As in a previous study on gestural enhancement of degraded speech comprehension

(Holle et al., 2010), we determined in our previous behavioral study (Drijvers and Ozyürek, 2017) which degradation level was optimal for gestural information to have the largest impact on enhancing degraded speech comprehension. In going beyond Holle et al., (2010), the only previous study on gestural enhancement of degraded speech, we did not cover the face of the actor and thus studied the gestural enhancement effect in a more natural context. This allowed us to investigate how gestures enhance degraded speech comprehension on top of the context of the (phonological) cues that are conveyed by visible speech. In Drijvers & Ozyurek (2017), participants completed an open-set identification task where they were asked to write down the verb they heard in videos that were either presented in 2-band noise-vocoding, 6-band noise-vocoding, clear speech, and visual only conditions that did not contain any audio.

Our previous results from Drijvers & Ozyurek (2017) demonstrated that listeners benefitted from gestural enhancement most at a 6-band noise-vocoding level. At this noise-vocoding level, auditory cues were still reliable enough to benefit from both visual semantic information and phonological information from visible speech. However, in 2-band noise-vocoding, listeners could not benefit from the phonological information that was conveyed by visible speech to couple the visual semantic information that was conveyed by the gesture. Instead, in 2-band noise-vocoding, the amount of correct answers was as high in the visual only condition that did not have audio.

In addition to clear speech, we thus created a 6-band noise-vocoding version of each clear audio file that was then recombined with the video, using a custom-made script in *Praat*, by bandpass filtering each sound file between 50 Hz and 8000 Hz and dividing the speech signal by logarithmically spacing the frequency bands between 50 and 8000 Hz. In more detail, this resulted in cutoff frequencies of 50 Hz, 116.5 Hz, 271.4 Hz, 632.5 Hz, 1473.6 Hz, 3433.5 Hz and 8000 Hz. We used half-wave rectification to extract the amplitude envelope of each band and multiplied the amplitude envelope with the noise bands before recombining the bands to form the degraded speech signal. The sound of the videos was presented through MEG compatible air-tubes.

In total, we included four conditions in our experiment: a clear speech only

condition (C), a degraded speech only condition (D), a clear speech + iconic gesture condition (CG) and a degraded speech + iconic gesture condition (DG) (see Figure 8A). All four conditions contained 40 videos, and none of the verbs in the videos overlapped. Note that we did not follow the design described in Drijvers & Ozyurek (2017), as using eleven conditions would have resulted in a very low number of trials per condition for source analyses.

Finally, to assess the participants' comprehension of the verbs, we presented participants with a 4-alternative forced choice identification task (see for details below) instead of the open-set identification task that was used in Drijvers & Ozyurek (2017), as an open-set identification task would have caused too many (motion) artifacts for the MEG analyses. Note that all stimuli can be found under Appendix I.

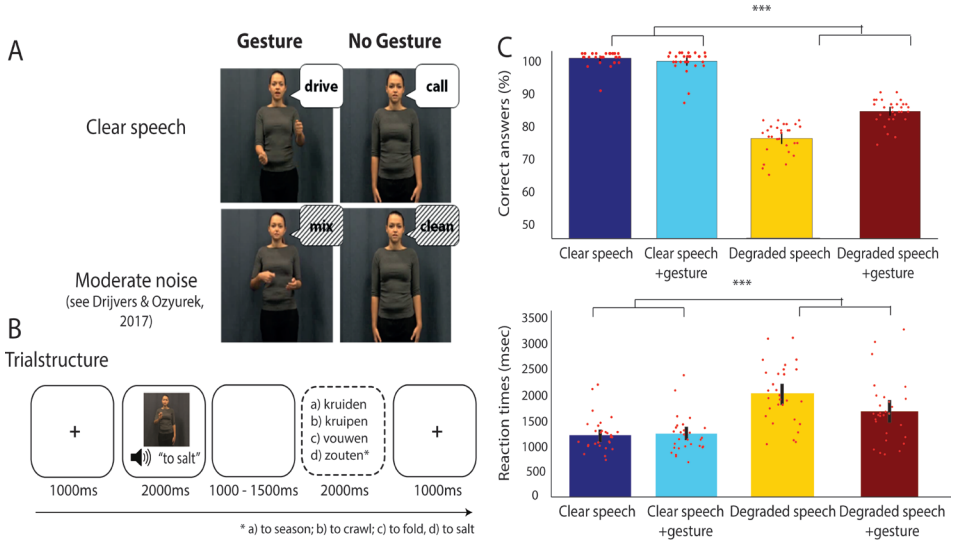


Figure 8 A: Illustration of the different conditions B: Trial structure. C: Upper panel: percentage of correct answers per condition. Error bars represent SD. *** = $p < 0.01$. Lower panel: reaction times (in milliseconds) per condition. Error bars represent SD.

3.3.3. Procedure

Participants were tested in a dimly-lit magnetically shielded room and seated 70 cm from the projection screen. All videos were projected onto a semi-translucent screen by back-projection using an EIKI LC-XL100L projector with a resolution of 1650x1080 pixels. The stimuli were presented full screen using Presentation software (Neurobehavioral Systems, Inc.) In the experiment, participants were asked to attentively listen and watch the videos. Each trial started with a fixation cross (1000 ms.), followed by the video (2000 ms.), a short delay (1000-1500 ms, jittered), followed by a 4-alternative forced choice identification task After watching the videos, participants were asked to identify what verb they had heard in the last video. Participants could indicate their choice by a right-hand button press on a 4-buttonbox where the 4 buttons represented the answering options for either a, b, c or d. These answering options always contained a phonological distractor, a semantic distractor, an unrelated answer and the correct answer. For example, the correct answer could be 'kruiden' (to season), the phonological distractor could be 'kruipen' (to crawl), the semantic distractor, which would fit with the gesture, could be 'zouten' (to salt), and the unrelated answer could be 'vouwen' (to fold) (see Figure 8B). The 4-alternative forced choice identification task ensured that participants were paying attention to the videos, and to check whether participants behaviorally resolved the verbs. Furthermore, the semantic competitors were included to investigate whether participants were focusing on the gesture only in the degraded speech conditions. We predicted that if this was indeed the case, they would choose the semantic competitors if they solely zoomed in on the gesture and ignored the degraded speech. After participants indicated their answers, a new trial would start after 1500 ms. (see Figure 8B). Participants were asked not to blink during the videos, but to blink after they had answered the question in the 4-alternative forced choice identification task.

Brain activity was recorded with MEG during the whole task, which consisted of 4 blocks of 40 trials. Participants had a self-paced break after each block. The whole experiment lasted about one hour, including preparation of the participant and instruction of the task. All participants were presented with a different pseudo-randomization of the stimuli, with the constraint that a specific condition

(e.g., two trials of DG) could not be presented more than twice in a row.

3.3.4. Experimental Design & Statistical Analyses

3.3.4.1. MEG data acquisition

MEG was recorded by using a 275-channel axial gradiometer CTF MEG system. An online lowpass filter with a cutoff at 300 Hz was applied, the data were digitized at 1.2 kHz and stored for offline analyses. Additionally, we recorded participants' eye gaze by using an SR Research Eyelink 1000 eye tracker, to monitor fixation during the task. Participants' electrocardiogram (ECG) and horizontal and vertical electrooculogram (EOG) were recorded for artifact rejection purposes. To measure and monitor the participants' head position with respect to the gradiometers, we placed three coils at the nasion and left/right ear canal. We monitored head position in real-time (Stolk et al., 2013). After the experimental session, we recorded structural magnetic resonance images (MRI) from 22 out of 32 subjects using a 1.5T Siemens Magnetom Avanto system with markers attached in the same position as the head coils, to align the MRIs with the MEG coordinate system in our analyses.

3.3.4.2. MEG data analyses: preprocessing and time-frequency representations of power

We analyzed the MEG data using FieldTrip (Oostenveld et al., 2011), an open-source MATLAB toolbox. First, the MEG data were segmented into trials starting 1 s before and ending 3 s after the onset of the video. The data were demeaned, a linear trend was fitted and removed. Line noise was attenuated using a discrete Fourier transform approach at 50 Hz and 100 Hz (first harmonic) and 150 Hz (second harmonic). We applied a third-order synthetic gradiometer correction (Vrba and Robinson, 2001) to reduce environmental noise, and rejected trials (on average 6.25%) that were contaminated by SQUID jump artifacts or muscle artifacts using a semi-automatic routine. Subsequently, we applied independent component analysis (Bell and Sejnowski, 1995; Jung et al., 2000) to remove eye-movements and cardiac related activity. Finally, the data were inspected visually

to remove artifacts that were not identified by these rejection procedures and resampled the data to 300 Hz to speed up the subsequent analyses (average number of trials per participant discarded: 9,97, SD = 3,08). To facilitate interpretation of the MEG data, we calculated synthetic planar gradients, as planar gradient maxima are known to be located above neural sources that may underlie them (Bastiaansen and Knösche, 2000). Here, the axial gradiometer data were converted to orthogonal planar gradiometer pairs, after which power was computed, and then the power of the pairs was summed.

The calculation of time-frequency representations (TFRs) of power per condition was carried out in two frequency ranges to optimize time and frequency resolution. First, we calculated the TFRs of the single trials between 2-30 Hz, by applying a 500 ms. Hanning window in frequency steps of 1 Hz and 50 ms. time steps. In the 30 - 100 Hz frequency range a multitaper (discrete prolate spheroidal sequences) approach was used (Mitra and Pesaran, 1999), by applying a 500 ms. window length, 2 Hz frequency steps, 50 ms time steps and 5 Hz frequency smoothing. To capture the gestural enhancement effect, we compared the differences in *Degraded Speech+Gesture* and *Degraded Speech* to the difference in *Clear Speech+Gesture* and *Clear Speech.* The four conditions (C, D, CG, DG) were averaged separately for each participant. TFRs were then log10 transformed and the difference between the conditions (D vs C, DG vs CG, DG vs D and CG vs C) was calculated by subtracting the log10 transformed power ('log ratio', e.g., log10(DG) - log10(D)). To calculate the effect of gestural enhancement, we compared the differences between DG vs D and CG vs C. (i.e., (log10(DG) - log10(D)) - (log10(CG) - log10(C)) Our time window of interest was between 0.7 and 2.0. which corresponded to the speech onset of the target word until the offset of the video. The range of our frequency bands of interest were selected on the basis of our hypotheses, as well as a grand average TFR of all conditions combined.

3.3.4.3. MEG data analysis: Source analyses

Source analysis was performed using Dynamic Imaging of Coherent Sources (DICS, Gross et al., (2001)) as a beamforming approach. We based our source

analysis on the data recorded from the axial gradiometers. DICS computes a spatial filter from the cross-spectral density matrix (CSD) and a lead field matrix. We obtained individual lead fields for every participant by spatially co-registering the individual anatomical MRI to sensor space MEG data by identifying the anatomical markers at the nasion and the two ear canals. We then constructed a realistically shaped single-shell head model on the basis of the segmented MRI for each participant, divided the resulting brain volume into a 10 mm spaced grid and warped it to a template brain (MNI). We also used the MNI template brain for the participants who did not come back for the MRI scan.

The CSD was calculated on the basis of the results of the sensor-level analyses: For the alpha band, we computed the CSD between 0.7-1.1, 1.1 - 1.5 and 1.6 -2.0 s at 10 Hz with +/- 2.5 Hz frequency smoothing. For the beta band, we computed the CSD between 1.3 - 2.0 s, centered at 18Hz with +/- 4Hz frequency smoothing and for the gamma band between 1.0 and 1.6s, between 65 and 80 Hz, with 10 Hz frequency smoothing. A common spatial filter containing all conditions was calculated and the data were projected through this filter, separately for each condition. The power at each gridpoint was calculated by applying this common filter to the conditions separately, and was then averaged over trials and log-10 transformed. The difference between the conditions was again calculated by subtracting the log-power for the single contrasts, and interaction effects were obtained by subtracting the log-power for the two contrasts. Finally, for visualization purposes, the grand average grid of all participants was interpolated onto the template MNI brain.

3.3.4.4. Cluster-based permutation statistics

We performed cluster-based permutation tests (Maris and Oostenveld, 2007) to assess the differences in power in the sensor and source-level data. The statistical tests on source level data were performed to create statistical threshold masks to localize the effects we observed on sensor level. A non-parametric permutation test together with a clustering method was used to control for multiple comparisons. First, we computed the mean difference between two conditions for each data sample in our dataset (sensor: each sample for sensor TFR analysis,

source: x/y/z sample for source space analysis). Based on the distribution that is obtained after collecting all the difference values for all the data samples, the observed values were thresholded with the 95th percentile of the distribution, which were the cluster candidates (i.e., mean difference instead of t-values), and randomly reassigned the conditions in participants 5000 times to form the permutation distribution. For each of these permutations, the cluster candidate who had the highest sum of the difference values was added to the permutation distribution. The actual observed cluster-level summed values were compared against the permutation distribution, and those clusters that fell in the highest or lowest 2.5% were considered significant. For the interaction effects, we followed a similar procedure and compared two differences to each other. Note that we do not report effect sizes for these clusters as there is not a simple way of translating the output of the permutation testing to a measure of effect size.

3.3.4.5. The relation between alpha, beta, and gamma oscillations and behavioral 4-alternative forced choice identification scores

We further tested whether power modulations in the alpha, beta, and gamma frequency band were related to the participants' individual scores on the 4-alternative forced choice identification task. Specifically, we quantified the individual's power modulation in each frequency band by averaging the power modulation over time points, frequencies and sensors in significant clusters of the interaction effects, resulting in an individual's modulation score per frequency band. Similarly, we calculated an interaction score for gestural enhancement on the behavioral task by comparing the difference in the percentage of correct answers of DG-D to the difference in CG-C, resulting in the amount of behavioral enhancement per participant. We then obtained Spearman correlation between this score and the power modulation per frequency band. As our hypotheses stated that the gestural enhancement effect would be supported by an alpha/beta suppression and a gamma power increase, we used one-tailed t-tests to test for this correlation.

3.4. Results

Participants were presented with videos that contained a gesture or no gesture, and listened to action verbs that were degraded or not (Figure 8 A, B). After each presentation, participants were prompted by a 4-alternative forced choice identification task and instructed to identify which verb they had heard in the videos (Figure 8B). We defined the 'gestural enhancement' as the interaction between the occurrence of a gesture (present/not present) and speech quality (clear/degraded), and predicted that the enhancement would be largest when speech was degraded and a gesture was present. Brain activity was measured using whole-head MEG throughout the whole experiment. The time interval of interest for the analysis was always 0.7 - 2.0s, from speech onset until video offset (Figure 10A).

3.4.1. Gestural enhancement is largest during degraded speech comprehension

Our behavioral data revealed, in line with previous behavioral studies (Drijvers and Ozyürek, 2017; Holle et al., 2010), that gesture enhanced speech comprehension most when speech was degraded. The percentage of correct answers in the 4-alternative forced choice identification task were analyzed by applying a repeated measures ANOVA with the factors Noise (clear speech vs. degraded speech) and Gesture (present vs. not present). This revealed a main effect of Noise, indicating that when speech was clear, participants were better able to identify the verb than when the speech was degraded ($F(1,28) = 83.79$, $p < 0.001$, $\eta^2 = .75$). A main effect of Gesture ($F(1,28) = 7.93$, $p = 0.009$, $\eta^2 = .22$), demonstrated that participants provided more correct answers when a gesture was present. Our main finding was a significant interaction between Noise and Gesture ($F(1,28) = 17.12$, $p < 0.001$, $\eta^2 = .38$), which indicated that gestures facilitated speech comprehension in particular in the degraded condition. A repeated measures ANOVA applied to the reaction times with the factors Noise (clear speech vs. degraded speech) and Gesture (present vs. not present) revealed a main effect of Noise, indicating that when the speech signal was clear, participants responded faster ($F(1,28) = 93.02$, $p < .001$, $\eta^2 = .77$). A main effect of Gesture ($F(1,28) = 5.66$, $p = .024$, $\eta^2 = .17$;

Figure 8C), indicated that when a gesture was present, participants responded faster. The data revealed an interaction between Noise and Gesture ($F(1,28)$ = 12.08, p <.01, η^2 = .30), which indicated that when speech was degraded and a gesture was present, participants were quicker to respond.

It should be acknowledged that these results seem attenuated as compared to the results from Drijvers & Ozyurek (2017). In this experiment, we for example reported a behavioral benefit when comparing DG to D of approximately 40%, as compared to approximately 10% in the current study. This can be explained by the type of task we used. In the open-set identification task, participants were unrestricted in their answers, whereas in the 4-alternative forced choice identification task, recognition was easier. This especially had an influence on the increased recognition of the verbs in the D condition, where participants were more able to correctly identify the verb when the answers were cued. Nevertheless, we see a similar pattern (DG-D) in the data of the current study and Drijvers & Ozyurek (2017). Note that the low number of errors in the current study, as well as the low amount of semantic errors (~3%, SD = 1.6%), confirmed that the participants did not solely attend to the gesture for comprehension in the DG condition.

3.4.2. Alpha power is suppressed when gestures enhance degraded speech comprehension.

Next we asked how oscillatory dynamics in the alpha band were associated with gestural enhancement of degraded speech comprehension. To this end, we calculated the time-frequency representations (TFRs) of power for the individual trials. These TFRs of power were then averaged per condition. The interaction was calculated as the log-transformed differences between the conditions. Figure 9 presents the TFRs of power in response to gestural enhancement at representative sensors over the left temporal, right temporal and occipital lobe. We observed a suppression of alpha power in the right temporal lobe at speech onset when speech was degraded and a gesture was presented, suggesting engagement of right-temporal areas in an early time window. Additionally, we predicted that alpha would be suppressed over visual regions to allow for more visual attention

to the gestures when speech was degraded. In line with our hypotheses, the TFR over occipital regions clearly showed a suppression of alpha power (8-12 Hz) over the full time interval. Lastly, the TFR of the left temporal lobe revealed a strong alpha suppression from 1.1s until the end of the video, suggesting engagement of the language system.

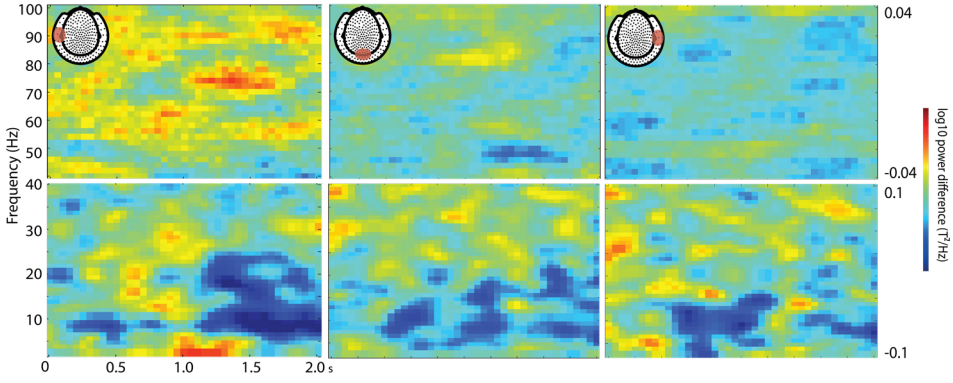


Figure 9. Time-Frequency representations (TFRs) of power of the interaction effect between Noise and Gesture ('Gestural enhancement effect') over three selected groups of representative sensors.

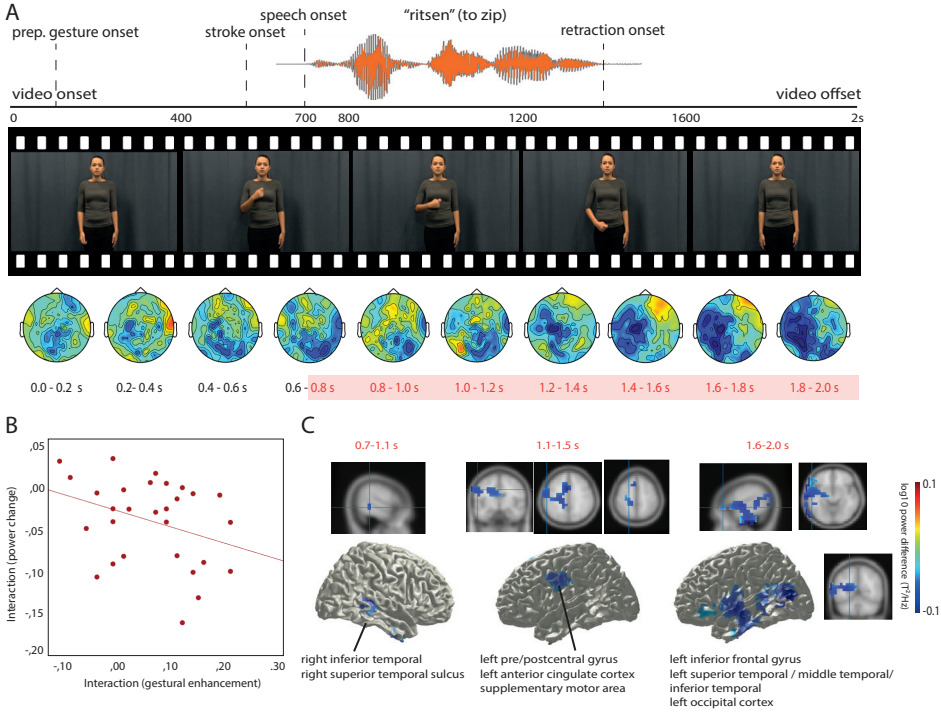


Figure 10. A: Illustration of the structure of the videos. Lower panel: Topographical distribution of oscillatory alpha power of the gestural enhancement effect in 200ms time bins from the start of the video until the end of the video. Shaded time windows denote significant clusters in sensor-level analyses. B: Individual's alpha power modulations as a function of individual's gestural enhancement scores on the 4-alternative forced choice identification task. C: Source-localized results of the interaction effect in the alpha-band, masked by statistically significant cluster.

To get more insight into these effects in space and time, we visualized the topographical distribution of the interaction in the alpha band over time (Figure 10A). The top panel represents structure of the videos, and the lower panel shows the topographical distributions over time of alpha power. These topographies reveal an early suppression of alpha power in the right temporal lobe (0.7 - 1.1 s), followed by an alpha suppression over left central regions (1.1 - 1.5 s) and left-temporal and occipital regions (1.6 - 2.0 s).

Sensor-level analyses of the interaction effect confirmed a larger suppression of alpha power in in response to DG-D as compared to CG-C, indicating that when

speech is degraded and a gesture is present, alpha power was more suppressed. A cluster randomization approach controlling for multiple comparisons in time and space revealed one negative cluster (0.7 – 2.0 s: $p < .001$, summed cluster statistic = -53.3).

Finally, we correlated the individual alpha power modulation with individual behavioral scores on the 4-alternative forced choice identification task, which revealed that the more a listener's alpha power was suppressed, the more a listener showed an effect of gestural enhancement during degraded speech comprehension (Spearman's rho = -,465, $p = 0.015$, one-tailed, FDR corrected Figure 10B).

3.4.3. Alpha suppression reveals engagement of rSTS, LIFG, language network, motor, and visual cortex.

To determine the underlying sources of this alpha power modulation during gestural enhancement of degraded speech comprehension, we used a frequency-domain spatial beamformer technique (DICS, Gross et al., (2001)). Instead of calculating the source of the negative cluster that was found in the sensor analysis over the whole time window (0.7 - 2.0 s), we divided this time window over three separate time windows, due to the distinct spatial sources that differed over time (0.7 – 1.1 s, 1.1 – 1.5 s, 1.6 – 2.0 s, see topographical plots in Figure 10). Furthermore, we applied a cluster-randomization approach to the source data to find a threshold for when to consider the source estimates reliable (note that the cluster-approach at sensor level constitutes the statistical assessment; not the source level approach).

Figure 10C shows that in the 0.7 – 1.1 s window, the source of the alpha power interaction was localized to the rSTS and to a lesser extent, the right inferior temporal lobe. This suggests engagement of the rSTS during gestural enhancement of degraded speech comprehension immediately after speech onset (one negative cluster, $p = 0.042$, summed cluster statistic = -9.64). In the 1.1 – 1.5 s time-window, the source of the alpha effect was localized to the left pre- and postcentral gyrus, as well as the supplementary motor area (SMA) and (anterior) cingulate cortex (ACC) (one negative cluster, $p = 0.016$, summed cluster statistic = -18.58). The axial plots in the second time window in Figure 10C reveal that this

alpha effect extends over a large part of the motor cortex and cingulate cortex. The alpha effect in the 1.6 – 2.0 s time window (one negative cluster, $p = 0.002$, summed cluster statistic = -26.65) was estimated in the LIFG, STG, MTG, ITG, and left occipital cortex. These results suggest engagement of an initially right lateralized source, followed by left central, temporal and occipital sources during gestural enhancement of degraded speech comprehension. For comparisons of the single contrasts, please see Supplementary Materials S1.

3.4.4. Beta power is suppressed when gestures enhance degraded speech comprehension.

Next, we investigated whether gestural enhancement induced modulations of oscillatory beta power. The TFRs of the interaction effect in Figure 9 revealed a left-lateralized beta power suppression (15-20 Hz) from 1.3 - 2.0s, possible extending to more posterior areas. We first plotted the topographical distribution of beta power over time to further investigate the spatiotemporal course of this effect (Figure 11A) and observed a larger beta power suppression from ~1000 ms, when the meaningful part of the gesture commences, which extended until the end of the video. Sensor-level analyses of the interaction effect confirmed a stronger suppression of beta power in DG-D than in CG-C from 1.3 – 2.0s (one negative cluster $p < .001$, summed cluster statistic = -32.85). We correlated the beta power modulation per participant with individual scores on the 4-alternative forced choice identification task, which demonstrated a significant relationship between the amount of beta power suppression and the benefit an individual had from gestures when speech was degraded (i.e., gestural enhancement, see Figure 11C) (Spearman's rho -,352, $p = 0.03$, one-tailed, FDR corrected).

3.4.5. Beta power suppression reflects engagement of LIFG, left motor, SMA, ACC, left visual and left temporal regions.

We then localized the gestural enhancement effect to test our hypotheses on the sources for this effect (Figure 11). This analysis demonstrated that the stronger suppression of beta power was localized (one negative cluster, 1.3-2.0 s, $p < .001$, summed cluster statistic = -26.13) in the left pre- and postcentral gyrus, ACC,

SMA, LIFG, but was also extended to more temporal sources, such as the left superior, medial and inferior temporal regions, the left supramarginal gyrus and the visual cortex. Note that the observed sources partially overlap with the sources in the alpha band (see Figure 10C). This might suggest that some of the beta sources are explained by higher harmonics in the alpha band. Note however that there is a clearer motor beta effect in the beta band than the alpha band: The cluster in the beta band is extending over a part of the motor cortex that corresponds to the hand region of the primary motor cortex, whereas the alpha effect in

Figure 10B is more pronounced over the arm-wrist region. This suggests that this beta power effect is possibly more motor-related than the observed alpha effect. For comparisons of the single contrasts, please see Supplementary Materials S2.

3.4.6. Gamma power is enhanced when gestures aid degraded speech comprehension

Finally, we investigated whether gestural enhancement induced reliable modulations of oscillatory power in the gamma band. The TFRs in Figure 9 revealed a left-temporal increase in gamma band power at 65 – 80 Hz. We plotted the topographical distributions of this interaction in the gamma band to investigate the spatiotemporal profile (Figure 11D). These topographical plots showed a similar gamma power increase in the 1.0 – 1.6 s interval. Cluster-based permutation tests on sensor-level data of the gestural enhancement effect revealed that this effect was larger in DG-D than in CG-C (one positive cluster, $p = 0.016$, summed cluster statistic = 11.56). Interestingly, these effects occur exactly when the meaningful part of the speech and the meaningful part of the gesture are unfolding. A listener's individual gamma power increase correlated positively with how much this listener could benefit from the semantic information conveyed by a gesture to enhance degraded speech comprehension (Figure 11F, Spearman's rho = .352, $p = 0.03$, one-tailed, FDR corrected).

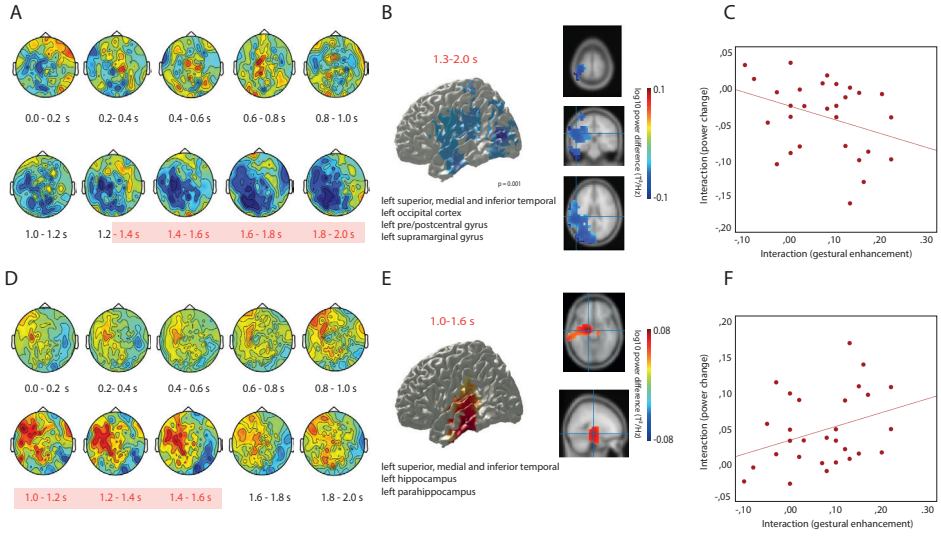


Figure 11 A: Topographical distribution of oscillatory beta power of the gestural enhancement effect in 200ms time bins from the start of the video until the end of the video. Shaded time windows denote significant clusters in sensor-level analyses. B: Source-localized results of the interaction effect in the beta-band, masked by statistically significant clusters. C: Individual's beta power modulations as a function of individual's gestural enhancement scores on the 4-alternative forced choice identification task D: Topographical distribution of oscillatory gamma power of the gestural enhancement effect in 200ms time bins from the start of the video until the end of the video. Shaded time windows denote significant clusters in sensor-level analyses. E: Source-localized results of the interaction effect in the gamma-band, masked by statistically significant clusters. F: Individual's gamma power modulations as a function of individual's gestural enhancement scores on the 4-alternative forced choice identification task.

3.4.7. Gamma power increases in left-temporal and medial temporal areas suggest enhanced neuronal computation during gestural enhancement of degraded speech comprehension.

We hypothesized that gamma power would be increased over LIFG and pSTS/STG/MTG, suggesting a facilitated integration of the visual and auditory information into a unified representation (Hannemann et al., 2007; Schneider et al., 2008; Wang et al., 2012). We therefore conducted source-level analyses to use as a statistical threshold for estimating the source of the observed sensor-level effect. In line with our hypotheses, this increase in gamma band power was

observed over left superior, medial and inferior temporal regions (Figure 11E, one positive cluster, $p = 0.01$, summed cluster statistic = 20.76), suggesting neuronal computation when speech is degraded and a gesture is present. This gamma power increase was also identified in sources in deeper brain structures, such as the medial temporal lobe which will be further discussed in Section 3.5.5. For comparisons of the single contrasts, please see Supplementary Materials S3.

3.4. Discussion

The current study investigated the oscillatory activity supporting gestural enhancement of degraded speech comprehension, to gain insight into the spatiotemporal neuronal dynamics associated with semantic audiovisual integration. When gestures enhanced degraded speech comprehension, we observed a stronger alpha and beta power suppression, suggesting engagement of the hand-area of the motor cortex, the extended language network (LIFG/pSTS/STG/MTG), medial temporal lobe and occipital regions. In the alpha band this effect displayed a spatiotemporal shift from rSTS, to left motor cortex, ACC, the language network, and visual cortex. The stronger suppression in the beta band occurred in the left hand area of the primary motor cortex, SMA, ACC, LIFG, left-temporal and visual cortex. Gestural enhancement was associated with enhanced gamma power over left-temporal and medial-temporal lobe regions. All individual oscillatory power modulations significantly correlated with an individual's behavioral score, demonstrating that individual oscillatory power modulations predict how much a listener could benefit from the semantic information conveyed by gestures to enhance degraded speech comprehension. Below we interpret these findings and discuss the putative role of the oscillatory dynamics in task-relevant brain areas during gestural enhancement of degraded speech.

3.5.1. Early alpha suppression reflects engagement of rSTS to optimally process the upcoming word

In an early time-window (0.7 – 1.1 s) we observed stronger alpha suppression in the rSTS when gestures enhanced degraded speech. In fMRI studies on auditory

degraded speech perception, the rSTS has shown to be sensitive to spectral fine-tuning (Scott, 2000; Zatorre et al., 2002) and pitch contours (Gandour et al., 2004; Kotz et al., 2003). In the (audio)visual domain, fMRI and EEG studies have demonstrated that the rSTS responds to motion and intentional action, and bilateral STS showed increased activation during audiovisual integration under adverse listening conditions (Saxe et al., 2004; Schepers et al., 2013). We propose that the rSTS is possibly engaged because the semantic information conveyed by the gesture is most informative during degraded speech comprehension, causing listeners to focus more on the preparation of a gesture early in the video. The larger engagement of the rSTS might thus reflect the increased uptake of gestural information when speech is degraded.

3.5.2. Listeners engage their motor system most when a gesture is presented in degraded speech.

During gestural enhancement of degraded speech comprehension, an alpha (1.1 - 1.5 s) and beta (1.3 - 2.0 s) power suppression were observed over the hand motor area, primary motor cortex, and SMA. This suggests that the involvement of the motor system might be modulated by the listener's interpretation of ongoing speech perception, resulting in the largest engagement when speech is degraded. This suggests that engaging the motor system during gestural observation in degraded speech might be a result of aiding interpretation, rather than simple mirroring of the observed action, or mere involvement limited to the production and perception of linguistic or sensory information (see for debate e.g., Toni, de Lange, Noordzij, & Hagoort, 2008). Rather, our results suggest that listeners might simulate the gesture more when speech is degraded, possibly to extract the meaning of the gesture to aid in interpreting the degraded speech, which is in line with previous studies on action observation (van Elk et al., 2010; Klepp et al., 2015; Weiss and Mueller, 2012) and embodied cognition (Barsalou, 2008; Pulvermüller, Hauk, Nikulin, & Ilmoniemi, 2005).

3.5.3. The ACC engages in implementing strategic processes to use gestural information to understand degraded speech.

The sources of the alpha and beta power suppression described in Section 3.4 both extended to the ACC. Caution should be taken when interpreting deep sources like the ACC when using MEG; however, our results are consistent with related brain imaging findings. Previous research using fMRI reported enhanced activity in the ACC when modality-independent tasks increased in difficulty, when listeners attended to speech, and during degraded speech comprehension (Eckert et al., 2009; Erb et al., 2013; Peelle, 2017), suggesting that these areas are involved in attention-based performance monitoring, executive processes and optimizing speech comprehension performance (Vaden et al., 2013). Additionally, previous research has reported that the ACC might subserve an evaluative function, reflecting the need to implement strategic processes (Carter et al., 2000). As the current effect occurs when the meaningful part of the speech and gesture are unfolding, we interpret the alpha and beta power suppression as engagement of the ACC to enhance attentional mechanisms and possibly strategically shift attention to gestures, and allocate resources to increase the focus on semantic information conveyed by the gesture.

3.5.4. A left lateralized network including IFG, pSTS/MTG, ITG, and STG is most engaged when gestures enhance degraded speech comprehension.

During gestural enhancement of degraded speech, an alpha (1.6 - 2.0 s) and beta (1.3 - 2.0 s) power suppression were observed in LIFG and left posterior temporal regions (pSTS/MTG, ITG, STG). Activation of left posterior temporal regions has been proposed to be involved in accessing lexical-semantic, phonological, morphological and syntactical information (Hagoort, 2013; Lau et al., 2008). The LIFG is thought to be involved in unification operations from building blocks that are retrieved from memory as well as selection of lexical representations and the unification of information from different modalities (Hagoort, 2013). A beta power suppression in LIFG has been related to a higher unification load that requires a stronger engagement of the task-relevant brain network (Wang et al., 2012). In line with this, we suggest that the larger alpha and beta power suppression

in LIFG reflects engagement during the unification of gestures with degraded speech. We tentatively propose that this larger engagement might facilitate lexical access processes by unifying speech and gesture. Here, the semantic information of the gesture might facilitate lexical activation of the degraded word, which simultaneously engages the language network in this process.

Note that this tentative explanation is also supported by analyses conducted over the single contrasts: In line with previous auditory literature (Obleser et al., 2012; Weisz et al., 2011) we observed enhanced alpha power in response to degraded speech, which has been suggested, in line with the functional inhibition framework, to possibly act as a 'gating mechanism' towards lexical integration, reflecting neural oscillators that keep alpha power enhanced to suppress erroneous language activations. However, we observed a larger alpha *suppression* in conditions that contained gestural information. We argue that the occurrence of a gesture thus seems to reverse the inhibitory effect that degraded speech imposes on language processing, by engaging task-relevant brain regions when the semantic information of the gesture facilitates lexical activation, and thus requires less suppression of potentially erroneous activations in the mental lexicon.

3.5.5. Semantic information from gestures facilitates a matching of degraded speech with top-down lexical memory traces in the MTL.

Gamma power was most enhanced when the meaningful part of the gesture and degraded speech were unfolding. This enhancement was estimated in the left (medial) temporal lobe. Enhanced gamma activity has been associated with the integration of object features, the matching of object specific information with stored memory contents and neuronal computation (Herrmann et al., 2004; Tallon-Baudry and Bertrand, 1999). In line with this, the observed gamma effect in the left temporal lobe might reflect cross-modal semantic matching processes in multisensory convergence sites (Schneider et al., 2008), where active processing of the incoming information facilitates an integration of the degraded speech signal and gesture. Next to left-temporal sources, enhanced gamma power was localized in deep brain structures, such as the medial temporal lobe. We tentatively propose that the observed gamma increases in medial temporal regions reflect that the

semantic information conveyed by gestures can facilitate a matching process with lexical memory traces that aids access to the degraded input.

3.5.6. Engagement of the visual system reflects that listeners allocate visual attention to gestures when speech is degraded.

We observed the largest alpha (1.6 - 2.0 s) and beta (1.3 - 2.0 s) suppression during gestural enhancement of degraded speech. We interpret these larger suppressions as engagement of the visual system and allocation of resources to visual input (i.e., gestures), especially when speech is degraded.

3.5.7. Individual oscillatory power modulations correlate with a listener's individual benefit of a gesture during degraded speech comprehension

We demonstrated a clear relationship between gestural enhancement effects on a behavioral and neural level: The more an individual listener's alpha and beta power were suppressed and the more gamma power was increased, the more a listener benefitted from the semantic information conveyed by a gesture during degraded speech comprehension. This gestural benefit was thus reflected in neural oscillatory activity and demonstrates the behavioral relevance of neural oscillatory processes.

3.6. Conclusions

The present work is the first to elucidate the spatiotemporal oscillatory neural dynamics of audiovisual integration in a semantic context and directly relating these modulations to an individual's behavioral responses. When gestures enhanced degraded speech comprehension, alpha and beta power suppression suggested engagement of the rSTS, which might mediate an increased uptake of gestural information when speech is degraded. Subsequently, we postulate that listeners might engage their motor cortex to possibly simulate gestures more when speech is degraded to extract semantic information from the gesture to aid degraded speech comprehension, while strategic processes are implemented by the ACC to allocate attention to this semantic information from the gesture when speech is degraded. We interpret the larger alpha suppression over visual areas as a

larger engagement of these visual areas to allocate visual attention to gestures when speech is degraded. In future eye-tracking research, we will investigate how and when listeners exactly attend to gestures during degraded speech comprehension to better understand how listeners direct their visual attention to utilize visual semantic information to enhance degraded speech comprehension. We suggest that the language network, including LIFG, is engaged in unifying the gestures with the degraded speech signal, while enhanced gamma activity in the MTL suggested that the semantic information from gestures can aid to access the degraded input and facilitates a matching between degraded input and top-down lexical memory traces. The more a listener's alpha and beta power were suppressed, and the more gamma power was enhanced, the more a listener demonstrated a benefit from gestures to enhance speech comprehension. Our results thus go beyond previous work by showing that low and high-frequency oscillations can predict the degree of integration of audiovisual information, also in a semantic context. Importantly, this work demonstrated a clear relationship between neural and behavioral responses during gestural enhancement of degraded speech comprehension.

3.7. Acknowledgements

This work was supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from Netherlands Organization for Scientific Research. We are very grateful to Nick Wood, for helping us in editing the video stimuli, and to Gina Ginos, for being the actress in the videos.

Chapter 4

Alpha and beta oscillations index semantic congruency between speech and gestures in clear and degraded speech

4.1. Abstract

Previous work revealed that visual semantic information conveyed by gestures can enhance degraded speech comprehension, but the mechanisms underlying these integration processes under adverse listening conditions remain poorly understood. We used MEG to investigate how oscillatory dynamics support speech-gesture integration when integration load is manipulated by auditory (e.g., speech degradation) and visual semantic factors (e.g., gesture congruency). Participants were presented with videos of an actress uttering an action verb in clear or degraded speech, accompanied by a matching (mixing gesture + 'mixing') or a mismatching gesture (drinking gesture + 'walking'). In clear speech, alpha/beta power was more suppressed in LIFG, motor and visual cortex when integration load increased in response to mismatching versus matching gestures. In degraded speech, beta power was less suppressed over pSTS and MTL for mismatching compared to matching gestures, showing that integration load was lowest when speech was degraded and mismatching gestures could not be integrated and disambiguate the degraded signal. Our results thus provide novel insights on how low-frequency oscillatory modulations in different parts of the cortex support the semantic audiovisual integration of gestures in clear and degraded speech: When speech is clear, LIFG and motor and visual cortex engage because higher-level semantic information increases semantic integration load. When speech is degraded, pSTS/MTG and MTL are less engaged because integration load is lowest when visual semantic information does not aid lexical access.

This chapter is based on Drijvers, L., Ozyurek, A., & Jensen, O. (2018). Alpha and beta oscillations index semantic congruency between speech and gestures in clear and degraded speech. *Journal of Cognitive Neuroscience*, 30(8), 1086-1097. doi:10.1162/jocn_a_01301

4.2. Introduction

Oscillatory dynamics are thought to subserve the integration of complex information from multiple modalities (Varela, Lachaux, Rodriguez, & Martinerie, 2001) such as during multisensory integration (Kayser & Logothetis, 2009; Schepers et al., 2013; Schroeder et al., 2008; Senkowski, Schneider, Foxe, & Engel, 2008). Low-frequency oscillatory power decreases in the alpha and beta band are often related to the engagement of brain areas, whereas increases are often related to disengagement or functional inhibition of task-irrelevant brain regions (Jensen & Mazaheri, 2010; Klimesch et al., 2007; Pfurtscheller & Lopes da Silva, 1999). In line with this, previous research revealed that oscillatory power increases can predict the degree of *non-semantic* audiovisual integration of an ambiguous stimulus (e.g., beeps and flashes, Hipp, Engel, & Siegel, 2011). However, it is poorly understood how these mechanisms translate to *semantic* audiovisual integration, such as in multimodal speech processing.

Investigating whether similar oscillatory mechanisms also support more realistic situations is particularly relevant when considering face-to-face communication, which integrates auditory (e.g. speech) and visual (e.g. gestures) modalities. Under adverse listening conditions, speech comprehension can be enhanced by the visual semantic information conveyed by iconic gestures (Drijvers & Ozyürek, 2017; Holle, Obleser, Rueschemeyer, & Gunter, 2010). These iconic gestures can illustrate object attributes, actions and space (McNeill, 1992) and are known to affect clear and degraded speech comprehension (e.g., Dick, Mok, Raja Beharelle, Goldin-Meadow, & Small, 2014; Drijvers & Ozyürek, 2017; Drijvers & Ozyurek, in press; Drijvers, Ozyurek & Jensen, 2018; Green et al., 2009; Holle et al., 2010; Straube, Green, Weis, & Kircher, 2012; Willems, Özyürek, & Hagoort, 2007, 2009, Josse et al., 2012; see for a review; Ozyurek, 2014). For example, when the semantic information that is conveyed by these gestures mismatches clear speech, previous studies have demonstrated that semantic integration load increases and audiovisual integration might be hindered. For example, previous fMRI studies have demonstrated more BOLD activation in LIFG when semantic integration load increased and gestures mismatched compared to matched clear speech (Willems et al., 2007, 2009). Similar effects have been demonstrated in EEG studies, where

the N400, an ERP component that is sensitive to semantic unification operations, was more negative when gestures mismatched than matched clear speech (e.g., Kelly et al., 2004). Extending on this, recent work demonstrated that the difference in N400 amplitude (i.e., the N400 effect) in response to mismatching compared to matching gestures is larger in clear than in degraded speech, which indicated that listeners are more hindered when integrating gestures with degraded speech (Drijvers & Ozyurek, 2018). These results suggest that when speech is degraded, the mismatching gesture cannot aid to disambiguate the remaining auditory cues to facilitate speech comprehension and integration load is lowest. However, it is unknown what neural mechanisms underlie speech-gesture integration in clear and adverse listening conditions, and it is unknown how semantic integration occurs when the integration load is manipulated by auditory factors (e.g., speech degradation) and visual semantic factors (e.g., congruency of gestures). Therefore, the current study aims to get insight in what oscillatory mechanisms support the semantic integration of speech and gestures in both clear and degraded speech.

Studies on unimodal degraded speech processing have consistently demonstrated less suppressed alpha power as a function of speech intelligibility (i.e., enhanced alpha power in response to degraded speech), which has been interpreted as possibly reflecting the allocation of resources and the functional inhibition of task-irrelevant neural activity during speech comprehension in challenging listening situations. This might be due to a higher auditory cognitive load when language processing is inhibited due to speech degradation (Becker, Pefkou, Michel, & Hervais-Adelman, 2013; Drijvers, Mulder, & Ernestus, 2016; Obleser & Weisz, 2012; Obleser, Wöstmann, Hellbernd, Wilsch, & Maess, 2012; Strauß, Wostmann, & Obleser, 2014; Weisz & Obleser, 2013; Wilsch, Henry, Herrmann, Maess, & Obleser, 2014). During audiovisual processing of speech in noise, other work has revealed that beta power localized in the STS was less suppressed in high noise compared to no or low noise, possibly reflecting disturbed or altered audiovisual speech processing (Schepers et al., 2013). The abovementioned studies however do not include a visual semantic component, such as iconic co-speech gestures. In the visual domain, previous research on speech-gesture integration has identified that during gestural enhancement of degraded speech comprehension, low- and

high-frequency oscillatory power modulations in LIFG, left-temporal, motor and visual regions predicted a listener's benefit from gestures during degraded speech comprehension (Drijvers, Ozyurek, & Jensen, 2018). However, it is unknown how oscillatory activity supports speech-gesture integration when this integration is modulated by auditory (speech degradation) and visual semantic factors (gesture congruency). The spatiotemporal characteristics of this integration process are needed to reveal which brain areas are engaged and disengaged in this process over time, such as when integration load is increased and a gesture mismatches compared to matches clear speech, but also when integration load is lowest, such as when a gesture mismatched compares to matches degraded speech.

Using magnetoencephalography (MEG), we investigated the spatiotemporal oscillatory neuronal dynamics underlying audiovisual integration in a multimodal semantic context. Participants were presented with videos of an actress uttering an action verb in clear or degraded speech, accompanied by a matching or a mismatching gesture, following the design of Drijvers & Ozyurek (2018). Based on the oscillatory modulations that we observed in Drijvers, Ozyurek & Jensen (2018), we expected that the neural integration of speech and gesture relies on an extended network, involving the language network (incl. LIFG/pSTS/MTG), the motor cortex and the visual cortex. In line with the functional inhibition framework, our general hypothesis was that a relative decrease of alpha and beta power would reflect engagement of task-relevant brain regions, whereas enhanced alpha and beta power would reflect areas that do not need to be engaged for the task at hand, or are less engaged in one condition compared to another condition (Jensen & Mazaheri, 2010; Klimesch et al., 2007). In clear speech, we thus expected that when visual semantic congruency would increase integration load (i.e., when a gesture would mismatch compared to match the clear speech, see Drijvers & Ozyurek, 2018), alpha and beta power would be more suppressed for mismatching compared to matching gestures. We expect that this larger suppression would occur in the language network, as well as the visual and motor cortex, reflecting increased visual attention to mismatching compared to matching gestures (Stothart & Kazanina, 2013; Drijvers, Ozyurek & Jensen, 2018), a larger engagement of the motor system during observation of mismatching compared to matching gestures

(Caetano et al., 2007; Kilner et al., 2009; Koelewijn et al., 2008), and a higher semantic unification load (Drijvers & Ozyurek, 2018; Drijvers, Ozyurek & Jensen, 2018). This higher semantic unification load then occurs because the mismatching semantic information of the gesture is harder to integrate with clear speech than matching semantic information (Wang et al., 2012; Drijvers & Ozyurek, 2018). Even though we thus expect that mismatching gestures increase integration load in clear speech, we expect that in degraded speech mismatching gestures result in the lowest integration. In degraded speech, the gestural information cannot be coupled to the remaining auditory cues in the degraded speech signal, which would hinder integration (Drijvers & Ozyurek, 2018). This is opposed to matching gestures in degraded speech, which can enhance recognition of degraded speech (as for example in Drijvers & Ozyurek, 2017; Holle et al., 2010). Therefore, in degraded speech, we expect that alpha and beta power will be less suppressed when a gesture mismatches compared to matches degraded speech, reflecting less engagement of task-relevant brain regions during speech-gesture integration.

4.3. Methods

4.3.1. Participants

Thirty-two right-handed Dutch native participants, recruited from Radboud University (mean age = 23,2, SD = 3,46, 14 males) participated in this experiment. All participants had normal hearing and normal or corrected-to-normal vision and no language, motor or neurological disabilities and gave written consent before participating in this experiment. Three participants (two female) were excluded due to technical failure (one) and excessive (head) motion artifacts (movement > 1 cm / > 60% of the trials affected). The final dataset included the data of twenty-nine participants.

4.3.2. Stimulus materials

Participants were presented with 240 short video clips of a female actress who uttered a Dutch action verb, which would be accompanied by a matching iconic gesture, a mismatching iconic gesture or no gesture. In the current chapter, we

report the results of a subset of this experiment, containing 160 video clips that either contained a matching iconic gesture or a mismatching gesture, to zoom in on the effect of 'integration load', by both manipulating auditory factors (by speech degradation), and visual semantic factors (by gesture congruency). Participants were thus presented with 160 video clips that contained an actress who uttered a highly-frequent action verb in clear or degraded speech, accompanied by a matching or a mismatching iconic gesture. All of these video clips were pretested as part of Drijvers & Ozyurek (2017). To ensure the verbs would fit with the gestures, we presented participants with the videos without their audio file and asked them to write down the verb they associated with the movement. We then showed participants the verb we originally matched the video with, and asked them to indicate on a 7-point scale how much this verb fitted with the movement in the video. The results revealed a mean recognition rate of 59% over all gesture videos, which indicates that the gestures are potentially ambiguous without speech, and thus might need speech for successful comprehension. Our rating task resulted in a mean score of "iconicity" of 6.1 (SD = 0.64), and all videos that scored under 5 on a 7-point scale were discarded.

In all videos that were used in this experiment (see Figure 12), the actress would always appear in the middle of the screen, where she was visible from her knees upwards. She wore neutrally-coloured clothing and was standing in front of a dark blue neutral background. The gestures that she made were not instructed but made by her on the fly. The actress did not get any feedback on the gestures she made. For the mismatching gestures, the experimenter would mention two verbs to the actress, of which the first one had to be the spoken verb, and the second one the to-be-gestured verb (e.g., 'to drive' and 'to mix', uttering 'drive' while making a mixing gesture). This method was chosen as our stimuli show the face of the actress, and we could therefore not replace the audio track of the video with another verb's audio track, as the visible speech would be different. To determine which verbs were used as mismatching gesture, we divided the list of verbs in the mismatching condition in two separate lists, and combined the verbs on the first list with the gesture that matched the verb on the second list. In all videos, the preparation of this gesture (counted as the first frame where the actress

moved her hand), was at 120 ms. At 550 ms, the stroke of the gesture would occur. Speech onset was at 680 ms, and the retraction of the gesture started at 1380 ms. The gesture offset was at 1780ms (See Figure 12B). As speech onset was at 680 ms and stroke onset at 550 ms, the overlap between the meaningful part of the gesture and the speech was optimal for mutual enhancement for comprehension (as previously demonstrated in Habets, Kita, Shao, Ozyurek, & Hagoort (2011)).

All audio-files were presented in clear speech or 6-band noise-vocoded speech. This noise-vocoding level was chosen as previous work showed that at a 6-band noise-vocoding level, participants are most able to use gestural information for comprehension (Drijvers & Ozyürek, 2017). From the video files, we extracted all audio-tracks, de-noised the speech and intensity-scaled the speech to 70 dB by using Praat (Boersma & Weenink, 2015). After degrading 80 out of the 160 sound files, all sound files were then recombined again with their corresponding video files, by using a custom-made script in *Praat*. To degraded the speech signals, we band-pass filtered each audio file between 50 Hz and 8000 Hz divided the speech signal in logarithmically spaced frequency bands with cutoff frequencies at 50 Hz, 116.5 Hz, 271.4 Hz, 632.5 Hz, 1473.6 Hz, 3433.5 Hz and 8000 Hz. These frequencies were used to filter white noise in order to obtain the six bands. Subsequently, the amplitude envelope of these bands were extracted using half-wave rectification. We then multiplied this envelope with the noise bands and recombined the bands, resulting in the degraded signal (Shannon et al., 1995). All sound was presented to participants through MEG-compatible air tubes. The total experiment consisted of four conditions: a clear speech + matching gesture condition (CM), a degraded speech + matching gesture condition (DM), a clear speech + mismatching gesture condition (CMM) and a degraded speech and mismatching gesture condition (DMM). Each condition consisted of 40 items, of which none were repeated in any other condition (see Figure 12A).

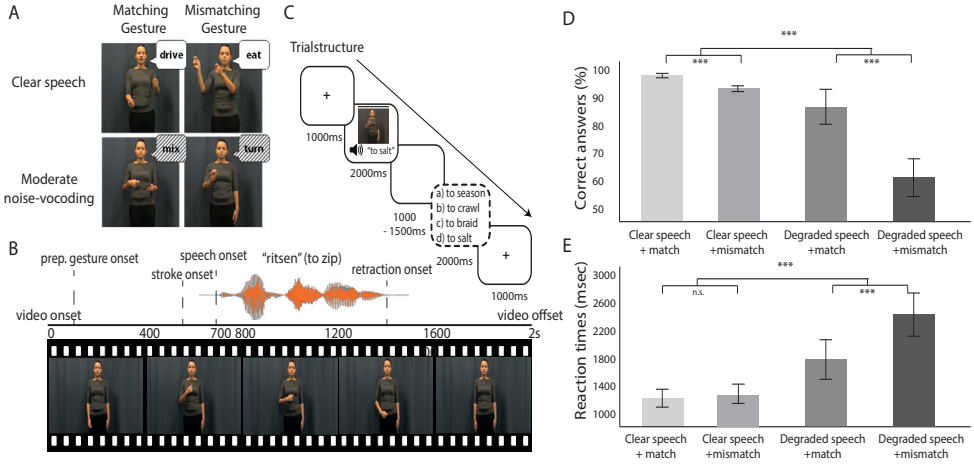


Figure 12 A: Illustration of the different conditions and stimuli. B: Illustration of the structure of the videos. C: Structure of trial. D: Upper panel: percentage of correct answers per condition. Error bars represent SD. *** = $p < .01$. Lower panel: reaction times in msec per condition.

4.3.3. Procedure

Participants were placed in the 275-channel axial gradiometer CTF MEG system, at 70 cm from the projection screen on which the videos were presented. All videos were projected full-screen onto a semi-translucent screen by back-projection using an EIKI LC-XL100L projector at a resolution of 1650 x 1080 pixels. The experiment was presented through Presentation software. Each trial would start with a fixation cross (1000 ms), followed by a video (2000 ms), a short delay period (1000 – 1500 ms, jittered) and ended with a 4-alternative forced choice identification task in which participants had to identify which verb they just heard in the videos. The answer options in this task would always consist of a semantic competitor, a phonological competitor, an unrelated answer and the correct answer. Participants had to indicate their choice by pressing a button with their right hand on a 4-buttonbox. After the participants had entered their response, a new trial would start after 1500 ms. (see Figure 12C). All participants were presented with an individual pseudo-randomization of the different videos that ensured that none of the conditions would occur more than twice in a row (e.g., two consecutive trials that had degraded speech and a mismatching gesture).

Participants were asked to sit as still as possible and not to blink during the videos, but after answering the 4-alternative forced choice identification task. We measured brain activity with MEG throughout the entire experiment. Participants were able to take a self-paced break per 40 trials.

4.3.4. MEG data acquisition

Whole-head MEG was recorded at a sampling rate of 1200 Hz by using a 275-channel axial gradiometer MEG system. Participants wore recording markers on the nasion and left and right ear canal to monitor their head position in real time, using a MATLAB toolbox (Stolk et al., 2013). During the breaks, this allowed us to readjust the subjects head position relative to the original position at the start of the experiment if the deviation was larger than 5 mm. We recorded electrocardiogram (ECG) as well as horizontal and vertical electrooculogram (EOG) for artifact rejection purposes. After the experiment, we invited the participants to record a structural magnetic resonance image of their brain, using a 1.5 T Siemens Magnetom Avanto system with markers attached in the same position as the head coils, to allow us to align the structural anatomy of the participants with the MEG coordinate system. We collected structural MRI's for 22 out of 32 participants.

4.3.5. MEG data analysis

All data in this experiment was analyzed by using FieldTrip (Oostenveld et al., 2011), an open-source MATLAB toolbox and custom Matlab scripts. We preprocessed the data by dividing the data in epochs from -1 s before video onset until 3 s after video onset. All data was demeaned and detrended, and line noise was attenuated by using a discrete Fourier transform approach at 50 Hz and its subsequent harmonics. In total, we rejected on average ~3 trials per condition, which were contaminated by SQUID jump artifacts and muscle artifacts by using a semi-automatic routine. We then applied independent component analysis to remove eye-movements and cardiac related activity (Bell & Sejnowski, 1995; Jung et al., 2000) to remove all remaining eye-movements and cardiac-related activity. Finally, we went through all single trials and removed any artifacts that were not

identified by using ICA or other rejection procedures. We then resampled the data to 300 Hz to speed up analyses.

We computed an approximation of the planar gradient by converting the axial gradiometer data to orthogonal planar gradiometer pairs, and computed and summed the power of the pairs. This approach might facilitate the interpretation of the MEG data, as planar gradient maxima are known to be located above the neural sources that might underlie an effect (Bastiaansen & Knösche, 2000).

4.3.6. Time-frequency analyses

Our frequencies of interest ranged from 2 Hz to 30 Hz, in frequency steps of 1 Hz. We applied a 500 ms Hanning window in 50 ms time steps. (Mitra & Pesaran, 1999). To calculate the differences between conditions, we compared oscillatory power by averaging the four conditions separately for each participant. TFRs were log10 transformed and the difference between the conditions was calculated by subtracting the log10 transformed power (= 'log ratio', e.g. log10(A) - log10(B), or, log10(CMM) – log10(CM) and log10(DMM) – log10(DM)). The time window of analysis was always between 0.7 – 2.0 s, which corresponds to speech onset until video offset.

4.3.7. Source analyses

To estimate the sources of our observed effects, we used dynamic imaging of coherent sources (DICS, Gross et al., (2001)) as a beamforming spatial filtering technique. For this part of the analysis, the axial gradiometer data was used. First, the algorithm computed a common spatial filter from the cross-spectral density matrix of the data and a lead field. For all frequency ranges of interest we used a single Hanning taper. All leadfields of the participants were constructed by using a realistically shaped single-shell head model based on the participants' own individual anatomical data and by identifying the anatomical markers at the nasion and the two ear canals. Each volume was then divided into a 10 mm spaced grid of points and warped to the template MNI brain, where the lead field was calculated for each grid point.

The time windows that were used as input for the source-analysis were based on the results from the sensor analysis. For the alpha band, we calculated the CSD between 1.3 - 2.0 s at 10 Hz, with 2 Hz frequency smoothing. For the beta band, we computed the CSD between 1.3 – 2.0, centered at 18 Hz with 4 Hz frequency smoothing. We used a common spatial filter containing all of the conditions to project the data through, separately per condition. We then averaged over trials; log-10 transformed the data and calculated the difference between conditions by subtracting the log power for the single contrasts. Finally, the grand-average grid of all participants was interpolated onto the template MNI for visualization purposes. Note that we included all trials in our sensor-level and source-level analyses, and did not differentiate between correct and incorrect trials, as the 4-alternative forced choice identification task might have masked the actual comprehension participants might have had when they were watching and listening to the video.

4.3.8. Cluster-based permutation statistics

We performed non-parametric cluster-based permutation tests (Maris & Oostenveld, 2007) across subjects to statistically quantify differences between the different conditions in power on source and sensor-level. We used the sensor-level statistics to create statistical threshold masks to localize the observed effects on source-level. We computed the mean difference between two conditions (e.g., CMM vs CM or DMM vs DM) for each x/y/z/ sample of our dataset in the frequency ranges (alpha: 8-12 Hz, beta: 15-20 Hz) and time window (0.7 - 2.0 s, i.e., from speech onset until the end of the video) we defined a priori and on the basis of a grand-average TFR of all conditions combined. After collecting all of the difference values of these comparisons (e.g. CMM vs CM or DMM vs DM), all values where thresholded with the 95th percentile of the entire distribution. The remaining values formed the cluster candidates. All conditions and their corresponding values were randomly reassigned 5000 times to form the permutation distribution. Out of this distribution, the cluster candidate who had the highest sum of the difference values was added to the permutation distribution. Finally, the actual observed cluster-level summed values were compared against this distribution, and all clusters that fell in the highest or lowest 2.5% were

considered significant.

4.4. Results

We presented participants with videos that showed an actress uttering a Dutch action verb, while she simultaneously made a matching or a mismatching gesture. Subsequently, participants had to indicate which verb they heard by a button press. Brain activity was recorded by MEG throughout the whole trial, but we focused on the time window from speech onset (0.7s) until the end of the video.

4.4.1. Behavioral results

A repeated-measures ANOVA with the factors Gesture (matching/mismatching) and Noise (clear/degraded) revealed that when speech was clear, participants were more able to identify a correct answer on the 4-alternative forced choice identification task than when speech was degraded ($F(1,28) = 94.97$, $p < 0.001$, $\eta^2 = .77$. Similarly, participants found it easier to identify a word when a gesture matched rather than mismatched the speech signal ($F(1,28) = 72.77$, $p < 0.001$, $\eta^2 = .72$, see Figure 12D). An interaction effect between Gesture and Noise confirmed that the difference in correct answers when comparing mismatching to matching gestures was larger in degraded speech than in clear speech ($F(1,28) = 58.45$. $p < 0.001$, $\eta^2 = .68$; CM: 97.2%, SD = 1.6%; CMM = 92.8%, SD = 2.1%; DM = 85.6%, SD = 12.1%; DMM = 61.4%, SD = 11.2%). Post-hoc t-tests on the relevant contrasts confirmed that participants were more able to correctly identify the verb when the verb was accompanied by a matching compared to a mismatching gesture (clear speech; $t(28) = -3.09$, $p < 0.01$, degraded speech; $t(28) = -8.42$, $p < 0.001$). We did not observe any reliable differences in the number of error responses that were selections of the semantic or phonological competitors.

A second repeated-measures ANOVA using the same factors revealed a similar pattern for the reaction times as for the correct answers: participants were quicker to answer when speech was clear compared to degraded ($F(1,28) = 143.63$, $p < 0.001$, $\eta^2 = .84$), and when a gesture matched compared to mismatch the speech signal ($F(1,28) = 59.90$, $p < 0.001$, $\eta^2 = .68$), see Figure 12E). The difference in reaction times when comparing mismatching to matching gestures was larger in

degraded speech than in clear speech ($F(1,28) = 46.40$, $p < 0.001$, $\eta^2 = .62$; CM = 1269.1 SD = 360.3 ; CMM = 1299.8, SD = 378.0; DM = 1849.9, SD = 578.5 DMM = 2492.4, SD = 673.5). Post-hoc t-tests on the relevant contrasts confirmed that participants were not quicker to identify the verb when the verb was accompanied by a matching compared to a mismatching gesture in clear speech $t(28) = .82$, $p = 0.41$), but were quicker to identify the verb when the verb was accompanied by a matching compared to a mismatching gesture in degraded speech ($t(28) = 8.02$, $p < 0.001$). These behavioral results reveal that gesture facilitates comprehension of degraded speech when the actress made a matching gesture, but hindered comprehension when she performed a mismatching gesture.

4.4.2. Semantic congruency effects in clear speech

4.4.2.1. Alpha and beta power are more suppressed when a gesture mismatches than matches clear speech.

We first conducted a sensor-level analysis over the full time window (0.7 – 2.0, from speech onset until video offset), to identify differences in oscillatory power between the conditions. We calculated the time-frequency representations (TFRs) of power for the individual trials, and averaged them per condition For TFRs of the single conditions, please see Figure S4 in Supplementary Materials S4. Figure 13A represents the TFRs of power in response to the contrast CMM vs CM between 2 and 30 Hz, at representative left-temporal sensors, based on the topographical plots that visualize this effect in time and space (see Figure 13B). Sensor-level analyses confirmed a larger alpha and beta power suppression over left-temporal, motor and occipital areas when speech was clear and a gesture mismatched compared to matched the speech signal (alpha; one negative cluster, $p = .04$, 1.3 - 2.0s, beta; one negative cluster, $p < .01$, 1.3 - 2.0s) suggesting engagement of these areas in response to the mismatching gesture.

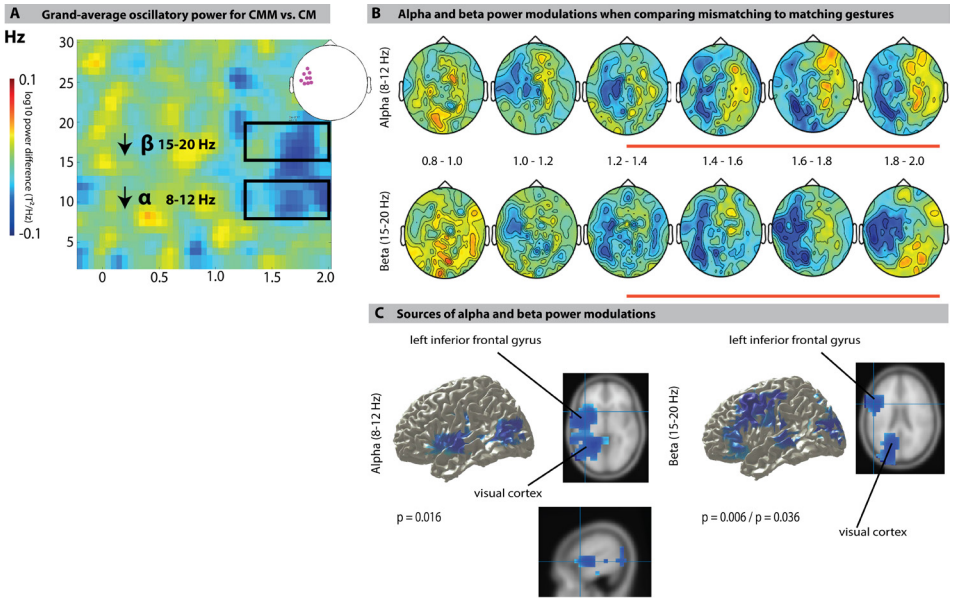


Figure 13 A) Time-frequency representation (TFRs) of power of the contrast between clear speech + mismatching gesture (CMM) vs. clear speech+match gesture (CM). B) Topographical distribution of alpha (upper) and beta (lower) power of the contrast CMM vs CM in 200 ms time bins. Orange bars denote significant clusters in the sensor-level analyses. C) Estimated source results of the contrast in the alpha (left) and beta (right) band, masked by statistically significant clusters.

4.4.2.2. Alpha power is more suppressed in LIFG, left insula and visual cortex when a gesture mismatches than matches clear speech.

We used the time-window of the significant clusters from the sensor analyses as input for our source analyses to estimate the sources of the alpha power modulation. Note that the statistical assessment was based on the sensor analysis, not the source-level analysis.

Nevertheless, we applied a cluster-randomization approach to the source data to find a threshold for when to consider the source estimates reliable. To investigate these underlying sources, we used a frequency-domain spatial beamformer technique (DICS, Gross et al., 2001). This analysis revealed that the source of the larger alpha power suppression in response to mismatching compared to matching gestures was localized in a widespread cluster including the LIFG, left insula and the visual cortex (one negative cluster, p = .04). These

results thus suggest engagement of the extended language network when a gesture mismatches clear speech.

4.4.2.3. Beta power is more suppressed in motor, visual regions and LIFG when a gesture mismatches compared to matches clear speech.

We then localized the sources of the sensor-level power difference in the beta band. We localized the beta power difference in the left pre- and postcentral gyrus, the left frontal midline/supplementary motor area, LIFG and the visual cortex (two negative clusters, $p < .01$, $p < .04$). In line with our hypotheses and earlier work (Drijvers, Ozyurek & Jensen, 2018), this larger beta power suppression over the motor cortex shows that listeners might engage their motor cortex more when a gesture mismatches compared to matches the clear speech signal.

4.4.3. Semantic congruency effects in degraded speech

4.4.3.1. Beta power is more enhanced when a gesture mismatches than matches degraded speech.

Next we investigated whether a similar pattern of oscillatory power modulations would emerge when we compared the same conditions in degraded instead of clear speech. For TFRs of the single conditions, please see Figure S1B in Appendix II. The TFR in Figure 14A suggests enhanced beta power, but no differences in alpha power. We plotted the topographical distribution of the contrast between mismatching and matching gestures in both frequency bands (see Figure 14B). We found no difference in alpha band power when comparing matching and mismatching gestures in degraded speech (no significant clusters, $p = 0.06$, see Figure 14C), but we did observe larger beta power over left-temporoparietal areas when a gesture mismatched compared to matched degraded speech (one positive cluster, $p < 0.001$, Figure 14B/C). Due to the lack of an alpha power difference in DMM vs DM, the difference in CMM vs CM was greater than the difference in alpha power in DMM vs DM (one positive cluster, $p = .012$). The difference in beta power between CMM vs CM was larger than in DMM vs DM (one positive cluster, $p = .004$).

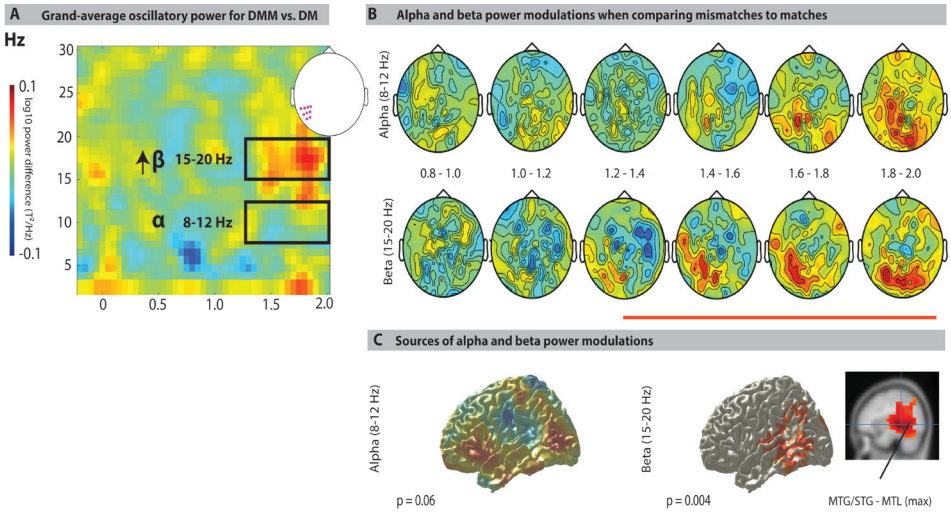


Figure 14 A) Time-frequency representation (TFRs) of power of the contrast between degraded speech + mismatching gesture (DMM) vs. degraded speech+match (DM) gesture. B) Topographical distribution of alpha (upper) and beta (lower) power of the contrast DMM vs DM in 200 ms time bins in our time-window of interest. Orange bars denote significant clusters in sensor-level analyses. C) Estimated source of the contrast in the alpha (left) and beta (right) band, masked by statistically significant clusters. Note that in the beta band this effect was not statistically significant, but the estimated sources of the difference are included for visualization purposes.

4.4.3.2. Enhanced beta power inhibits STS and MTL when a gesture mismatches degraded speech.

We localized the enhanced beta power in response to mismatching compared to matching gestures in degraded speech in the left auditory cortex, superior temporal sulcus, middle temporal gyrus, and medial temporal lobe (one positive cluster, $p < .01$).

4.5. Discussion

We investigated how oscillatory dynamics support the semantic integration of speech and gestures in clear and degraded speech, and what the spatiotemporal dynamics are that are associated with speech-gesture integration. We manipulated semantic integration load by presenting participants with videos of an actress who uttered an action verb in clear or degraded speech, accompanied by a matching or

mismatching gesture. Our behavioral results demonstrated a semantic congruency effect and speech degradation effect on performance; participants were slower and less able to correctly identify the verb when gestures mismatched speech and speech was degraded. These results replicate previous findings and underline the additive effect of speech degradation (e.g., Holle et al., 2010) and semantic congruency between speech and gestures (e.g., Willems et al., 2007; 2009; Ozyurek et al., 2007; Kelly et al., 2004; Drijvers & Ozyurek, 2018) on integration load and subsequently, behavioral performance.

Our neurophysiological results demonstrate that semantic congruency and speech degradation modulated oscillatory activity in the alpha and beta band. When speech was clear, we observed a larger alpha power suppression over LIFG and visual cortex, and a beta suppression over LIFG, (pre)motor cortex, and visual cortex when a gesture mismatched compared to matched speech. When speech was degraded, we observed no difference in alpha power when comparing degraded speech and a mismatching gesture to a matching gesture. However we did observe enhanced beta power over pSTS when a gesture mismatched compared to matched degraded speech. In both the alpha and the beta band, we observed a larger difference between mismatching and matching gestures in clear than in degraded speech, suggesting that integration load was lowest in degraded speech (in line with Drijvers & Ozyurek, 2018).

4.5.1. Alpha/beta power is more suppressed over visual cortex to allow for increased visual attention to mismatching compared to matching gestures during clear speech

Both alpha and beta power were more suppressed over visual regions when a gesture mismatched compared to matched clear speech. This effect occurred from when the meaningful part of the gesture and speech were unfolding until the end of the video (1.3 - 2.0s). The larger alpha/beta suppression over visual regions suggests that the visual system is more engaged when a listener observes a mismatching gesture than a matching gesture, and that more visual attention is allocated to a mismatching gesture compared to a matching gesture. We suggest that when all auditory cues are still intact, a mismatching gesture will

generate a larger mismatch response, causing increased visual attention to these mismatching gestures compared to matching gestures as a result of the detection of mismatching semantic information. Similar results have been found by Stothart & Kazanina (2013), who reported a post-stimulus alpha suppression for deviant visual stimuli, potentially reflecting a shift in attentional resources following the detection of change. Note that the loci of the clusters in the alpha and the beta band slightly seem to differ: the maximum of the cluster in the beta band can be localized to BA18, whereas the maximum of the alpha cluster is estimated in BA19. This suggests that the observed beta effect is not simply a harmonic of the observed alpha activity.

4.5.2. Alpha/beta power is more suppressed over LIFG due to increased semantic unification load in clear speech

We observed a larger suppression of alpha and beta power when a gesture mismatched compared to matched clear speech. This larger suppression in response to a mismatching gesture was localized in the LIFG. Previous studies have proposed that the LIFG is sensitive to unification operations from units that are retrieved from memory, the unification of information from different modalities, and lexical access operations (Hagoort, 2013). For example, in a study on unimodal speech comprehension, sentences with incongruent sentence endings yielded larger beta power over LIFG. This was interpreted as reflective of a higher semantic unification load that was evoked by the incongruent sentence endings, which required a stronger engagement of the task-relevant brain network (Wang, Jensen, Van den Brink, et al., 2012). Similarly, we demonstrated in a previous study that alpha/beta power is more suppressed in LIFG when integration load increases (Drijvers, Ozyurek & Jensen, 2018). In line with this work, we suggest that the larger alpha and beta power suppression over LIFG is reflective of the larger engagement that is required from LIFG when a mismatching gesture increases semantic unification load to resolve the mismatch between the auditory information and the visual semantic information.

Note that previous studies (e.g., Dick et al., 2014; Green et al., 2009; He et al., 2015; Holle et al., 2010; Straube, Green, Bromberger, & Kircher, 2011; Willems et

al., 2009) have discussed the role of the LIFG in speech-gesture integration, and that earlier work has argued that LIFG is sensitive to the semantic relationship of speech-gesture pairs when a new and unified representation of the gestural input and speech needs to be constructed (Willems et al., 2009), which is the case for incongruent gestures. This interpretation was later extended by Holle et al., (2010) who argued that LIFG activity reflects a modulation or a revision of the integrated speech-gesture information. This interpretation partially fits our findings. When post-hoc visualizing the oscillatory modulations in the single conditions, we observed that LIFG revealed suppressed activity compared to baseline in all conditions. The contrast between the mismatching and matching gestures thus shows that this suppression is larger when a mismatching gesture is paired with clear speech than when a matching gesture is paired with clear speech. However, the unification of these movements with the speech signal involved engagement of the LIFG in all single conditions (i.e., degraded speech+matching gesture, degraded speech+mismatching gesture, clear speech+matching gesture, clear speech+mismatching gesture). This suggests that LIFG possibly has a unifying function of the different inputs irrespective of congruency, but that an increased integration load also increases engagement of the LIFG to unify the inputs.

4.5.3. Motor beta suppression reveals stronger simulation of mismatching gesture in clear speech

Beta power was more suppressed over pre-central cortex and supplementary motor area when a gesture mismatched compared to matched clear speech. This effect occurred in a similar time-window as the alpha modulation (1.3 - 2.0s) and lasted from when the speech and gesture were unfolding until the end of the video. The larger suppression for mismatching compared to matching gestures suggests that engagement of the motor system is modulated by the semantic fit of the information that is conveyed by the gestures, which is in line with previous studies on action observation (e.g., Klepp et al., 2015; Schaller, Weiss, & Müller, 2017; van Elk et al., 2010; Weiss & Mueller, 2012). We interpret this effect as showing that listeners more strongly engage their motor system to 'simulate' the mismatching gesture to re-evaluate whether it fits with the processed speech

signal. Note that we did not observe a similar effect when speech was degraded. This suggests that when speech is degraded, matching and mismatching gestures are simulated equally when auditory cues are not reliable and a re-evaluation of the fit of the gesture is hindered. This would be in line with current and previous work that suggests that integration load is lowest when speech is degraded (Drijvers & Ozyurek, 2017).

4.5.4. Enhanced beta power over STS/MTG and MTL reveals hindered semantic integration and lexical access when gestures mismatch degraded speech

When speech was degraded, we did not observe reliable differences in alpha power when comparing mismatching gestures and matching gestures. However, beta power was less suppressed in response to mismatching compared to matching gestures (1.3 - 2.0s) when speech was degraded. This is in line with previous work on non-semantic audiovisual speech processing, that demonstrated a similar smaller beta suppression in noisy speech when comparing audiovisual to audio-only conditions. This effect was localized to the STS (Schepers et al., 2013). This underlines that modulations of oscillatory activity in the STS play a role in audiovisual speech processing under clear and adverse listening conditions. Previous studies have shown that suppressed beta band activity plays a role in tasks where information from different modalities needs to be integrated (Kopell, Kramer, Malerba, & Whittington, 2010), and in naturalistic audiovisual processing (Kayser & Logothetis, 2009). When speech is degraded and the semantic information that is conveyed by the gesture cannot be matched to the degraded auditory cues, pSTS/MTG might be less engaged because of the hindered audiovisual integration. Similarly, as the meaningful information from a mismatching gesture will not aid in resolving the degraded speech signal, lexical retrieval might be hindered (Hannemann et al., 2007), which is demonstrated by less involvement of the MTL when listeners process mismatching as compared to matching gestures.

Our current results also contribute to recent discussions over the role and involvement of pSTS/MTG and LIFG in speech-gesture integration. Although

the role of pSTS and LIFG has been discussed, MTG has often been found to be involved in speech-gesture integration. Some studies have shown that MTG is modulated by semantic congruency (Dick, Goldin-Meadow, Solodkin, & Small, 2012; Green et al., 2009, see Ozyurek, 2014) and activity in the MTG has been linked to coupling sound and meaning. However, the role of (p)STS has been debated. Some studies have argued that STS is sensitive to semantic aspects of speech-gesture integration. For example, in an fMRI study, stronger activation for ambiguous words that were paired with iconic (i.e., semantic) compared to grooming gestures (non-semantic) was observed (Holle et al., 2008). Moreover, a larger involvement of the (p)STS has been reported in response to congruent iconic gestures coupled with degraded speech compared to clear speech (Holle et al., 2010), but not when comparing complementary versus redundant gestures (Dick et al., 2012). Other studies have argued that pSTS is mostly involved in the mapping and coupling of lower-level audiovisual information, which might already have a stable common object representation, but not to semantic congruency in speech-gesture integration (e.g. Willems 2007; 2009, Dick et al., 2012). Similarly, studies on audiovisual integration (e.g., lips and speech) have suggested that the STS might be related to associating the auditory and visual modalities at a lower level stage of multimodal matching (e.g., Beauchamp, 2005; Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp, Lee, Argall, & Martin, 2004; Callan et al., 2004; Calvert, 2001). Our current results suggest that indeed pSTS is sensitive to hindered audiovisual integration, but that this is not solely caused by semantic congruency. Note that although we observed a difference in oscillatory power in pSTS/MTG in degraded speech when comparing mismatching to matching gestures, we did not observe a modulation of oscillatory activity in pSTS/MTG when speech was clear. This suggests that pSTS/MTG is less engaged when speech is degraded. This might occur because integration processes are hindered when the visual semantic information cannot help to retrieve or disambiguate the degraded lexical item, which increases integration load. We thus tentatively propose that LIFG and pSTS indeed work together to integrate speech and gestures (cf. Willems et al., 2009), but that the role of LIFG is not solely modulatory or revising in nature (see e.g., Willems et al., 2009; Holle et al., 2010). Instead, LIFG unifies higher-level semantic information from multiple

inputs, irrespective of whether a stable common representation exists on which the input streams can be mapped (see Willems et al., 2009; Holle et al., 2010). However, when speech is degraded and a gesture mismatched speech, integration load was lowest when the gesture could not be integrated and disambiguate the degraded signal, resulting in less engagement from MTL and lower-level areas such as the pSTS/MTG.

4.6. Conclusion

The present work is the first study that investigated how oscillatory modulations can inform us about the processes underlying speech-gesture integration in clear and degraded speech, as well as what the spatiotemporal dynamics are that are associated with this process. We set out to investigate how the semantic integration of speech and gestures is supported when integration load that is manipulated by auditory (e.g. degraded speech) and visual (e.g. gesture congruency) factors. Our results provide novel insight by revealing how low-frequency oscillations support semantic audiovisual integration in clear and degraded speech: when gestures mismatch clear speech, listeners engage LIFG, motor and visual regions when semantic unification load increases due to the gesture. When speech is degraded, pSTS/MTG and MTL are less engaged, possibly reflecting the hindered integration of gestures and the degraded signal when the gesture does not disambiguate the degraded speech or aid lexical access. Our results thus reveal that low-frequency oscillatory modulations can index congruency between speech and gestures in a semantic context. Our results demonstrate that low-frequency power modulations do not only support *non-semantic* audiovisual integration, but also *semantic* integration. This suggests a domain-general mechanistic role of brain oscillations in enabling integration of different modalities and engaging/inhibiting brain areas that do not contribute to this integration process.

4.7. Acknowledgements

This work was supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from Netherlands Organization for Scientific Research. OJ was supported by James S. McDonnell Foundation Understanding Human

Cognition Collaborative Award (220020448) and the Royal Society Wolfson Research Merit Award. We are very grateful to Nick Wood (†), for helping us in editing the video stimuli, and to Gina Ginos, for being the actress in the videos.

Chapter 5

Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension

5.1. Abstract

Native listeners benefit from both visible speech and iconic gestures to enhance degraded speech comprehension (Drijvers & Ozyürek, 2017). We tested how highly-proficient non-native listeners benefit from these visual articulators compared to native listeners. We presented videos of an actress uttering a verb in clear, moderately or severely degraded speech, while her lips were blurred, visible, or visible and accompanied by a gesture. Our results revealed that unlike native listeners, non-native listeners were less likely to benefit from the combined enhancement of visible speech and gestures, especially since the benefit from visible speech was minimal when the signal quality was not sufficient.

This chapter is based on Drijvers, L., & Ozyurek, A., (2019). Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language & Speech.* doi.org/10.1177/0023830919831311

5.2. Introduction

As a non-native listener, understanding speech can be challenging, especially under adverse listening conditions. Previous research has shown that for native speakers, speech comprehension in adverse listening conditions can be enhanced by iconic gestures (Drijvers & Ozyürek, 2017; Holle et al., 2010; Obermeier, Dolk, & Gunter, 2011). Iconic gestures can be described as hand movements illustrating object attributes, actions and space (e.g., McNeill, 1992). Similarly, phonological cues conveyed by visible speech, consisting of information conveyed by lip movements, tongue and teeth can enhance comprehension in adverse listening conditions (Ross et al., 2007; Sumby & Pollock, 1954). However, it is unknown how much non-native listeners can benefit from these visual semantic and phonological cues, or how these visual articulators interact to enhance non-native speech comprehension in clear and adverse listening conditions.

There are several reasons why the processing of information that is conveyed by visual articulators could differ in non-native listeners compared to native listeners. For example, non-native listeners might try to focus more heavily on the information that is conveyed by visual articulators to compensate for their poorer comprehension skills in the non-native language. Previous work has indeed shown that visual articulatory information conveyed by visible speech improves non-native language learning and comprehension (Hannah et al., 2017; Hazan et al., 2006; Jongman, Wang, & Kim, 2003; Kawase, Hannah, & Wang, 2014; Kim & Davis, 2014; Summerfield, 1983; Wang, Behne, & Jiang, 2008, 2009) This could especially be relevant when phoneme perception (e.g. the difference between confusable phonemes as /æ/ and /ɛ/), which is thought to be especially challenging for non-native listeners, impedes comprehension (Broersma & Cutler, 2011). Additionally, previous work indicated that both facial (i.e., visible speech) and especially gestural input (i.e., hand gestures) can help non-native tone perception, especially when phonetic demands are high, such as in noise (Hannah et al., 2017). However, most work has focused on gestures improving non-native pitch perception, tonal distinctions, and intonational patterns (Hirata & Kelly, 2010; Hirata, Kelly, Huang, & Manansala, 2014; Kelly, Bailey, & Hirata, 2017; Kelly, Hirata, Manansala, & Huang, 2014; Morett & Chang, 2015). It remains unknown

how gestural semantic information enhances speech comprehension on top of visual speech in non-native listeners on a word level, and how this compares to native listeners, especially when speech is degraded.

Studies on non-native gesture processing have demonstrated that in clear speech, iconic gestures enhance non-native language comprehension and improve non-native language learning (e.g., Dahl & Ludvigsen, 2014; Hardison, 2010; Kelly, McDevitt, & Esch, 2009; Macedonia & Kriegstein, 2012; Sueyoshi & Hardison, 2005). For example, Kelly et al., (2009) have demonstrated that at initial stages, non-native listeners learn and remember novel words better when instructed with iconic gestures. In spite of this possible benefit, it is also plausible that non-native listeners might not be able to benefit from the semantic information that is conveyed by gestures on top of visible speech, especially when speech is unclear. Previous behavioral work on unimodal auditory speech processing has suggested that non-native listeners might not be able to use sentential semantic context to resolve phonemic information loss when signal quality is not clear enough. For instance, Bradlow & Alexander (2007) presented native and non-native listeners with sentences ending in low/high predictable words in plain or clear speech. Non-native listener's comprehension of sentence-final words was only improved when both semantic and acoustic information were available (i.e., when the sentence was highly predictable and produced in clear speech), whereas native listeners could benefit from both semantic and acoustic information in combination or separately (Bradlow & Alexander, 2007; Golestani, Rosen, & Scott, 2009; Oliver, Gullberg, Hellwig, Mitterer, & Indefrey, 2012). This suggests that non-native listeners might not be able to benefit from both visual articulators when speech is unclear and auditory information is insufficient to combine the information conveyed by these two visual articulators with the auditory input.

Previous literature thus suggests that non-native listeners might utilize visual semantic and phonological cues differently than native listeners, especially in adverse listening conditions, but the contribution of these two types of visual information to speech comprehension has not been studied in a joint context yet. A naturally following question is then whether non-native listeners can benefit from visual semantic and phonological cues in an multimodal context

in a similar way as native listeners do, but also whether and how these cues interact with phonological cues that are conveyed by visible speech. The aim of our study was therefore two-fold: we asked whether and to what extent visible speech contributes to the enhancement of degraded speech comprehension for non-native listeners, but also whether non-native listeners experience an additive benefit from semantic information conveyed by gestures on top of enhancement of visual speech. Additionally, we aimed to compare these results to data that we have collected in previous work on native listeners (Drijvers & Ozyurek, 2017).

In Drijvers & Ozyurek (2017), we presented participants with videos of an actress uttering a clear or degraded (2- or 6-band noise-vocoding) verb while her lips were blurred, visible speech was present, or visible speech and a gesture were present. We demonstrated that native listeners benefit most from having both visible speech and an iconic gesture present as compared to having just visible speech present, or seeing the actress with her lips blurred. This effect was most prevalent at a moderate noise-vocoding level (6-band noise-vocoding) as compared to a severe noise-vocoding level (2-band noise-vocoding). These results suggested that there is a range that allows for optimal integration when the language system is weighted to an optimal reliance on auditory inputs (speech) and visual inputs (gestures and visible speech) to enhance degraded speech comprehension.

In the current study, we presented highly proficient non-native listeners of Dutch with the same stimuli and design of Drijvers & Ozyurek (2017), and compared the data from the current study to data on the native listeners reported in Drijvers & Ozyurek (2017). We focused on highly proficient participants because low proficient participants would not be able to recognize the verbs and only try to pick up information from visual cues. Investigating highly-proficient listeners would thus allow us to study the enhancement that both visual articulators contribute to clear and degraded speech comprehension. We hypothesized that non-native listeners would show a similar optimal integration range around 6-band noise-vocoding due to their high proficiency, but that differences in the amount of enhancement per visual articulator might arise when they cannot effectively use visual semantic information due to their non-native listener status. For example, non-native listeners might show a similar yet diminished pattern

compared to the native listeners described in Drijvers & Ozyurek (2017), because using the semantic cues that are conveyed by the gestures might be more difficult when the phonological information cannot be resolved when the speech is degraded (Bradlow & Alexander, 2007; Golestani et al., 2009; Oliver et al., 2012). Alternatively, non-native listeners might show a similar or increased use of visual articulators compared to native listeners, because their poorer comprehension of the non-native language restricts them to rely on solely auditory cues.

5.3. Methods

5.3.1. Participants

Twenty-three right-handed highly-proficient German listeners of Dutch (7 males, MeanAge = 22.35, SD = 1.69) with no neurological, language, hearing or motor disorders and normal or corrected-to-normal vision participated in this study. All participants had lived in the Netherlands for at least 1 year, regularly used Dutch for their studies/personal lives, and acquired Dutch after the age of twelve (meanAoA = 18.25, SD = 2.71) All participants partook in the LexTALE (LexTALE1) (Lemhöfer & Broersma, 2012), a 5-minute non-speeded visual lexical decision test, to determine whether they indeed were highly-proficient in Dutch, and only German listeners with a score above 60% were allowed to participate in the main experiment. A score of 60% and higher is thought to correlate with a B2 level or higher (upper-intermediate level). After the experiment, participants filled in an adapted version of the LexTALE test (LexTALE2) that contained the verbs that were used in the experiment, to ensure that the participants were familiar with the verbs that were used in the video (LexTALE1: MeanScore = 74.42%, SD = 7.97%, LexTALE2: MeanScore = 77.5%, SD = 11.45%). If a participant made a mistake in any of the verbs contained in the second LexTALE test, this verb was removed from the main analyses.

5.3.2. Stimulus materials

The materials in this experiment are identical to the set of stimuli used in Drijvers & Ozyurek (2017) and consisted of 220 videos of a Dutch non-professional actress

uttering a highly frequent Dutch action verb while she was displayed with either having her lips blurred, visible, or visible and accompanied by an iconic gesture (see Figure 15). All verbs were unique and only displayed in 1 condition. All gestures that were made by the actress were iconic and depicted the action verbs that she uttered (e.g., a mixing gesture resembling a whisking movement that accompanied the action verb 'mixing'). The actress was asked to spontaneously make a gesture that she found representative for the verb. None of the gestures were performed in front of the face to avoid blocking the lips of the actress. The fit with the verb and iconicity of these gestures was extensively pretested and described in Drijvers & Ozyurek (2017). We only included iconic gestures that were potentially ambiguous in the absence of speech, as this is how they are normally perceived in everyday communication (Krauss et al., 1991). The gestures we used had a mean recognition rate of ~59% (range: 37% - 81%). We pre-tested these gestures by asking participants to write down which verb they associated the video with. Moreover, we asked participants to rate the gesture for how iconic that gesture was for the verb (iconicity ratings >5 on a 7-point scale (mean 6.1), for more details, see: Drijvers & Ozyurek, 2017).

Every video was 2000ms long and speech onset started on average at 680ms after video onset. The preparation of the iconic gestures that the actress made started on average 120 ms after video onset, the stroke started on average at 550 ms., gesture retraction started at 1380 ms and gesture offset was at 1780ms. Speech onset was on average at 680 ms., meaning that the stroke onset started on average 130 ms. before speech onset, maximizing the overlap between the meaningful part of the gesture and speech for mutual enhancement and comprehension (Habets et al., 2011).

The speech in the videos was presented in clear speech, 2-band noise-vocoding and 6-band noise-vocoding. Noise-vocoding manipulates the spectral/temporal detail of the speech while preserving the amplitude envelope of the speech signal. Noise-vocoding results in a fairly intelligible speech signal, depending on the number of bands that are used for vocoding, with more vocoding bands resulting in a more intelligible signal. For example, in 2-band noise-vocoding the signal is band-pass filtered between 50 and 8000 Hz and divided into 2 logarithmically

spaced frequency bands, resulting in cut-off frequencies at 50, 632.5, and 8000 Hz. These frequencies were used to filter white noise in order to obtain 2 noise bands. The amplitude envelope of each band was extracted using half-wave rectification, multiplied with the noise bands, and recombined to form the degraded signal.

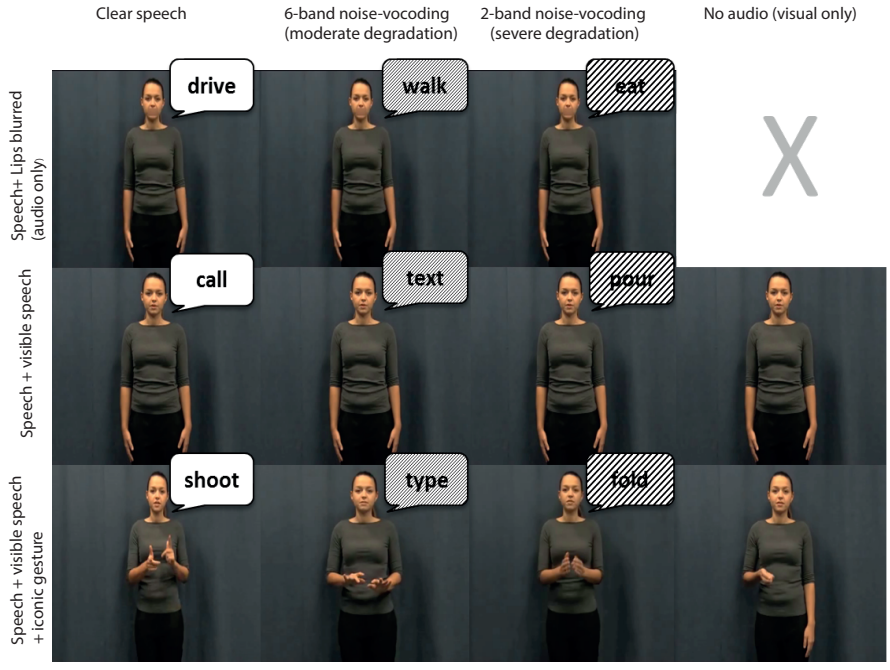


Figure 15 Overview of the design and conditions used in the experiment (picture taken from Drijvers & Ozyurek, 2017).

To test the different contributions of visible speech, gestures and both articulators combined to clear and degraded speech comprehension, we divided the 220 videos over eleven conditions (20 videos per condition, identical to Drijvers & Ozyurek, 2017). In the same 3x3 design, we manipulated the number of visual articulators present in the videos (Speech + Lips blurred; Speech +VisibleSpeech; Speech+VisibleSpeech+Gesture) and the different sound levels (2-band noise-vocoding, 'severe' degradation; 6-band noise-vocoding, 'moderate' degradation and clear speech). Two 'Visual only' conditions without audio files (VisibleSpeech+Gesture; VisibleSpeech Only (similar to lip-reading) were included to test how much information participants could obtain from the visual input by itself.

5.3.3. Procedure

Participants were tested in a dimly lit soundproof booth, and fitted with headphones. The experimenter gave a short verbal instruction to the participant to describe the different conditions that the video could be presented in. All stimuli were presented full-screen on a 1650x1080 monitor using Presentation (Neurobehavioral systems, Inc.), at a 70cm distance in front of the participant (identical to native listeners described in Drijvers & Ozyurek, 2017). All trials started with a fixation cross (1000ms), followed by a free-recall task that prompted the participants to type in the verb that they perceived in the videos. After they had submitted their answer, a new trial would start after 500 ms. An answer was 'correct' when the correct verb was written down, or minor spelling mistakes were made. Synonyms and category-related verbs (e.g. 'to bake for 'to cook' were coded as incorrect). All participants received a different pseudo-randomization of the stimuli, and no videos were presented more than twice in a row. The stimuli were divided over blocks of 55 trials, with a self-paced rest after every block. All participants completed the experiment within 45 minutes.

5.4. Results

Following Drijvers & Ozyurek (2017), we tested whether there were differences in the number of correct answers by using a 3 (VisualArticulator: (Speech+Lips blurred; Speech+VisibleSpeech; Speech+VisibleSpeech+Gesture) by 3 (Noise-vocoding Level: (2-band, 6-band, clear speech) repeated measures ANOVA. The VisualOnly conditions were tested separately. We found a significant main effect of Noise-Vocoding ($F(2,38) = 1543.23$, $p = < .001$, partial $\eta^2 = .99$). Pairwise comparisons (Bonferroni corrected) revealed that participants significantly gave more correct answers in the clear conditions than in the 2-band noise-vocoding condition ($p = < .001$), and significantly more correct answers in the 6-band noise-vocoding condition than in the 2-band noise-vocoding condition ($p = < .001$). This thus suggests that the more noise-vocoded the signal was, the less likely it was that participants gave a correct answer.

Second, we observed a main effect of VisualArticulator ($F(1.41, 26.75) = 184,78$, $p = < .001$, partial $\eta^2 = .91$, Greenhouse-Geisser corrected). Pairwise comparisons

(Bonferroni corrected) revealed that participants significantly gave more correct answers in the conditions both visible speech and gestures than in conditions containing only containing visible speech (p = < .001), and more correct answers in conditions containing visible speech than conditions in which the lips were blocked (p = < .001). This thus suggests that non-native listeners were more likely to provide a correct answer when both visual articulators were present.

We observed a significant interaction between VisualArticulator and Noise-vocoding (F(3.06, 58.06) = 52.71, p = < .001, partial η^2 = .74, Greenhouse-Geisser corrected). Contrasts confirmed that the difference in correct answers in conditions containing visible speech and conditions containing visible speech and gestures interacted for both clear speech vs. 6-band noise-vocoding (F(1,19) = 88.45, p = < .001 partial η^2 = .83), and 2-band noise-vocoding and 6-band noise-vocoding (F(1,19) = 10.46, p = < .01 partial η^2 = .10). As can be observed from Figure 16, the difference between conditions containing visible speech and gestures and conditions containing solely visible speech seems largest in 6-band noise-vocoding. This is similar as the interaction effect that is observed in Drijvers & Ozyurek (2017), and suggests that listeners benefit most from both visible articulators at 6-band noise-vocoding.

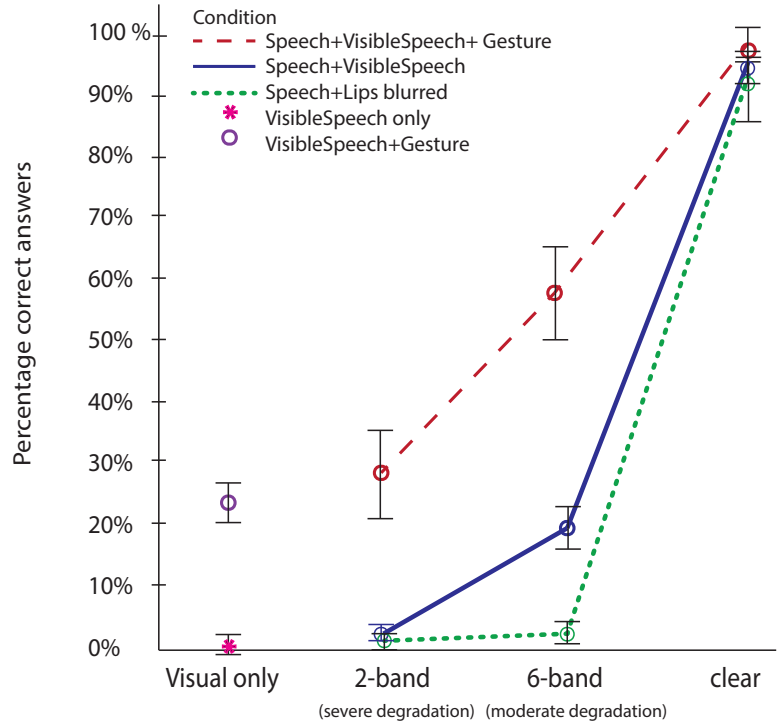


Figure 16 Percentage of correctly identified verbs (% correct) per condition.

Following previous studies (Drijvers & Ozyürek, 2017; Sumby & Pollock, 1954), we tried to confirm this effect by comparing the differences in enhancement per VisualArticulator as well as the enhancement per noise-vocoding level by defining three relative difference scores (A- B/100-B, i.e., enhancement types) for a) VisibleSpeech enhancement: Speech+VisibleSpeech - Speech+Lips blurred; b) Gestural enhancement: Speech+VisibleSpeech+Gesture - Speech+VisibleSpeech; and c) Double enhancement: Speech+VisibleSpeech+Gesture - Speech+Lips blurred, (see Ross et al., 2007 for a discussion of other calculation methods) divided by the maximal possible enhancement (e.g., for VisibleSpeech enhancement: 100 -Speech+Lips blurred). Note that although all enhancement types contained visible speech, we aimed to differentiate between the contribution of visible speech by itself (VisibleSpeech enhancement), gestural information occurring in the presence of visible speech (Gestural enhancement), and the contribution of both articulators combined (Double enhancement). Double

enhancement could be informative about whether or not one visual articulator enhanced the enhancement of the other articulator (e.g,, does the occurrence of gestural information on top of visible speech enhance comprehension on top of the enhancement of visible speech?). Moreover, these combinations more clearly mimic real-life occurences of gesture than for example a condition that would contain Speech+Lips blurred+Gesture.

These difference scores were compared by means of a repeated measures ANOVA with the factors EnhancementType (VisibleSpeech, Gestural or Double enhancement) and Noise-vocoding (2-band, 6-band, clear speech). We observed a significant main effect of Noise-Vocoding (F(1.35, 25,56) = 131.283, p = < .001, partial η^2 = .87, Greenhouse-Geisser corrected). Pairwise comparisons (Bonferroni corrected) revealed that enhancement was larger in 2-band noise-vocoding than in clear speech (p = < .001), and larger in 6-band noise-vocoding than in 2-band noise-vocoding (p = < .001). This thus suggests that enhancement was largest in 6-band noise-vocoding.

We observed a significant main effect of EnhancementType (F(2,38) = 85.93, p = < .001, partial η^2 = .82). Pairwise comparisons (Bonferroni corrected) revealed that enhancement was larger when both visual articulators were present than when one visual articulator was present (p < .001), and larger when both visual articulators were present than when no visual articulator was present (p < .001). This thus suggests that the more visual articulators were present, the more enhancement occurred.

We observed a significant interaction effect of Noise-vocoding x EnhancementType (F(2.71, 51.43) = 29.242, p = <.001, partial η^2 = .62, Greenhouse-Geisser corrected). Pairwise comparisons (Bonferroni corrected) revealed a significant difference between Gestural enhancement and Visible Speech enhancement in both 6-band noise-vocoding (t(19) = -8.441, p = < .001) and 2-band noise-vocoding (t(19)=7.644, p = < .001), see Figure 17). There was no difference between Gestural enhancement and Double speech enhancement in 2-band noise-vocoding (t(19) = -1.658, p = .11), but we did observe a difference in 6-band noise-vocoding (t(19) = -6.998, p < .001), see Figure 3). In more detail, this thus suggests that participants benefitted most from having both visual articulators

present in 6-band noise-vocoding than in 2-band noise-vocoding or clear speech, similar as was observed in Drijvers & Ozyurek (2017). However, although we observed a difference in Gestural enhancement and Double enhancement in 2-band noise-vocoding for native listeners, we did not observe this for non-native listeners.

In the Visual Only conditions, we observed a difference between VisibleSpeech only and VisibleSpeech+Gesture ($t(19) = -15.25$, $p = < .001$), indicating that more correct answers were given when both visible speech and gesture were present. Subsequently, we compared this Gestural enhancement (VisibleSpeech+Gesture – VisibleSpeech only/100 - VisibleSpeech Only) to Gestural enhancement in 6-band noise-vocoding and 2-band noise vocoding. We only observed a difference in Gestural enhancement in the 6-band noise-vocoding condition ($t(19) = 4.683$, $p = < .001$) but not in 2-band noise-vocoding $t(19) = 1.566$, $p = .13$), confirming that gestural enhancement was largest in 6-band noise-vocoding.

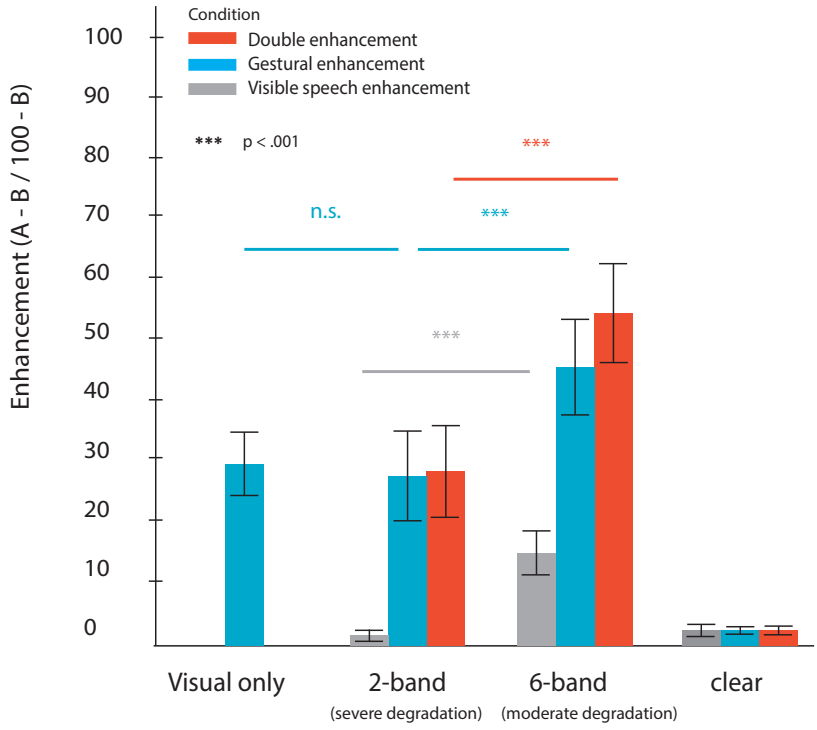


Figure 17 Enhancement effect (A-B/100-B) per visual articulator. Error bars represent SD, n.s. = not significant.

5.4.1. Comparison of native and non-native listeners

Although comparisons between the data from the current dataset and data from Drijvers & Ozyurek (2017) should be made carefully, we compared the enhancement effects for gestural enhancement, visible speech enhancement and double enhancement in native listeners (data from Drijvers & Ozyurek, 2017) and non-native listeners in a mixed repeated-measures ANOVA with one between-subjects factor (NativeLanguage), and two within-subject factors (EnhancementType: VisibleSpeech/Gesture/Double; Noise-Vocoding: 2-/6-band). This comparison revealed an interaction effect for EnhancementType & NativeLanguage ($F(2, 37) = 19{,}08$, $p = < .001$, partial $\eta^2 = .508$). Contrasts confirmed that the difference in visible speech enhancement and gestural enhancement was larger for native than non-native listeners ($F(1,38) = 9.77$, $p = < .01$ partial $\eta^2 = .21$), and that the difference in gestural enhancement and double enhancement was larger for

native than non-native listeners ($F(1,38) = 4.34$, $p = .044$, partial $\eta^2 = .10$). Second, we observed an interaction effect for Noise-Vocoding & NativeLanguage ($F(1, 38) = 4,299$, $p = .045$, partial $\eta^2 = .102$), revealing that the difference between enhancement in 2-band and 6-band noise-vocoding was larger for native listeners than non-native listeners (see Figure 18).

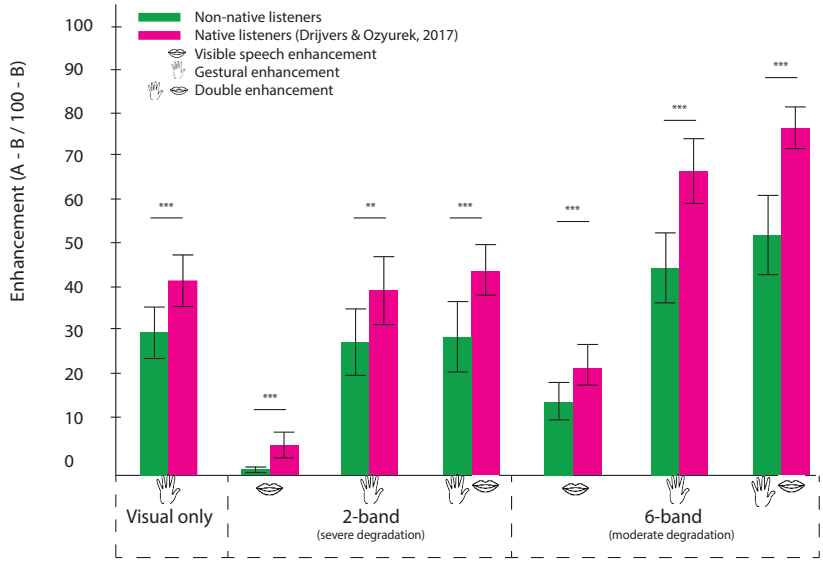


Figure 18: Enhancement effect (A-B)/100-B) per language (non-native/native, taken from Drijvers & Ozyurek, 2017). Error bars represent SD.

5.5. Discussion

The first aim of this study was to investigate whether and how non-native listeners can benefit from phonological cues from visible speech and semantic cues from iconic gestures to enhance speech comprehension in clear and adverse listening conditions. Second, we addressed the question of how these different visual articulators interact to enhance comprehension when they are presented in a joint context, and whether there is an optimal range where there is an equal reliance on auditory inputs (e.g., speech) and visual inputs (e.g., visible speech, gestures) when non-native listeners process degraded speech. In line with previous studies (Drijvers & Ozyürek, 2017; Holle et al., 2010; Ross et al., 2007; Sumby & Pollock, 1954), we found the largest enhancement of visible speech and gestures for non-

native listeners at a moderate level of noise-vocoding (6-band) as compared to severe level of noise-vocoding (2-band). However, the enhancement effects for visible speech, gesture and both visible articulators were less pronounced for non-native listeners than the effects observed in natives in Drijvers & Ozyurek (2017).

As we observed in native listeners (Drijvers & Ozyurek, 2017), non-natives benefit most when iconic gestures were presented on top of visible speech in all noise-vocoding conditions. We did not observe differences in clear speech, which is probably due to a ceiling effect. The overall enhancement pattern for non-natives largely followed what we observed in native listeners: when both iconic gestures and visible speech were present, participants found it easiest to identify the spoken verb in the video. This enhancement benefit was larger in 6-band noise-vocoding than 2-band noise-vocoding. Note that our gestures were not completely unambiguous, which allowed speech and gesture to mutually disambiguate each other especially in a moderately degraded context (6-band noise-vocoding). In a moderately degraded context, there are still enough auditory cues present to map the information that is conveyed by visual articulators to, in contrast to in 2-band noise-vocoding (in line with Kelly, Ozyürek, & Maris, 2010).

Interestingly however, unlike in native listeners, we observed a difference in Gestural enhancement and Double enhancement in 6-band noise-vocoding, but not in 2-band noise-vocoding. This suggests that in 2-band noise-vocoding, the enhancement that non-native listeners experience from iconic gestures is not aided by the presence of visible speech or the speech signal. This is most probably due to the fact that non-native listeners cannot couple the phonological cues that are conveyed by visible speech to the degraded auditory cues when noise-vocoding is too severe, which is in line with previous work on unimodal auditory speech processing (Bradlow & Alexander, 2007; Golestani, Rosen & Scott, 2009; Oliver et al., 2012). Subsequently, non-native listeners seem to rely primarily on the semantic information of the gesture during comprehension, but they cannot make use of the extra enhancement they would get from the phonological cues that are conveyed by visible speech to benefit from double enhancement. This is corroborated by the results that were observed in the visual-only conditions. Here, native listeners were more able to correctly identify the verb than non-native

listeners, suggesting that for non-native listeners it is more difficult to map both the information that is conveyed by visible speech and the information conveyed by gestures to a word that is familiar to them.

5.6. Conclusion

The results of the present study support our previous findings in native listeners from Drijvers & Ozyurek (2017), but also revealed some interesting differences. Whereas the enhancement from gestures was similar yet smaller for non-native listeners as compared to native listeners, the enhancement from visible speech was absent in 2-band noise-vocoding for non-native listeners. This thus indicates that in contrast to native listeners, non-native listeners might require a more intelligible signal than native listeners to benefit from visual phonological information, and subsequently, benefit from both visual articulators in a joint context. This effect was already observable at 6-band noise-vocoding, but was even larger in 2-band noise-vocoding, when even less phonological cues were present and signal quality was worse. This demonstrates that the optimal range for integration and reliance on auditory and visual inputs might be less liberal for non-native than native listeners because they need more phonological cues to optimally make use of the enhancement of both visible speech and gestures. This is in line with theories on speech-gesture comprehension that postulate that speech and gesture mutually enhance comprehension (Kelly et al., 2010), and explains why multimodal input, especially conveyed by visible speech and gesture, benefits non-native listeners less than native listeners.

5.7. Acknowledgements

We thank Natascha Roos and Natalia Dubinkina for help in pre-testing. This research was supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from Netherlands Organization for Scientific Research. We are very grateful to Nick Wood, for helping us in editing the video stimuli and to Gina Ginos, for being the actress in the videos.

Chapter 6

Native language status of the listener modulates the neural integration of speech and iconic gestures in clear and adverse listening conditions

6.1. Abstract

Native listeners neurally integrate iconic gestures with speech, which can enhance degraded speech comprehension. However, it is unknown how non-native listeners neurally integrate speech and gestures, as they might process visual semantic context differently than natives. We recorded EEG while native and highly-proficient non-native listeners watched videos of an actress uttering an action verb in clear or degraded speech, accompanied by a matching ('to drive'+driving gesture) or mismatching gesture ('to drink'+mixing gesture). Degraded speech elicited an enhanced N400 amplitude compared to clear speech in both groups, revealing an increase in neural resources needed to resolve the spoken input. A larger N400 effect was found in clear speech for non-natives compared to natives. In degraded speech, an N400 effect was only observable in native listeners. Non-native listeners might thus process gesture more strongly than natives when speech is clear, but need more auditory cues to facilitate access to gestural semantic information when speech is degraded.

This chapter is based on Drijvers, L., & Ozyurek, 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. doi:10.1016/j.bandl.2018.01.003.

6.2. Introduction

During face-to-face communication, a listener's brain constantly integrates information from auditory inputs, such as speech, and visual inputs, such as iconic co-speech gestures. For example, a listener might see a speaker making a drinking gesture (i.e., a hand mimicking a glass that is moved towards the mouth) when she is asking whether someone wants a drink. Iconic gestures, like that drinking gesture, can be described as hand movements that illustrate object attributes, actions, and space, and can carry semantic information that is relevant to what is conveyed in speech (e.g. Goldin-Meadow, 2005; McNeill, 1992). This semantic information can affect the processing of speech comprehension in normal and adverse listening conditions, such as in degraded speech (Drijvers & Ozyürek, 2017). So far, how the brain processes gestural information in the context of degraded speech has only been investigated in native listeners (Holle et al., 2010; Obermeier et al., 2012). However, the neural mechanisms that support speech-gesture integration in non-native listeners in clear and degraded speech have never been investigated.

Previous studies have reported that non-native listeners can make use of auditory semantic-contextual cues (e.g., a previous sentence context) in adverse listening conditions to aid comprehension, but only when the auditory signal is of sufficient quality to facilitate access to semantic information (Bradlow & Alexander, 2007; Mayo et al., 1997; Zhang et al., 2016). Non-native listeners can also benefit from visual semantic cues from gestures (Dahl & Ludvigsen, 2014; Sueyoshi & Hardison, 2005), but this has been only studied behaviorally in clear speech and with low-proficient non-native listeners. It remains unclear whether and how the semantic cues from iconic co-speech gestures can influence the neural processing of degraded speech comprehension in highly-proficient non-native listeners with sufficient vocabulary knowledge of a language. Whereas non-native listeners might process gestural information more strongly in clear speech than non-natives, they might require more auditory cues to benefit from gestures in degraded speech than native listeners. They might show more processing difficulties when coupling the semantic information from gesture to the degraded speech signal (see similar mechanisms proposed for difficulty in comprehension of reduced speech in non-

natives, Ernestus, Dikmans, & Giezenaar (2017)). To investigate this, the present study uses behavioral measures and event-related potentials (ERPs) as a measure of online neural semantic integration to investigate how native and non-native listeners integrate gestures with clear and degraded speech.

6.2.1. Native speech-gesture processing in clear and adverse conditions

There is ample evidence from both behavioral and neuroimaging studies that native listeners process and integrate gestures with clear speech (e.g. Beattie & Shovelton, 1999; 2002; Holle & Gunter, 2007; Holler et al., 2014; Holler, Kelly, Hagoort, & Ozyurek, 2010; Holler, Shovelton, & Beattie, 2009; Kelly, Barr, Church, & Lynch, 1999; Kelly, Healey, Özyürek, & Holler, 2015; Obermeier, Holle, & Gunter, 2011; Holle et al., 2012; Josse et al., 2012; see for a review, Özyürek, 2014), even when the gesture is irrelevant for the listeners' task (Kelly, Creigh, & Bartolotti, 2010), or when the gesture has no semantic content (beat gestures) (Biau & Soto-Faraco, 2013, 2015; Biau, Torralba, Fuentemilla, de Diego Balaguer, & Soto-Faraco, 2015; Dimitrova, Chu, Wang, Özyürek, & Hagoort, 2016; Wang & Chu, 2013). Furthermore, fMRI studies have studied speech-gesture integration from a spatial perspective, and reported involvement of bilateral posterior superior temporal sulcus/middle temporal gyrus (pSTS/MTG) (integration processes) and left inferior frontal gyrus (LIFG) (demanding semantic unification operations, revision/modification) (Dick, Mok, Raja Beharelle, Goldin-Meadow, & Small, 2014; Green et al., 2009; He et al., 2015; Holle, Gunter, Rueschemeyer, Hennenlotter, & Iacoboni, 2008; Holle et al., 2010; Willems, Özyürek, & Hagoort, 2007, 2009).

An alternative approach has been to investigate the temporal character of the brain mechanisms that support speech-gesture integration by measuring ERPs in the EEG signal. ERPs can be seen as deflections in voltage that are measured and recorded from electrodes placed on the scalp. Previous studies on the neural integration of iconic gestures and clear speech in native listeners (Cornejo et al., 2009; Habets, Kita, Shao, Ozyurek, & Hagoort, 2011; Holle & Gunter, 2007; Kelly et al., 1999; Kelly, Kravitz, & Hopkins, 2004; Obermeier et al., 2011; Wu & Coulson, 2005, 2007) have focused on the N400 component to assess differences

in semantic processing. The N400 is a negative-going ERP component between 300 - 600 ms. that peaks around 400 ms. The amplitude of the N400 is interpreted to reflect the ease of semantic integration and the extent to which neural resources are needed to integrate information. The N400 amplitude is smaller when semantic unification operations are easier (Kutas & Federmeier, 2000, 2014). Previous ERP studies on gesture processing have shown modulations of the N400 amplitude in mismatch paradigms (e.g., Cornejo et al., 2009; Habets et al., 2011; Kelly, Kravitz, & Hopkins, 2004; Kelly, Ward, Creigh, & Bartolotti, 2007; Kelly & Lee, 2012; Ozyürek et al., 2007; Sheehan, Namy, & Mills, 2007; Wu & Coulson, 2007), with more negative N400 amplitudes in response to speech that was presented with a mismatching gesture as compared to a matching gesture. This indicates that the brain is sensitive to the way gesture relates to speech, and that gesture is processed semantically. For example, Habets et al., (2011) investigated the degree of asynchrony in speech and gesture onsets that are optimal for semantic integration. They presented participants with videos where gestures were semantically congruent or incongruent, and where gesture and speech were presented either simultaneuos (SOA = 0), or the speech was delayed by 160 ms or 360 ms, and showed an N400 effect for the SOA 0 and SOA 160 conditions, but not the SOA 360 condition. Their results implied that speech and gesture are integrated most efficiently when they occur within a certain time span, because iconic gestures need speech to be disambiguated to fit within the speech context.

Contrary to the numerous studies on speech-gesture integration during clear speech processing, less is known about how native listeners integrate speech and gestures in adverse listening conditions. Previous research has shown that visual semantic cues that are conveyed by iconic gestures can enhance clear speech comprehension when speech is ambiguous (Holle & Gunter, 2007) and when speech is degraded (Drijvers & Ozyurek, 2017; Holle et al., 2010). For example, in an fMRI study, Holle et al., (2010) investigated which brain areas are responsive to speech-gesture integration, bimodal enhancement, and inverse effectiveness. They presented participants with videos that could either contain speech in a good signal-to-noise ratio, a moderate signal-to-noise ratio, or no speech. Simultaneously, the actor in these videos would either make an accompanying

iconic gesture or no gesture. Their results showed that speech-gesture integration could enhance speech comprehension in noise (especially at a moderate noise level) and that this bimodal enhancement was reflected by an increased activation of left pSTS/STG. Similarly, in a recent experiment (Drijvers & Ozyurek, 2017), we presented participants with videos with varying levels of visual information: videos could either contain a speaker with her lips blurred, a speaker with visible speech, or a speaker with visual speech and a gesture. The sound in these videos was presented either clear, moderately degraded by noise-vocoding (6-band) or severely degraded by noise-vocoding (2-band). The results revealed that listeners benefit more from having two visual articulators (i.e., visual speech and iconic gestures) present as compared to one (i.e., visible speech only), and that this benefit was largest at a moderate vocoding level, where listeners can still benefit from both the phonological cues from visible speech and semantic cues from iconic gestures to disambiguate the speech. However, although Holle et al., (2010) have demonstrated the spatial neural correlates of speech-gesture integration in adverse listening conditions, it remains unclear what the *online temporal neural correlates* are of how the semantic information from iconic gestures enhances the comprehension of degraded speech, and whether matching and mismatching gestures have an effect on the N400 amplitude in clear and degraded listening conditions. Second, Holle et al., (2010) have presented gestures in head-occluded conditions, and not in a context where all visual articulators are visible to participants. It remains unknown whether the semantic information conveyed by gestures is used as much when both visible speech and gestures are available as visible cues to enhance speech comprehension.

In the auditory domain, previous ERP studies have mostly focused on degraded speech comprehension in an auditory semantic context (e.g., a previous sentence context). These auditory electrophysiological studies have demonstrated that the N400 amplitude of a native listener is reduced in response to incongruent items that are acoustically degraded (e.g., a negative N400 amplitude when unifying an incongruent word with a preceding context in clear speech is less negative during degraded speech), or even absent when speech is too severely degraded (Aydelott et al., 2006; Boulenger et al., 2011; Obleser & Kotz, 2011; Strauß et al., 2013).

For example, Obleser & Kotz (2011) demonstrated that the N400 amplitude in response to low-cloze sentence-final words (indexing semantic integration load) decreased linearly with more signal degradation. In line with this, a similar EEG study by Strauß et al., (2013) on the influence of expectancies under degraded speech comprehension proposed that an adverse listening condition might narrow expectancies about the speech signal. By diminishing the sensory input, the neural system might rely more on signal-driven expectancies than contextual information.

The question remains however whether the neural resources that are needed to integrate a word with semantic information are similarly modulated by the imposed perceptual load of degraded speech when visual semantic context (e.g., iconic gestures) instead of auditory semantic context is provided. Unlike the sentential semantic context provided in the studies above, gestures might provide visual semantic context and semantic expectencies about a word when speech is degraded. This means that in response to degraded speech, the N400 amplitude might be more enhanced compared to clear speech, as a listener might recruit more neural resources when speech is degraded, such as visual semantic information that is conveyed by gestures to try to resolve the auditory input (in line with Skipper et al., 2006; 2007). Furthermore, the N400 effect in degraded speech might be smaller than in clear speech, due to the fact that gestures also need speech for their disambiguation (see Habets et al., 2011), and speech quality is diminished when speech is degraded.

6.2.2. Non-native speech-gesture processing in clear & adverse listening conditions

The next question is how gestures can enhance clear and degraded speech comprehension in non-native listeners. Non-native listeners might utilize visual semantic cues that are conveyed by gestures more than native listeners due to their lack of full proficiency. Behavioral studies have shown that iconic co-speech gestures can enhance non-native language comprehension and non-native language learning (e.g., Dahl & Ludvigsen, 2014; Macedonia & Kriegstein, 2012; Sueyoshi & Hardison, 2005). However, up to date, there are no studies on the

neural correlates of how visual semantic cues that are conveyed by gestures might enhance clear or degraded speech comprehension for non-native listeners.

Previous behavioral research on non-native degraded speech comprehension has been mostly tested in an auditory context, using only auditory semantic information in a verbal context as a modulating factor. These studies reported differences between native and highly proficient non-native listeners in terms of how previous auditory semantic context is taken into account during adverse listening conditions (Bradlow & Alexander, 2007; Bradlow & Bent, 2002; Gat & Keith, 1978; Golestani et al., 2009; Mayo et al., 1997; Oliver et al., 2012; Shimizu, Makishima, Yoshida, & Yamagishi, 2002; Wijngaarden et al., 2002; Zhang et al., 2016). However, how these differences are reflected in neural activity remains unknown. For example, in a behavioral study, Bradlow & Alexander (2007) presented native and non-native listeners with sentences in which the final word would either be highly predictable or not and produced in plain or clear speech. The results demonstrated that non-native listeners' comprehension was only aided when *both* semantic and acoustic information were available (e.g., in a sentence that was highly predictable and produced in clear speech). Conversely, native listeners could benefit from acoustic and semantic information both in combination and separately. One of the explanations for this difference between native and non-native listeners is that non-native listeners might not be able to use semantic contextual information to resolve the information loss at the phoneme level when the signal clarity was insufficient (e.g., Bradlow & Alexander, 2007; Golestani et al., 2009; Oliver et al., 2012; Zhang et al., 2016). In line with this, another audiovisual behavioral study by Hazan et al., (2006) demonstrated that non-native listeners effectively incorporate and use visual cues from visible speech that are related to phonological features in the auditory signal to enhance speech comprehension in noise, and that increasing auditory proficiency is linked to an increased use of visual cues by non-native listeners. Based on this previous research, one might expect differences in the way speech and gesture are integrated in non-natives and natives in clear and degraded speech contexts. Therefore, to get a detailed insight into possible processing differences between native and non-native listeners during this semantic integration, an on-line method that can

monitor the possible differences in neural integration is needed to investigate how the native language status of the listener influences the extent to which an iconic gesture is semantically integrated with clear and degraded speech.

6.2.3. The present study

We present an EEG study that aims to further our understanding of how native and non-native listeners integrate information online from speech and iconic co-speech gestures during both clear and degraded speech comprehension. Here, we measure the brain's electrophysiological response to the speech and gesture videos by focusing on ERPs in the EEG signal, to exploit the excellent temporal resolution this method offers. In line with previous electrophysiological research on the neural integration of speech and iconic gestures (e.g., Habets, Kita, Shao, Ozyurek, & Hagoort, 2011; Holle & Gunter, 2007; Ozyürek, Willems, Kita, & Hagoort, 2007; Wu & Coulson, 2007), we focused on the N400 component to neurally assess differences in how visual semantic information is integrated with clear and degraded speech in native and non-native listeners. To this end, we presented native and highly proficient non-native listeners with videos of an actress uttering Dutch action verbs (see Drijvers & Ozyurek, 2017), while she simultaneously made an iconic gesture that could either match or mismatch with the speech signal. The sound in these videos was either clear or degraded. All participants completed a behavioral 4-alternative forced choice identification task after each item that asked which verb they had heard in the videos.

Behaviorally, and in line with previous work (Holle et al., 2010; Drijvers and Ozyürek, 2017), we expected that native listeners would benefit from gestures during degraded speech comprehension, resulting in more correct answers on the 4-alternative forced choice identification task when a gesture matched the speech signal, and faster reaction times for matching than mismatching gestures during degraded speech comprehension. On an electrophysiological level, we expected that integrating gestures with degraded speech is more effortful and requires more neural resources than in clear speech because there are less auditory cues available. This would then result in higher N400 amplitudes in degraded speech as compared to clear speech. Furthermore, we expected a typical N400 effect when

comparing a matching and mismatching gesture in clear speech, with a more negative N400 amplitude in response to mismatching gestures. We expected a similar N400 effect in degraded as in clear speech, resulting in a more negative N400 amplitude in response to mismatching compared to matching gestures. However, we predicted this N400 effect to be smaller in degraded speech, because semantically coupling degraded speech with gestures will be more effortful due to the fact that the diminished auditory input will not always be resolved by gestures, especially not when the gesture mismatches the signal. This is in line with speech and gesture comprehension theories that claim that speech and gesture interact to enhance comprehension and that gestures also need speech to be disambiguated (Habets et al., 2011; Kelly et al., 2010).

For non-native listeners we expected similar behavioral results for all conditions due to their high proficiency. We recruited highly proficient non-native listeners with enough vocabulary knowledge of the words we presented. Low proficient participants would not recognize all of the verbs, and possibly be able to only pick up information from gestures. This would not be sufficient to study gestural enhancement of degraded speech comprehension.

On an electrophysiological level, we expected a similar typical N400 effect for highly-proficient non-native listeners during clear speech comprehension when comparing matching and mismatching gestures as in native listeners. Based on previous research however showing that non-natives might make more use of gestural context (e.g., Dahl & Ludvigsen, 2014) we expected that this N400 effect might be stronger in non-natives than in natives. However, non-native listeners' electrophysiological responses might differ from natives when speech is degraded. Non-native listeners might require more neural resources than natives to resolve degraded speech, resulting in a lesser ability to rely on visual semantic information to resolve the phonetic input than native listeners. This, in turn, might diminish how much non-natives can benefit from gestural information, resulting in no or a reduced N400 effect when comparing degraded speech and a mismatching gesture to degraded speech and a matching gesture. This would fit with previous behavioral results that suggested that a certain signal clarity is required for non-natives for semantic information to be effective (e.g., Bradlow & Alexander, 2007;

Hazan et al., 2006).

6.3. Methods

6.3.1. Participants

Twenty-four Dutch participants (mean age = 21.6, SD = 1.97, 9 males) and twenty-three German advanced learners of Dutch (mean age = 22.4, SD = 2.35, 8 males) participated in this experiment. All participants were right-handed and reported no language impairments, normal hearing, no motor disabilities and normal or corrected-to-normal vision. All participants gave informed written consent before the start of the experiment and received a financial compensation for participation.

All participants were students at Radboud University. The German participants ('non-native listeners') were recruited on the basis of the following criteria: They had lived or studied in the Netherlands for at least 1 year, had to use Dutch regularly (minimally once per week) for their studies and/or their personal lives, and acquired Dutch after age 12 (range: 12-23, mean age = 18.7, SD = 2.5). One participant from the Dutch participant group was excluded from analyses due to having excessive artifacts.

6.3.1.1. LexTALE assessment

Before the main experiment, the Dutch proficiency level of all participants was assessed by the Dutch version of the Lexical Test for Advanced Learners of English (LexTALE), a vocabulary test using non-speeded visual lexical decision (Lemhöfer & Broersma, 2012). Participants are presented with 40 Dutch words and 20 nonwords. Nonwords were nonsense strings created either by changing a number of letters in an existing word, or by recombining existing morphemes. Only German participants with a proficiency level of 67.5% and higher were allowed to participate in the main experiment. A score of 60% and higher is predicted to correlate with a B2 level or higher (Lemhöfer & Broersma, 2012). After the main EEG experiment (described below), participants were presented with an adapted version of the LexTALE to assees their knowledge of the specific

verbs that we used in the main experiment. Again, this version consisted of 40 real words from the main experiment and 20 nonwords.

6.3.2. Stimulus materials

The materials in this experiment are partially based on a subset of pretested stimuli which are described in more detail in Drijvers & Ozyurek (2017). We presented participants with 160 video clips of a female, native Dutch actress uttering a highly frequent Dutch action verb. All videos were recorded with a JVC GY-HM100 camcorder and had an average length of 2 seconds (See Figure 19B). The actress was visible from the knees up, wore neutrally colored clothing, and was standing in front of a unicolored background. The onset of each video was the same: The actress in the videos would stand in the middle of the screen with her arms hanging casually on each side of her body. The actress always produced an iconic co-speech gesture that could either match or mismatch with the spoken verb (e.g., the verb 'drive' and a driving gesture in the match conditions, and the verb 'eat' with a mixing gesture in the mismatch conditions, see Figure 19A).

The preparation of these gestures always started 120 ms after video onset, the stroke of the gesture started on average at 550 ms, gesture retraction at 1380 ms, and gesture ended at 1780 ms. Speech onset was on average at 680ms, which means that stroke onset started 130 ms before speech onset, maximizing the overlap between the meaningful part of the gesture and speech for mutual comprehension (Habets et al., 2011) (see Figure 19C).

Since our videos showed the face of the actress and we could therefore not recombine a mismatching auditory track to a video to create the mismatch condition, we asked the actress to utter a verb and produce a mismatching gesture with it. These mismatching gestures were created by dividing the list of verbs in the mismatch conditions in two lists, and combining the verbs on the first list with the gesture corresponding to a verb on the second list, and vice versa (e.g., a verb on the first list ('drink') would be coupled with a verb on the second list ('salt'), so the actress would utter the word 'drink' while making a salting gesture). Iconicity ratings of the gestures were conducted as part of Drijvers & Ozyurek (2017) and revealed a mean recognition rate of 59% when speech was absent. This reveals

that these gestures were potentially ambiguous without speech, which is mostly the case in spontaneous speech-gesture production (Krauss et al., 1991), and that they were to an extent dependent on speech to be disambiguated (Habets et al., 2011, see Drijvers & Ozyurek, 2017).

All auditory sound files were intensity-scaled to 70 dB, de-noised in *Praat* (Boersma & Weenink, 2015) and recombined with their corresponding video files in Adobe Premiere Pro. From every cleaned audio-file, a 6-band noise-vocoded version was created by using a custom-made Praat script. Noise-vocoding degrades the spectral content of the speech signal while pertaining the temporal envelope (Shannon et al., 1995). The speech signal then remains intelligible to a certain extent, with more bands corresponding to a more intelligible speech signal. Since our previous experiment (see Drijvers & Ozyurek, 2017) identified a 6-band noise-vocoding level as the optimal range in which iconic gestures can enhance degraded speech comprehension the most, this was also the speech degradation level that was used in this experiment (see Drijvers & Ozyurek (2017)).

In total, four conditions were created for this experiment: a clear speech + matching gesture condition ('clear-match', e.g., 'to eat' in clear speech combined with a co-speech gesture for 'to eat'), a clear speech + mismatching gesture condition ('clear-mismatch', e.g., 'to call' in clear speech combined with a mismatching co-speech gesture for 'to drive'), a degraded speech + matching gesture condition ('degraded-match', e.g., 'to mix' in degraded speech combined with a matching co-speech gesture for 'mixing') and a degraded speech + mismatching gesture condition ('degraded-mismatch', e.g., 'to turn' in degraded speech with a mismatching co-speech gesture for 'salting') (See Figure 19 for an overview). All conditions consisted of 40 unique videos with unique verbs and gestures.

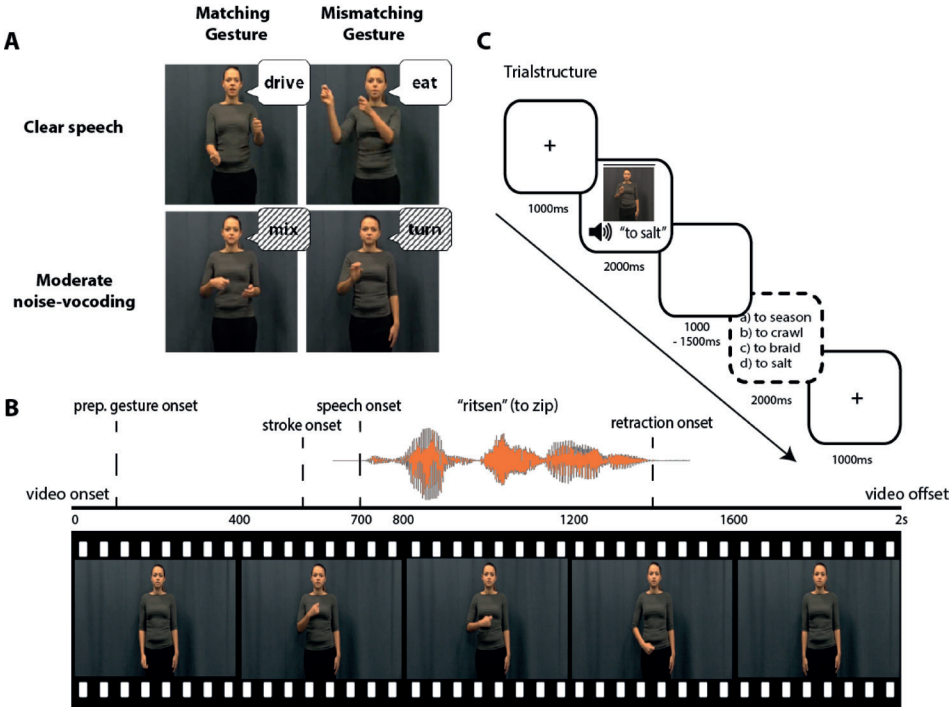


Figure 19 Experimental overview. A) Overview of conditions. B) Video structure. C) Trialstructure

6.3.3. Procedure

Upon arrival, participants first completed a consent form and participated in the LexTALE test before they were fitted with an EEG cap. Participants sat in front of a computer monitor while holding a four-button box in an acoustically and electrically shielded room. Stimuli were presented full screen on a 1650x1080 monitor by using Presentation software (version 16.4; Neurobehavioral Systems, Inc.) Participants were explained that the videos would contain a girl who would utter a Dutch action verb and asked to attentively watch and listen to the stimuli. Each trial would start with a fixation cross (1000 ms), after which the video started (2000 ms). After a short delay (1500 ms) participants were presented with a 4-alternative forced choice identification task and asked to identify which verb (out of four alternatives: correct answer, phonological competitor, semantic competitor, unrelated answer) they heard in the video by pressing a 4-button

box. The order of the stimuli was pseudo-randomized for all participants and presented in four blocks of 40 trials. The constraint on this randomization was that a condition could not be presented more than twice in a row. After each block, participants could take a self-paced break. All participants completed the experiment within 30 minutes. After the experiment, participants filled in the adapted version of the LexTALE to test their knowledge of the verbs used in these videos.

6.3.4. EEG data acquisition

The participants EEG was continuously recorded throughout the experiment from 32 AG-AgCI electrodes, of which 27 were mounted in a cap (actiCap) according to the 10-20 standard system, one was placed on the right mastoid for re-referencing and 4 were used for bipolar horizontal and vertical electrooculograms (EOG). The ground electrode was placed on the forehead. Electrode impedance was kept below 5 KQ. The EEG was filtered through a 0.02 - 100 Hz band-pass filter and digitized on-line with a sampling frequency of 500 Hz.

6.3.5. EEG data analysis

We analyzed the EEG data by using Fieldtrip (Oostenveld et al., 2011) a toolbox running under MATLAB (MathWorks, Natick, MA). First, we re-referenced the EEG data offline to the average of the right and left mastoid and filtered the data with a high-pass filter at 0.01 Hz and a low pass filter at 35 Hz. The data was segmented into epochs from -1 to 3.5 s relative to the onset of the videos. We used a baseline window of -0.4s to -0.2s. Artifacts were removed by using a semi-automatic rejection routine. On average, we excluded 8,1 % of the trials for each participant (13/160). One participant from the Dutch participant group was excluded from analyses due to having excessive artifacts.

To calculate the event-related potential, the time-locked average (time-locked to video onset) over all remaining trials was computed separately for the four conditions for each participant. We used non-parametric cluster-based permutation tests (Maris & Oostenveld, 2007) to evaluate the differences between conditions within each listener group separately. Using a multi-level statistical approach, a

dependent samples t-test was executed for every data point of two conditions (time by individual by electrode) for the within-group results. All adjacent data points that exceeded a pre-set threshold of 5% were grouped into clusters. In each of these clusters, the t-statistics were summed in order to calculate the cluster-level statistics. Then, a Monte-Carlo permutation distribution was created by randomly assigning a participant's average to one of the two conditions (1000 times) and calculating the largest cluster-level statistic for every permutation. The highest cluster-level statistic from each randomization was entered into the Monte Carlo permutation distribution and cluster-level statistics were calculated for the measured data and compared against this permutation distribution. Clusters that fell in the highest or lowest 2.5th percentile of the distribution were considered significant (see Maris and Oostenveld, 2007).

6.4. Results

6.4.1. Behavioral results - LexTALE

Non-native listeners scored within the high-proficiency range, but performed lower than native listeners on the first LexTALE test (mean = 92.8 (SD = 4.86) for native listeners vs. mean = 76.4 (SD = 5.38) for non-native listeners, $t(44) = 10.892$, $p = < 0.001$) and in the second, adapted LexTALE test (mean = 96.41, (SD = 3.60), for native listeners vs. mean = 86.58, (SD = 5.32) for non-native listeners, $t(44) = 7.34$, $p < 0.001$). The second test assessed their knowledge about the words we used as stimuli, and revealed that they were highly familiar with them, reaching almost native-like levels.

6.4.2. Behavioral results - 4-alternative forced choice identification task

For the within-group differences, we tested the difference in correct answers and reaction times in two 2 x 2 (Noise-vocoding (clear, degraded) x Gesture (match, mismatch)) repeated measures analysis of variance (ANOVA) per group (native/non-native) (see Figure 20). For the between-group differences, we tested the difference in correct answers and reaction times in a mixed repeated-measures ANOVA with Group as a between-group factor (native/non-native) and Noise-

vocoding and Gesture as within-group factors.

6.4.3. Behavioral results - Native listeners

We observed a significant effect of Noise-vocoding, indicating that when the speech signal was clear, native listeners' response accuracy was higher than when the speech was degraded ($F(1,22) = 140.95$, $p = < 0.001$, Wilks' Lambda = 0.135, $\eta^2 = 0.87$). We also found an effect of Gesture ($F(1,22) = 128.87$, $p = < 0.001$, Wilks' Lambda = 0.146, $\eta^2 = 0.85$), indicating that when the gesture matched the speech signal, native listeners were more able to correctly identify the verb. We found a significant interaction between Noise-vocoding and Gesture ($F(1,22) = 112.20$, $p = < 0.001$, Wilks' Lambda = 0.164, $\eta^2 = 0.83$), indicating that when the speech signal was clear and the gesture matched the speech signal, participants demonstrated higher response accuracy. Bonferroni corrected post-hoc analyses revealed a difference between clear-match and clear-mismatch $t(22) = 6.67$, $p_{bon} = < 0.001$, between degraded-match and degraded-mismatch $t(22) = 11.12$, $p_{bon} = < 0.001$, between clear-match and degraded-match $t(22) = 7.89$, $p_{bon} = < 0.001$ and between clear-mismatch and degraded-mismatch $t(22) = 12.42$, $p_{bon} < 0.001$.

We found a similar pattern in terms of reaction times and found a main effect of Noise-vocoding ($F(1,22) = 74.11$, $p = < 0.001$, Wilks' Lambda = 0.22, $\eta^2 = 0.77$), indicating that when the speech signal was clear, native listeners answered more quickly. We found a significant main effect of Gesture ($F(1,22) = 69.20$, $p = <0.001$, Wilks' Lambda = 0.24, $\eta^2 = 0.76$), indicating that when the gesture matched with the speech signal native listeners answered more quickly. Lastly, there was a significant interaction between Noise-vocoding and Gesture ($F(1,22) = 43.87$, $p = < 0.001$, Wilks' Lambda = 0.23, $\eta^2 = 0.66$), indicating that when the signal was clear and the gesture matched with the speech signal, native listeners answered more quickly. Bonferroni corrected post-hoc analyses revealed no significant difference between clear-match and clear-mismatch, $t(22) = -0.96$, $p_{bon} = 0.348$, but did show significant differences between degraded-match and degraded-mismatch, $t(22) = -7.80$, $p_{bon} = < 0.001$, between clear-match and degraded-match $t(22) = -6.97$, $p_{bon} = < 0.001$, and between clear-mismatch and degraded-mismatch, $t(22) = -8.73$, $p_{bon} = < 0.001$.

6.4.4. Behavioral results - Non-native listeners

In general, non-native listeners showed similar behavioral results as natives regarding the differences in conditions. Our analysis revealed a significant main effect of Noise-vocoding, indicating that when speech was clear, non-native listeners had a higher response accuracy than when speech was degraded ($F(1,22) = 165.47$, $p < 0.001$, Wilks' Lambda = 0.11, $\eta^2 = 0.88$) and a significant main effect of Gesture, indicating that when a matching gesture was present, non-native listeners were more able to correctly identify the verb than when a mismatching gesture accompanied the verb ($F(1,22) = 69.65$, $p < 0.001$, Wilks' Lambda = 0.24, $\eta^2 = 0.76$). Lastly, we found a significant interaction between Noise-vocoding and Gesture, indicating that when speech was clear and the gesture matched the speech signal, non-native listeners showed a higher response accuracy ($F(1,22) = 82.91$, $p < 0.001$, Wilks' Lambda = 0.21, $\eta^2 = 0.79$). Post-hoc analyses (Bonferroni corrected) showed revealed no significant difference in response accuracy between clear-match and clear-mismatch, $t(22) = -0.92$, $p = 0.367$, but did show significant differences between degraded-match and degraded-mismatch, $t(22) = -9.55$, $p < 0.001$, between clear-match and degraded-match $t(22) = -6.74$, $p < 0.001$, and between clear-mismatch and degraded-mismatch, $t(22) = -15.29$, $p < 0.001$.

We observed a similar pattern in reaction times as in response accuracy: We observed a significant main effect of Noise-vocoding ($F(1,22) = 104.554$, $p < 0.001$, Wilks' Lambda = 0.174, $\eta^2 = 0.82$), indicating that non-native listeners were quicker to respond when the speech signal was clear and a significant main effect of Gesture, indicating that when the gesture matched the speech signal, non-native listeners responded quicker than when the gesture mismatched with the speech signal ($F(1,22) = 53.42$, $p < 0.001$, Wilks' Lambda = 0.29, $\eta^2 = 0.70$). We observed a significant interaction between Noise-vocoding and Gesture, indicating that when speech was clear and the gesture matched the speech signal, non-native listeners were quicker to respond ($F(1,22) = 59.53$, $p < 0.001$, Wilks' Lambda = 0.27, $\eta^2 = 0.73$). Bonferroni corrected post-hoc analyses revealed no significant difference between clear-match and clear-mismatch, $t(22) = -0.71$, $p_{bon} = 0.483$, but did show significant differences between degraded-match and degraded-mismatch, $t(22) = -8.576$, $p_{bon} < 0.001$, between clear-match and degraded-match $t(22) = -8.48$, $p_{bon} <$

0.001, and between clear-mismatch and degraded-mismatch, $t(22) = -10.76$, $p_{bon} < 0.001$ (see Figure 20).

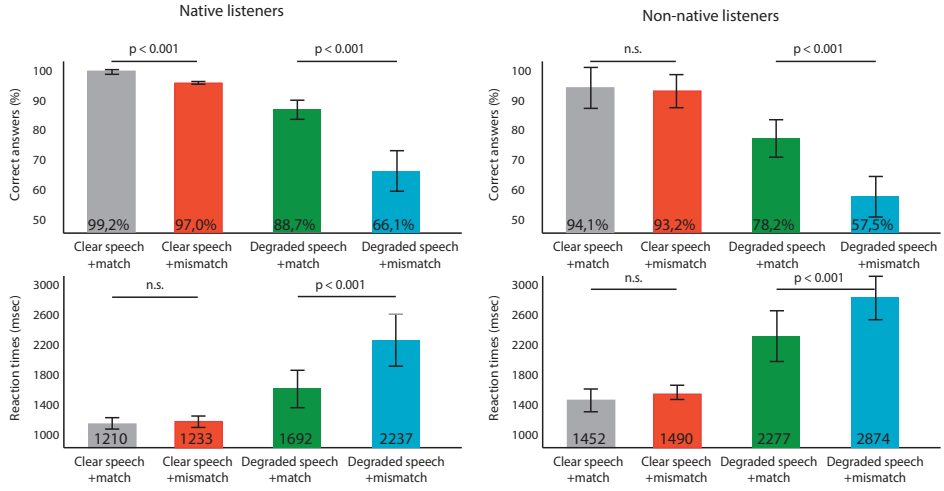


Figure 20 Behavioral results of the 4-alternative forced choice identification task. Top panels represent correct answers in percentages per listener group. Error bars present standard deviations. Lower panels represent reaction times (ms).

6.4.5. Behavioral results - Native versus non-native listeners

The analysis of the correct answers did not reveal a significant difference on any of the interaction terms that contained the between-group factor, indicating that native and non-native listeners scored similar on the number of correct answers in the clear and degraded speech trials ($F(1,44) = 3.116$, $p = .084$) and trials that contained a matching or a mismatching gesture ($F(1,44) = .282$, $p = .598$). We also observed no interaction between Noise-Vocoding, Gesture and Group ($F(1,44) = .001$, $p = .971$).

However, we observed different results in the reaction times. Native listeners answered more quickly than non-native listeners on clear and degraded speech trials ($F(1,44) = 6.965$, $p = .011$, Wilks' Lambda = 0.863, $\eta^2 = 0.14$), but not quicker on trials containing a matching or mismatching gesture ($F(1,44) = .266$, $p = .609$). We did not observe an interaction between Noise-Vocoding, Gesture and Group ($F(1,44) = .182$, $p = .67$).

In sum, these results thus show that non-native listeners thus demonstrate a

similar behavioral pattern as native listeners, but that overall they have slower reaction times in the degraded speech conditions.

6.4.6. EEG data - native participants

For the analyses of our EEG data, we defined our time-window of interest (1.0 - 1.7, which corresponds to 300ms after speech onset (at ~ 680 ms) until 1000 ms after speech onset, based on previous research on speech-gesture integration and N400 effects, and visual inspection of the waveforms (e.g., Habets et al., 2011; Kutas & Federmeier, 2014). We compared the ERPs of the four conditions time-locked to the onset of the video and averaged over all 23 native participants.

For native listeners, we observed a significant difference between the clear-match and the clear-mismatch condition (clear-mismatch>clear-match, $p < 0.001$), the degraded-match and the degraded-mismatch condition (degraded-mismatch>degraded-match, $p < 0.05$), between the clear-match condition and degraded-match condition (clear-match<degraded-match, $p < 0.001$) and the clear-mismatch and degraded-mismatch condition (clear-mismatch<degraded-mismatch, $p < 0.001$). Figure 21 shows the grand average event-related potentials for all four conditions, as well as the topographical plots of the N400 effects in clear and degraded speech. Degraded-mismatch elicited the largest N400 amplitude, followed by degraded-match, clear-mismatch and clear-match. In the clear speech conditions, the N400 effect was most pronounced over central-parietal electrodes, but in the degraded conditions, this effect was more widespread over left and right temporoparietal electrodes. To compare the N400 effects in clear and degraded speech, we subtracted the averages of the clear-match from the clear-mismatch condition, and the averages of the degraded-match condition from the degraded-mismatch condition. The N400 effect was larger in clear than in degraded speech ($p = 0.041$).

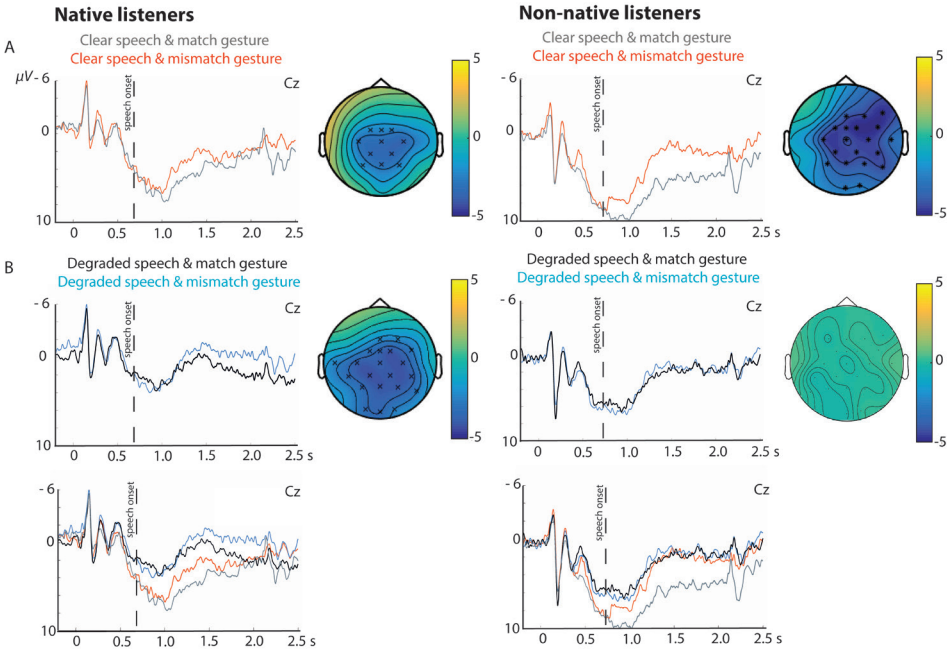


Figure 21 Grand-average waveforms for ERPs elicited in the different conditions at electrode Cz. Negativity is plotted upward. Waveforms are timelocked to the onset of the video. Bottom row of plot represents the plots from A and B overlaid to optimally view differences between the conditions.

6.4.7. EEG data - non-native listeners

In non-native listeners, we observed a significant difference between clear-match and clear-mismatch (clear-mismatch>clear-match, $p < 0.001$), but not between degraded-match and degraded-mismatch ($p = 0.16$). We observed a significant difference between clear-match and degraded-match (clear-match<degraded-match, $p < 0.001$) and between clear-mismatch and degraded-mismatch (clear-mismatch<degraded-mismatch, $p < 0.05$) Degraded-mismatch and degraded-match elicited the largest N400 amplitude, followed by clear-mismatch and clear-match. The topographical plots revealed that the N400 effect in clear speech extends over central-parietal as well as right lateralized electrodes, unlike what was observed in natives.

6.4.8. EEG data - native versus non-native listeners

Although the difference between native and non-native listeners in separate ERP waveforms per condition could not be compared, we did compare the N400 effects found in clear and degraded speech between the two groups (i.e., the interaction effect of nativeness and gesture congruence). We observed a larger N400 effect in clear speech for non-native listeners as compared to native listeners ($p<.05$) and a larger N400 effect for native as compared to non-native listeners in degraded speech ($p<.05$), which was driven by the absence of an N400 effect in the non-native listeners group.

6.5. Discussion

The current study examined whether and how (non)-native listeners neurally integrate iconic gestures with clear and degraded speech. Even though native and non-native listeners demonstrated similar behavioral results, our EEG results suggested that native and non-native listeners neurally integrate speech with gestures differently in both clear and degraded speech. Natives, but not non-natives revealed an N400 effect in degraded speech. Non-natives however, revealed a larger N400 effect in clear speech than native listeners. Below we will discuss these results in more detail.

6.5.1. Behavioral results - native & non-native listeners

Native and non-native listeners showed similar behavioral results and were more able to correctly identify a verb when speech was clear as compared to degraded, and when a gesture matched compared to mismatched speech. Reaction times revealed a similar pattern, but no difference in reaction times was observed between clear speech and a matching compared to a mismatching gesture for both native and non-native listeners, which is possibly due to a ceiling effect in both conditions. On a behavioral level, this thus suggests that both native and non-native listeners attempt to integrate gestures with both clear and degraded speech. Listeners seem to use the semantic information from gestures to boost comprehension when speech is degraded.

However, although the behavioral patterns of both groups look similar with regard to the differences between the conditions, non-native listeners seem to have slower reaction times in the degraded conditions than the clear conditions when compared to natives. This could indicate that it is more difficult for them to resolve the remaining auditory cues and couple the semantic information that is conveyed by the gesture to the speech signal (similar to results on reduced speech, such as Ernestus, Dikmans & Giezenaar (2017)). Our EEG results provided more evidence for this claim.

6.5.2. EEG results - native listeners

We observed a more negative N400 amplitude when gestures mismatched compared to matched clear speech (in line with e.g., Ozyurek et al., 2007; Kelly et al., 2004, Holle & Gunter, 2007; Habets et al., 2011), suggesting that integrating mismatching gestures requires more neural resources than integrating matching gestures. When speech was degraded we observed a similar pattern with more negative N400 amplitudes than in clear speech, suggesting more neural resources were required to integrate gestures when speech was degraded. Here, listeners might need more neural effort or semantic unification operations to disambiguate both the degraded auditory cues and the visual semantic information that is conveyed by the gesture.

Interestingly, previous research on auditory degraded speech comprehension has reported a *reduced* N400 amplitude when auditory target words were (increasingly) degraded as compared to clear and when they were presented in a low-cloze probability context (e.g., when semantic expectations about the upcoming word are low) (Aydelott et al., 2006; Obleser & Kotz, 2011; Strauß et al., 2013). Although we did not provide listeners with prior auditory context similar to the studies mentioned above, we did provide listeners with a visual semantic context. Note that this visual semantic context was not completely unambiguous, as both the gesture and the speech could mutually disambiguate each other. We expected that, to some extent, gestures could therefore elicit predictions about the degraded word, which, in turn, could have enhanced degraded speech comprehension, resulting in the recruitment of more neural resources compared to clear speech. In line with this tentative explanation, we observed an *increased* N400 amplitude in response to degraded compared to clear speech. Note that our data revealed stepwise differences between the conditions: the degraded-mismatch condition yielded the largest N400 amplitude, followed by degraded-match, clear-mismatch and clear-match. We suggest that this shows an increase in neural resources that are required to resolve the speech signal and couple the semantic information conveyed by the gesture, resulting in an additive effect of both speech degradation and semantic incongruency of the gesture on the amplitude of the N400.

We also observed an N400 effect in both degraded speech and clear speech, which shows that gestures exert a visual semantic context effect. This N400 effect was reduced in degraded speech, which is possibly due to the fact that listeners have less auditory cues to their disposal to couple the gestural information to. This is also in line with Obleser & Kotz (2011), who find that effortful semantic computation is more visible in less degraded signals, that is, when signal quality is good enough for semantic manipulations to have an effect on comprehension. Similarly, this is also the case for gestural information, which is partially ambiguous without the speech context.

In Figure 20 A and B, a possible latency of the N400-effect can be observed when comparing the effect in degraded and clear speech. However, post-hoc

analyses of the N400 peak latency did not reveal any difference between clear and degraded speech, but only in the onset of the N400 effect (1000ms for clear speech vs. 1280 ms for degraded speech) Previous studies (e.g., Obleser & Kotz, 2011; Strauß et al., 2013) did report significant differences in peak latency in response to degraded speech, suggesting a delayed semantic integration. Since we did not find a difference in N400 peak latency but only in the onset of the N400 effect, we suggest this is due to the auditory cognitive load that degraded speech imposes on the listener (Connolly, Philips, Stewart, & Brake, 1992).

6.5.3. EEG results - non-native listeners

Similar to native listeners, non-native listeners showed a more negative N400 amplitude in clear speech for mismatching than matching gestures. In degraded speech, non-native listeners revealed no difference between matching and mismatching gestures, nor did these N400 amplitudes differ from the N400 amplitude of mismatching gestures in clear speech.

These results seem in line with theoretical explanations of why differences between native and non-native listeners arise under adverse listening conditions. Possibly, non-native listeners cannot fully make use of the semantical cues of the gesture when the auditory cues are too difficult to resolve (Gat and Keith, 1978; Mayo et al., 1997; Bradlow and Bent, 2002; Bradlow and Alexander, 2007; Golestani et al., 2009; Oliver et al., 2012). Compared to native listeners, non-native listeners may have required more neural resources to resolve the degraded auditory cues. In turn, this may have caused a limited benefit from visual information for comprehension, especially when the degraded auditory cues were not reliable enough to couple the visual semantic information to or for the visual information to boost comprehension of the degraded auditory cues. This might have resulted in a similar N400 amplitude of the degraded conditions and the clear speech and mismatching gesture condition. In the 4-alternative forced choice identification task however, the unreliable degraded auditory cues might be more easily recognized when the four answer options were presented. This might have masked the actual comprehension difficulties the listeners had when they watched the video.

In line with this interpretation of our data, we also observed a smaller N400 effect for natives in degraded compared to clear speech. Similarly, this result suggested that the neural processing of semantic integration already suffered from having less auditory cues present to map the semantic information from the gestures to. We therefore suggest that this effect is even more enhanced for non-native listeners: when signal quality suffers and there are less auditory cues to map semantic information to, non-native listeners are less able than native listeners to benefit from semantic information from the gesture to boost comprehension and resolve the degraded auditory input. Note that a mismatching gesture in degraded speech can possibly also have a deleterious effect, when the visual information was difficult to integrate with the remaining auditory cues and the semantic information did not aid resolving the auditory cues.

A direct comparison of the ERP waveforms of native and non-native listeners was not possible because of the many differences there could exist between these groups that are irrespective of the experimental manipulation, such as motivation (which might have been larger for the non-native group, as they completed a Dutch language proficiency test upon arrival). For example, post-hoc analyses of the N1/P2 complex at the start of the video revealed differences between the groups that could not be explained by stimulus characteristics. However, we did compare the N400 effects in the two groups, and found a larger N400 effect in clear speech for non-native compared to native listeners, and a larger N400 effect in degraded speech for native listeners (due to the absence of an N400 effect in non-native listeners). This revealed that in clear speech, non-native listeners possibly recruit the visual semantic information more than native listeners, which is possibly due to the fact that they pay more attention to gestures when they are unsure about their language proficiency. As we did not observe an N400 effect in degraded speech, we suggest that non-native listeners might employ different neural processing strategies for semantic information than native listeners when speech is degraded. One possibility is that non-native listeners first try to resolve the degraded auditory cues and recruit more visual information when resolving the degraded cues is too taxing. If however the remaining auditory cues are not reliable enough, they cannot benefit from these semantic cues. Native listeners on

the other hand use and attempt to integrate the visual semantic information to immediately sharpen their perception to resolve the degraded speech signal, and can benefit more from this information than non-natives.

Although differences in the distribution of the N400 component should be carefully made on the basis of ERP scalp topographies, we observed a more right-lateralized topography of the N400 effect in clear speech for non-native as compared to native listeners. Right-hemisphere effects have been found in a range of studies that reported sensitivity of the right hemisphere during speech-gesture integration (especially in pSTS/MTG, Green et al., 2009; Holle et al., 2010; Holler, Kokal, et al., 2014; Skipper et al., 2009; Straube et al., 2012; Willems et al., 2007, 2009), when semantic contexts are indirectly related (Kiefer, Weisbrod, Kern, Maier, & Spitzer, 1998) and when gestures were semantically more distant (i.e., mismatching) (Kelly et al., 2004, 2007). In clear speech, non-native listeners might attempt to exploit and process the semantic information from gestural input more than native listeners, resulting in the recruitment of right-lateralized areas in the heightened processing of the semantic information that is provided by the gesture. A similar pattern is observed in the N400 effect in degraded speech for native listeners, where we observed a widespread negativity over both left and right lateralized electrodes. Previous literature has hypothesized that the N400 could reflect reverberating neural activity that is instantiated by a network consisting of memory/storage (MTG/STG), unification (LIFG) and control retrieval (dorsolateral prefrontal cortex) areas (Baggio & Hagoort, 2011). Especially when speech is degraded, the dynamic reverberating circuits involved between (L)IFG and pSTS/MTG might be more widespread to recruit more top-down information to enhance degraded speech comprehension and facilitate unification of the two input streams. This more extended network would also fit with the account that when speech processing becomes more taxing, additional neural resources are recruited to aid in comprehension (Skipper, Nusbaum, & Small, 2006; Skipper, Wassenhove, Nusbaum, & Steven, 2007).

In future work, we aim to address these questions by including a baseline condition where there is no gesture present, to investigate whether the semantic information from the gesture enhances recognition depending on semantic

congruency. Future studies could also test the current paradigm in a more sentential context, or whether similar results will hold when participants have a lower proficiency level, to test how a possible larger dependence on visual semantic information affects comprehension.

6.6. Conclusion

Our data revealed that native and non-native listeners differ in the extent to which the semantic information from the gesture is coupled to the degraded speech signal on a neural level. Non-native listeners might recruit additional neural resources to process gestural information when speech is clear, by focusing more on gestural information than native listeners. While both native and non-native listeners use more neural resources to disambiguate the degraded speech signal, non-native listeners were more hindered in their ability to neurally couple the semantic information from the gesture to degraded auditory cues, possibly because they need more auditory cues to facilitate access to gestural information. Thus, although gestures enhance degraded speech comprehension, highly-proficient non-native listeners benefit less from visual semantic context than native listeners and integrate speech and gestures differently.

6.7. Acknowledgements

This research was supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from Netherlands Organization for Scientific Research. We are very grateful to Nick Wood, for helping us in editing the video stimuli, to Mary-Jo Diepeveen, for help collecting the data, to Gina Ginos, for being the actress in the videos and to Peter Hagoort and Ole Jensen for helpful discussions.

Chapter 7

Native and non-native listeners show similar yet distinct oscillatory dynamics when using gestures to access speech in noise

7.1. Abstract

Listeners are often challenged by adverse listening conditions during language comprehension induced by external factors, such as noise, but also internal factors, such as being a non-native listener. Visible cues, such as semantic information conveyed by iconic gestures, can enhance language comprehension in such situations. Using magnetoencephalography (MEG) we investigated whether spatiotemporal oscillatory dynamics can predict a listener's benefit of iconic gestures during language comprehension in both internally (non-native versus native listeners) and externally (clear/degraded speech) induced adverse listening conditions. Proficient non-native speakers of Dutch were presented with videos in which an actress uttered a degraded or clear verb, accompanied by a gesture or not, and completed a 4-alternative forced choice identification task after every video. The behavioral and oscillatory results obtained from non-native listeners were compared to an MEG study where we presented the same stimuli to native listeners (Drijvers et al., 2018a). Non-native listeners demonstrated a similar gestural enhancement effect as native listeners, but overall performed significantly more slowly on the 4-alternative forced choice identification task. In both native and non-native listeners, an alpha/beta power suppression revealed engagement of the extended language network, motor and visual regions during gestural enhancement of degraded speech comprehension, suggesting similar core processes that support unification and lexical access processes. An individual's alpha/beta power modulation predicted the gestural benefit a listener experienced during degraded speech comprehension. Importantly, however, non-native listeners showed less engagement of the mouth area of the primary somatosensory cortex, left insula (beta), LIFG and ATL (alpha) than native listeners, which suggests that non-native listeners might be hindered in processing the degraded phonological cues and coupling them to the semantic information conveyed by the gesture. Native and non-native listeners thus demonstrated similar yet distinct spatiotemporal oscillatory dynamics when recruiting visual cues to disambiguate degraded speech.

This chapter is based on Drijvers, L., Van der Plas, M., Ozyurek, A., Jensen, O. (in press). Native and non-native listeners show similar yet distinct oscillatory dynamics when using gestures to access speech in noise. *NeuroImage*.

7.2. Introduction

Adverse listening conditions during language comprehension can be caused by external factors, such as noise (Peelle, 2017b), but also internal factors, such as when understanding language as a non-native listener (Lecumberri et al., 2010). Especially under such adverse listening conditions, listeners can improve comprehension by integrating information from auditory modalities, such as speech, and visual modalities, such as visible speech and co-speech gestures. Brain oscillations are thought to have a mechanistic role in enabling the integration of information from these different auditory and visual modalities (Kayser & Logothetis, 2009; Schroeder et al., 2008; Senkowski, Saint-Amour, Hofle, & Foxe, 2011; Varela, Lachaux, Rodriguez, & Martinerie, 2001). The engagement of brain areas that are relevant for this integration process is often thought to relate to a suppression of low-frequency oscillatory power in the alpha (8 – 12 Hz) and beta (13 – 30 Hz) band (Jensen & Mazaheri, 2010; Klimesch et al., 2007; Payne & Sekuler, 2014). Oscillatory power modulations have shown to be predictive of the degree of non-semantic (e.g., beeps and flashes, Hipp et al., (2011)) and semantic audiovisual integration of an ambiguous stimulus (e.g., speech degradation, (Drijvers et al., 2018a; Drijvers et al., 2018b). Here, we investigate how brain oscillations support semantic audiovisual integration when listeners face adverse listening conditions induced by both internal factors (i.e., non-nativeness) and external factors (i.e., speech degradation).

When listeners face adverse listening conditions induced by an external factor, such as speech degradation, studies on unimodal auditory degraded speech comprehension have demonstrated less suppressed alpha power when speech was degraded, possibly reflecting an increased auditory cognitive load when language processing is inhibited (Obleser & Weisz, 2012; Weisz et al., 2011; Wostmann et al., 2015). In multimodal adverse listening conditions, however, semantic information conveyed by iconic gestures has been shown to enhance language comprehension (Drijvers & Ozyürek, 2017; Holle et al., 2010). These iconic gestures (e.g., a 'mixing' gesture when describing a recipe) can convey semantic information that illustrates objects, actions or spatial relationships (McNeill, 1992) and are thought to be automatically integrated with speech (Kelly, Creigh, & Bartolotti, 2010) on

both a neural and behavioral level (see for an overview, Özyürek, 2014). Imaging studies relying on fMRI that investigated the spatial correlates of this process suggested that the semantic integration of speech involves left-inferior frontal gyrus (LIFG), posterior middle temporal gyrus (pMTG), superior temporal sulcus (STS), visual and motor regions (Dick et al., 2014; Green et al., 2009; Straube et al., 2012; Willems et al., 2009, 2007; Zhao et al., 2018.). EEG studies on the temporal correlates of speech-gesture integration reported low-frequency oscillatory modulations to gestures that had both a semantic and non-semantic relation to speech (Biau & Soto-Faraco, 2015; Biau et al., 2015; He et al., 2015; He, Gebhardt, Str, Rondinone, & Straube, 2011; He et al., 2018). In line with studies on non-semantic audiovisual integration (Hipp, Engel & Siegel, 2011) and studies on the neural correlates of speech-gesture integration, we demonstrated in a previous MEG study that oscillatory power modulations in LIFG, left-temporal, motor and visual regions can predict how much a listener can benefit from gestures during degraded speech comprehension (Drijvers et al., 2018a). However, it is unknown whether similar oscillatory modulations can predict how much a listener can benefit from the semantic information conveyed by gestures when internal factors cause an adverse listening condition during language comprehension, such as when understanding language as a non-native listener.

When an internal factor, such as non-nativeness, impacts language comprehension, previous research demonstrated that semantic information conveyed by gestures can enhance language comprehension (Dahl & Ludvigsen, 2014; Sueyoshi & Hardison, 2005). However, in a recent EEG study that investigated how the N400 component was modulated by the semantic congruency of gestures in clear and degraded speech, an N400 effect for non-native listeners was observed only when speech was clear but not when speech was degraded. Thus, although non-native listeners seem to benefit from gestural enhancement during degraded speech comprehension, speech-gesture integration seems to be hindered for non-native listeners when speech is degraded (Drijvers & Özyürek, 2018). A potential explanation for these findings is that non-native listeners need more phonological cues to benefit from the semantic information that is conveyed by the gesture. This is in line with previous behavioral work that demonstrated that non-native

listeners can only utilize auditory semantic-contextual cues for comprehension when the auditory signal is of sufficient quality to allow access to semantic cues (Bradlow & Alexander, 2007; Golestani et al., 2009; Hazan et al., 2006; Mayo et al., 1997; Oliver et al., 2012; Zhang et al., 2016). However, it is unknown which brain areas engage in this process over time, and how this differs from native listeners, who are not challenged by internally induced adverse listening conditions when understanding language.

The current paper investigates whether spatiotemporal oscillatory dynamics can predict how much a listener can benefit from semantic information conveyed by gestures in internally induced (i.e., non-nativeness) and externally induced adverse listening conditions (i.e., speech degradation). Using the same paradigm as in Drijvers et al., (2018a) where only external factors induced an adverse listening condition, we presented participants with videos of an actress who uttered an action verb in clear or degraded speech, while making a gesture or not. After watching the videos, the participants had to indicate which verb they heard in a 4-alternative forced choice identification task. An internally induced adverse listening condition was created by testing highly proficient non-native speakers of Dutch with sufficient vocabulary knowledge of Dutch, as low-proficient participants would not recognize all verbs, or be focused solely on the gesture. An externally induced adverse listening condition was created by manipulating speech quality by noise-vocoding. We used the already acquired MEG data from native listeners (described in Drijvers et al., 2018a) to compare to the oscillatory activity observed in non-native listeners during semantic audiovisual integration.

We expected that non-native listeners would show a similar gestural enhancement effect as native listeners on the 4-alternative forced choice identification task. However, we predicted that non-native listeners would overall be less accurate and slower than native listeners when answering what verb they heard in the videos. This would, in line with previous literature (Drijvers & Ozyurek, 2017; in revision) indicate that although non-native listeners benefit from gestures during degraded speech comprehension, they might be hindered in resolving the degraded auditory cues and coupling those cues to the semantic information that is conveyed by the gesture.

Our central hypothesis was that a suppression of alpha (8-12 Hz) and beta power (15 - 20 Hz) would reflect engagement of brain regions that are relevant for comprehension during gestural enhancement of degraded speech comprehension. Similar to what was observed in previous work on native listeners (Drijvers et al., 2018a), we predicted that for non-native listeners, gestural enhancement would rely on the engagement of the extended language network (LIFG, and left-temporal regions), motor and visual regions to perform this semantic audiovisual integration. As for native listeners, we expected that for non-native listeners a larger alpha power suppression in the extended language network would reflect stronger engagement of these regions when unification load is higher (Wang, Jensen, van den Brink, et al., 2012; Drijvers et al.,, 2018a, 2018b). We expected larger alpha power suppression over visual regions, reflecting a larger allocation of visual attention to gestures when speech is degraded. A larger beta power suppression over motor regions would reflect a larger engagement of these regions during gestural observation when speech is degraded (Caetano et al., 2007; Kilner et al., 2009; Koelewijn et al., 2008). However, as previous work suggested that non-native listeners might only be able to utilize semantic cues when the auditory signal is of sufficient quality to allow access to these semantic cues (Bradlow & Alexander, 2007; Golestani et al., 2009; Hazan et al., 2006; Mayo et al., 1997; Oliver et al., 2012; Zhang et al., 2016), we expected less engagement of the LIFG for non-native listeners compared to native listeners. This would reflect that when speech is degraded, it is more difficult for non-native listeners to unify the degraded auditory cues with the semantic information that is conveyed by the gesture. Lastly, we expected that the observed oscillatory power modulations in non-native listeners would correlate with the benefit a non-native listener would experience during degraded speech comprehension, similar as to what was observed for native listeners (Drijvers et al., 2018a).

7.3. Methods

7.3.1. Participants

The non-native listener group was formed by thirty-two right-handed German

advanced learners of Dutch (mean age = 23.09, 15 males) who reported normal hearing, normal or corrected-to-normal vision, no language, motor or neurological impairments. All participants were students at Radboud University who were paid to participate in the study, and were recruited on the basis of the following criteria: They had lived or studied in the Netherlands for at least 1 year, had to use Dutch regularly (minimally once per week) for their studies and/or their personal lives, and acquired Dutch after age 12. We excluded two participants due to unreported metal (1) in their bodies and left-handedness (1). The data of the non-native listener group was compared to the data of the native listener group (n=30) reported in Drijvers et al., (2018a). All participants gave written consent before participation.

7.3.1.1. LexTALE assessment

As we aimed to recruit highly-proficient non-native German speakers of Dutch to introduce internal ambiguity, we assessed the proficiency level of our (potential) participants with the Dutch version of the Lexical Test for Advanced Learners of English (LexTALE), a vocabulary test using non-speeded visual lexical decision (Lemhöfer & Broersma, 2012). In this test, participants are presented with 60 words (40 Dutch words, 20 nonwords) of which they have to decide whether is a real word in Dutch or not. Nonwords are constructed of strings created by either changing a few letters in a real Dutch word, or by recombining existing Dutch morphemes. As we were aiming for an intermediate-high proficiency level of our participants, we only included participants who scored at a B2 level or higher (above 67.5%). After the experiment, we used an adapted version of the LexTALE test consisting of 40 verbs that were used in the experiment, and 20 non-words that were constructed on the basis of the stimuli used in the experiment to ensure that the German participants were familiar with the verbs that were used in the MEG experiment (similar to Drijvers & Ozyurek, 2018).

7.3.2. Stimulus materials

We used the same stimuli as in Drijvers et al., (2018a). These stimuli consisted of 160 2-second video clips of a woman uttering Dutch verbs in either clear or

degraded speech, while producing a gesture or not. All verbs that were used were highly frequent Dutch action verbs (see for pre-tests and earlier behavioral experiment, Drijvers & Ozyurek (2017), and for German participants, (Drijvers & Ozyurek, 2019, Drijvers & Ozyurek, 2018)). The actress in the videos was visible from the knees up, and was wearing neutral-colored clothing and stood at a neutrally colored background (see Figure 22A).

In short, all videos had an average length of 2000ms. The preparation of the gesture (i.e., the first frame in which the actress moved her hand), was at 120 ms. The stroke of the gesture commenced at approximately 550ms, followed by speech onset at approximately 680 ms, gesture retraction at 1380 ms and gesture offset at 1780 ms. Note that the stroke of the gesture started on average 130 ms before speech onset, maximizing the overlap between speech and gesture for mutual enhancement and comprehension (Habets et al., 2011).

The sound in the videos was intensity-scaled to 70 dB, de-noised with *Praat* (Boersma & Weenink, 2015) and recombined with their corresponding video files in Adobe Premiere Pro. All the 'cleaned' audio files were noise-vocoded by using 6 noise-vocoding bands (see for pretests; Drijvers & Ozyurek (2017) and Drijvers & Ozyurek (under review)). We used 6 noise-vocoding bands because pretests had shown that 6-band noise-vocoding allowed for the most gestural enhancement in both non-native and native speakers. Noise-vocoding obtained by band-pass filtering each speech file between 50 and 8000 Hz, and dividing the signal by 6 logarithmically spaced bands between 50 and 8000 Hz. This resulted in cut-off frequencies at 50 Hz, 116.5 Hz, 271.4 Hz, 632.5 Hz, 1473.6 Hz, 3433.5 Hz and 8000 Hz. We used half-wave rectification to extract the amplitude envelope and multiplied the amplitude envelope with the noise-bands, before recombining the bands to create the degraded speech signal (Shannon et al., 1995). The speech sounds from the presented videos were presented to the participant through plastic MEG compatible air tubes.

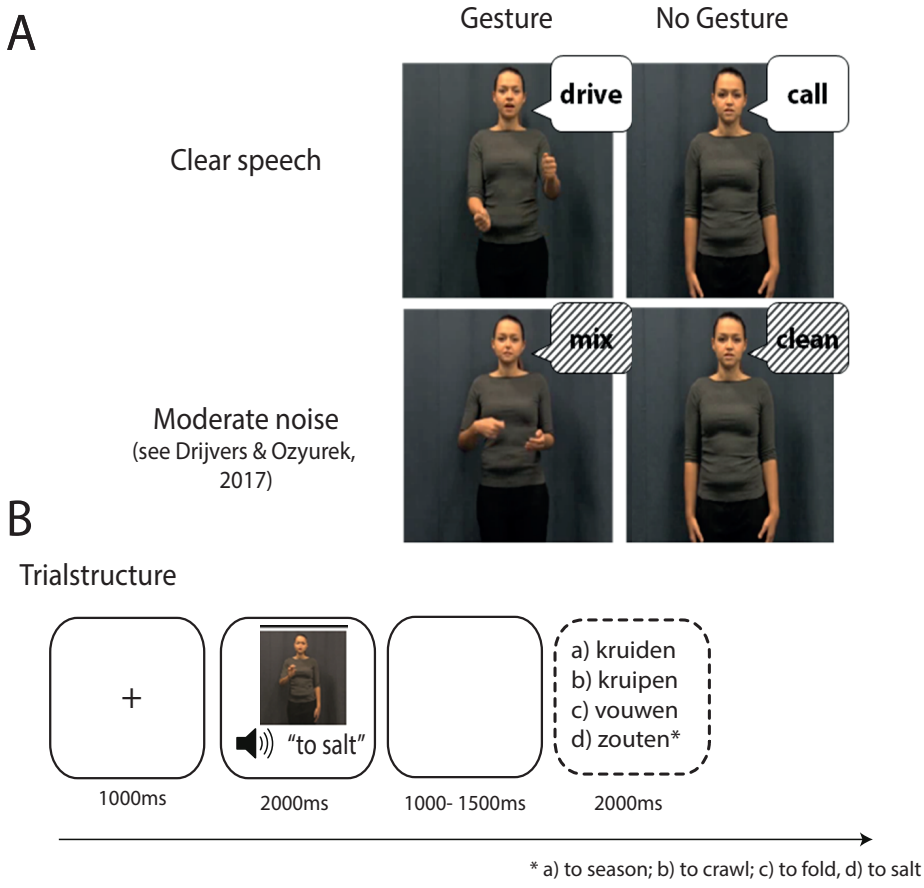


Figure 22 A: Illustration of the different conditions and stimuli, based on Drijvers et al., (2018a). B: Structure of a trial: Participants were presented with a fixation cross and watched and listened to the video. After a delay period, they had to indicate out of 4 options which verb they heard in the video.

We presented the stimuli in four conditions to probe gestural enhancement of degraded speech comprehension (as in Drijvers et al., 2018a): a clear speech condition with no gesture (CO, clear speech only), a degraded speech condition with no gesture (DO, degraded speech only), a clear speech condition with a matching gesture (CM, clear speech + matching gesture) and a degraded speech condition with a matching gesture (DM, degraded speech + matching gesture). All four conditions contained 40 videos, and none of the verbs overlapped in any condition.

7.3.3. Procedure

All participants were required to take an online LexTALE test to see whether they met the participation criteria. If a participant scored above 60%, the participant was invited for the MEG experiment. For the MEG experiment, participants were asked to attentively listen to and watch the videos. All participants were instructed that they would encounter a 4-alternative forced choice identification task after each video where they would be asked to indicate which verb they heard in the videos by means of a right-hand button-press on a 4-button box. We included the 4-alternative forced choice identification task to ensure that participants were paying attention to the videos, and to calculate whether these behavioral responses could be predicted by their oscillatory modulations (as was found for native listeners in Drijvers et al., 2018a).

Every trial started with a fixation cross (1000ms), which was followed by the experimental video (2000ms). After a short interval (1000 - 1500 ms, jittered), the subject had to indicate which verb they heard. Following the response, there was a 1000 ms pause upon which the next trial would start. The order of the stimuli was pseudo-randomized per subject, with the constraint that the same condition could not occur more than twice in a row along with the constraint that each video would only be presented once. The videos were divided into four mixed blocks of 40 trials each. After each block, the participants could take a self-paced break. If any significant head-movement occurred (> 5 mm), the experiment was paused and the subject was brought back to the original starting position.

7.3.4. MEG data acquisition

We followed all procedures described in Drijvers et al., (2018a). We recorded MEG with a 275-channel axial gradiometer CTF MEG system. All data were filtered online with a 300 Hz low pass filter, digitized at 1.2 kHz and stored for offline data analyses. The head position of the participants with respect to the gradiometers was measured by using three tracking coils (placed at the left and right ear canal and at the nasion to monitor head position in real-time (Stolk et al., 2013). Four channels of the CTF system were malfunctioning throughout all recordings (MLC11, MLC32, MLF62, MRF66). We recorded all participants' eye gaze by using

an Eyelink 1000 eyetracker, to monitor eye-blinks during the task. Participants' electrocardiogram (ECG) and horizontal and vertical electrooculogram (EOG) were recorded for artifact rejection purposes. A neck brace was applied to reduce head-movements in the MEG (Lozano-Soldevilla, Ter Huurne, Cools, & Jensen, 2014). In the MEG, the subject was positioned in a seated position at 70 cm distance to the screen, similar as in Drijvers et al., (2018a). All stimuli were back-projected onto a semi-translucent screen by using a PROPixx projector with a resolution of 1920x1080 and a refresh rate of 120 Hz. All stimuli were presented at full screen through Presentation software (Neurobehavioral Systems, Inc.).

7.3.5. MEG data analyses: preprocessing and time-frequency representations of power

We analyzed all MEG data in FieldTrip, an open-source MATLAB toolbox (Oostenveld et al., 2011), and followed the exact same procedure as in Drijvers et al., (2018a). The data were segmented into trials starting 1s before and ending 3 s after the onset of the video. The data was demeaned, detrended and band-stop filtered at 50, 100 and 150 Hz to remove any line noise that could contaminate the data. We then visually inspected the data for overt muscle artifacts, movement artifacts, SQUID jump artifact and other irregular artifacts. All trials with overt artifacts were rejected. We used a semi-automatic rejection routine and removed 4 trials per condition on average which were contaminated by SQUID jump artifacts and muscle artifacts. We then applied independent component analyses to attenuate the signals generated from eye-blinks, eye-movements and cardiac-related activity (Bell & Sejnowski, 1995). As a final step, we went through all single trials again to remove any artifacts that were not removed by ICA or our semi-automatic rejection procedure. We then resampled the data to 300 Hz to speed up the subsequent analyses. For a more intuitive interpretation of the data, we calculated a synthetic planar gradient, as planar gradient maxima are known to be located above the neural sources that may underlie them (Bastiaansen & Knösche, 2000). An approximation of the planar gradient was computed by converting the axial gradiometer data to orthogonal planar gradiometer pairs, and summing the power of the pairs.

7.3.6. Time-frequency analyses of power

The time-frequency analyses of power were the same as described in Drijvers et al., (2018a). Over a frequency range of 2 – 30 Hz, we applied a 500-ms Hanning window in frequency steps of 1 Hz and time steps of 50 ms. As we were interested in the gestural enhancement, we compared the difference in power in the Degraded speech + Matching Gesture (DM) and Degraded Speech (DO) conditions to the difference in the Clear Speech + Matching Gesture (CM) and Clear Speech (CO) condition. The power in these conditions was averaged separately for each participant and log10 transformed. We compared the within-group differences between the conditions (DO vs CO, DM vs. CM, DM vs. DO, CM vs. CO), by subtracting the log10 transformed power (i.e., the log ratio, log10(A) - log10(B). Similarly, to calculate the gestural enhancement effect, we calculated the difference between DM/DO and CM/CO as (log10(DM) - log10(DO) - (log10(CM) - log10(CO)). As a time-window of interest, we used the whole window in which speech and gesture were unfolding (0.7 – 2.0s) for both the within-group and between-group comparisons. To compare the effects of the non-native listeners to the native listeners, we compared this time-window of interest between groups in both the alpha (8-12 Hz) and beta (14 - 22 Hz) band. Note that in Drijvers et al., (2018a), effects in the gamma band were described. As we did not observe any modulations in the gamma band in the non-native listener group, we did not perform a between group comparison in the gamma band.

7.3.7. Source analyses

We estimated the sources of observed effects on sensor-level by using dynamic imaging of coherent sources (DICS, Gross et al., (2001)), a beamforming spatial filtering technique. Note that our source analysis served to localize the observed effects on sensor-level, but not to form an additional statistical assessment. Axial gradiometer data was used to perform these analyses. First, a spatial filter was calculated from the cross-spectral density matrix, as well as a lead field matrix. We constructed individual lead fields from our participants by using a realistically shaped single-shell head model based on the participants' own anatomical data from a segmented structural MRI, by dividing the brain volume in a 120 mm

spaced grid and warping it to a template brain (MNI).

All within-group source analyses used the time windows in which conditions were found to statistically differ in the sensor analyses. For the alpha band, we thus calculated the CSD at 10 Hz, with 2 Hz frequency smoothing. For the beta band, this effect was centered at 18 Hz, with 4 Hz frequency smoothing. Note that these settings, except for the time windows, are similar to the analyses described in Drijvers et al., (2018a). As the time-windows slightly differed for the non-native and native listeners, we performed a between-group comparison over the whole time window of interest, in both the alpha and beta frequency band to test for between-group differences. We used a common spatial filter over all conditions to project the data through. This common filter was then separately applied to each condition to calculate the power at each gridpoint. This was averaged over trials and log10 transformed. For visualization purposes, we interpolated the grand-average grid of all participants onto the template MNI brain.

7.3.8. Cluster-based permutation statistics

Non-parametric cluster-based permutation tests were performed across subjects to statistically assess oscillatory power differences between the different conditions and between the non-native and native listener group (Maris & Oostenveld, 2007). The source-level statistics were computed to create thresholded masks to localize any effects that were observed on the sensor-level. We computed the mean difference between two conditions for each x/y/z sample (source) or sample for sensor TFR analysis (sensor), in the frequency ranges of interest (alpha; 8 - 12 Hz, beta; 14 - 22 Hz, as determined by a grand-average TFR of all conditions combined) and time window of interest (0.7 - 2.0 s, from speech onset to video offset). After collecting the difference values of the comparisons, all adjacent values exceeding the threshold of 5% percent were grouped into clusters. This resulted in a distribution of different cluster candidates. The cluster candidates were randomly reassigned 5000 times across all conditions and participants. The cluster with the highest sum of difference values was added to a distribution, resulting in a permutation distribution. The observed cluster values were then compared to this newly created permutation distribution. The clusters that were

in the highest or lowest 2.5% were considered significant.

7.3.9. The relation between alpha and beta oscillations and behavioral 4-alternative forced choice identification scores.

In Drijvers et al., (2018a) we observed a clear correlation between oscillatory power modulations and the amount of gestural enhancement participants experienced during an externally induced adverse listening condition. As non-native listeners might choose different strategies to process the degraded speech signal or use the gestural information to enhance comprehension, we again correlated an individual's oscillatory power with the behavioral scores that we obtained from the 4-alternative forced choice identification task. We calculated this by averaging the power modulation over time points, frequencies and sensors in significant clusters of the interaction effects, which resulted in an individual's power modulation score per frequency band. For the behavioral scores, we calculated an interaction score for the reaction times and amount of correct answers, which was similar to how we calculated the gestural enhancement in oscillatory power. We computed difference scores between the conditions (e.g., DM-DO, CM-CO) and compared these differences to each other, resulting in the amount of behavioral enhancement per participant. Subsequently, we obtained Spearman correlation between these scores an individual's power modulation per frequency band. As our hypotheses were specific on the direction of the power modulation per frequency band, we used one-tailed tests.

7.4. Results

Highly-proficient non-native listeners of Dutch watched videos in which an actress would utter an action verb in clear or degraded speech, while making a gesture or not. After every video participants completed a 4-alternative forced choice identification task in which they identified what verb they heard in the video. We recorded MEG during the whole experiment, but were interested in oscillatory modulations of power in the alpha and beta frequency band while participants watched the videos, and how the oscillatory dynamics during this time interval related to their behavioral benefit on the 4-alternative forced choice

identification task.

Our analysis was twofold. First, similar to Drijvers et al., (2018a), we were interested in the behavioral responses as well as oscillatory modulations during gestural enhancement of degraded speech comprehension in non-native listeners (within group). Gestural enhancement was calculated as the interaction between the occurrence of a gesture (present/not present) and speech degradation (clear/degraded). Second, we compared the observed behavioral results and oscillatory modulations in non-native listeners to those observed in native listeners, as reported in Drijvers et al., (2018a) (between-group).

7.4.1. Behavioral results - Non-native listeners (within-group)

7.4.1.1. Gestural enhancement of speech comprehension is largest when speech is degraded

Non-native listeners experienced the most gestural enhancement when speech was degraded, mirroring earlier work on gestural enhancement in native speakers (Drijvers et al., 2018a), and behavioral work on gestural enhancement of degraded speech comprehension in non-native speakers (Drijvers & Ozyurek, 019). A repeated-measures ANOVA with the factors Noise-vocoding (clear speech vs. degraded speech) and Gesture (present vs. not present) on the percentage of correct answers revealed that participants were more able to correctly identify the verb in clear than in degraded speech ($F(1,29) = 246.896$, $p < .001$, $\eta^2 = .895$), and when a gesture was present compared to not present ($F(1,29) = 13.88$, $p = .001$, $\eta^2 = .324$). A significant interaction between Noise-vocoding and Gesture ($F(1,29) = 14.238$, $p = .001$, $\eta^2 = 329$), indicated that gestural enhancement was largest when speech was degraded. A similar pattern was observed in the reaction times, where listeners were faster when speech was clear than degraded ($F(1,29) = 121.38$, $p < .001$, $\eta^2 = .807$) and a gesture was present compared to not present ($F(1,29) = 41.629$, $p < .001$, $\eta^2 = .589$). Gestural enhancement was largest when gestures were present and speech was degraded ($F(1.29) = 15.113$, $p = .001$, $\eta^2 = .343$), which caused reduced reaction times that was more evident in degraded than in clear speech (see Figure 23).

7.4.2. Behavioral results - Non-native listeners vs. native listeners (between-group)

We compared the results of the two groups in a 2 (group; non-native / native) x 2 (gesture; present/not present) x 2 (noise-vocoding; clear/degraded speech) repeated-measures ANOVA for both the correct answers and the reaction times. The analysis of the correct answers revealed no significant differences on any of the interaction terms that contained the between-group factor, indicating that non-native listeners and native listeners had a similar number of correct answers on clear and degraded speech trials ($F(1,57) = 3.778$, $p = .057$) and trials containing a gesture or no gesture ($F(1,57) = 0.447$, $p = .507$). Gestural enhancement of degraded speech comprehension was not larger for native listeners compared to non-native listeners ($F(1,57) = 3.778$, $p = .306$).

The results of the reaction times revealed different results: native listeners were quicker to answer than non-native listeners on clear and degraded speech trials ($F(1,57) = 15.091$, $p = < .001$), as well as quicker to answer on gesture and no-gesture trials ($F(1,57) = 8.78$, $p = < .001$). Again, there was no three-way interaction of Gesture, Noise-vocoding and Group ($F(1,57) = 0.354$, $p = .554$), indicating that although native and non-native listeners show similar behavioral effects, non-native listeners overall answer more slowly than native listeners. In conclusion, our behavioral results thus revealed that although gestural enhancement of degraded speech comprehension was similar for native and non-native listeners, non-native listeners answered more slowly and were trending towards more incorrect answers.

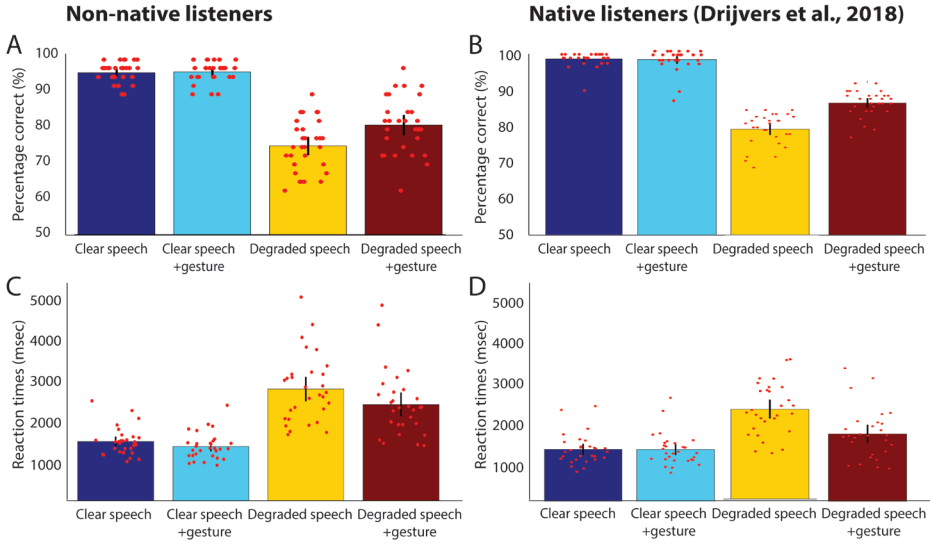


Figure 23 A/B: Percentage of correct answers per condition for non-native (A) and native listeners (B). Error bars represent SE. Red dots represent each individual participant's data. C/D: Reaction times (in milliseconds) per condition for non-native (C) and native listeners (D). Error bars represent SE. Red dots represent each individual participant's data. Gestural enhancement of degraded speech was similar for non-native and native listeners, but non-native listeners were significantly slower.

7.4.3. MEG results - Non-native listeners (within-group)

7.4.3.1. Alpha power is suppressed in pSTS/MTG, motor and visual regions during gestural enhancement of degraded speech comprehension in non-native listeners

Next we asked how oscillatory activity in the alpha band was modulated during gestural enhancement of degraded speech comprehension in non-native listeners. We first conducted a sensor-level analysis over the full time-window of interest (0.7 - 2.0, from speech onset until the end of the video) to test for an interaction effect between noise-vocoding and gesture occurrence. This 'gestural enhancement effect' was calculated by comparing the differences between the DMDO and CMCO contrasts (i.e., (log10(DM) - log10(D)) - (log10(CM) - log10(C)). Time-frequency representation (TFRs) of power of individual trials were calculated and averaged per condition. Figure 24A and Figure 24B represent the TFRs of power during gestural enhancement of degraded speech comprehension at representative

sensors within the non-native listener group. We then visualized the effect in time and space by plotting the topographical distribution of the interaction in the alpha band over time (see Figure 25A). Sensor-level analyses revealed that alpha power was more suppressed when speech was degraded and a gesture was present (one negative cluster, $p = 0.006$). This difference between DMDO and CMCO showed a central-parietal onset (0.7 - 1.0) that progressed over left-temporal and occipital (1.0 - 1.4) areas to right-temporal areas (1.4 - 2.0).

To localize the observed effect from the sensor analysis, we conducted source analyses to determine the underlying sources of the negative cluster. We applied a cluster-randomization approach to the source data and used the outcome of this analysis as a threshold for when to consider the source estimates reliable (the statistical assessment of the effect was thus formed by our sensor analyses, not the source analyses). As can be observed by the topographical alpha power distribution plots in Figure 25, the effect observed in the sensor analysis commences at left-central and parietal regions and progresses over left-temporal and occipital regions to right-temporal regions. We therefore assessed the sources of this cluster in three time windows instead of one to reliably capture the sources of the effect. In the first time-window, from 0.7 - 1.0 s, we observed a larger alpha power suppression when gestures enhanced degraded speech comprehension over STS/MTG, pre/postcentral regions and angular gyrus ($p = 0.04$, one negative cluster, see Figure 24C). In the time window from 1.0 - 1.4 s, we observed a larger alpha power suppression over left-temporal (pSTS/STG/MTG) and occipital regions ($p = 0.02$, one negative cluster, see Figure 24D), and in a final time window from 1.4 to 2.0 s we observed a larger alpha power suppression in right-temporal (pSTS/STG/MTG) regions and, the left temporoparietal junction and left- and right-occipital regions ($p = 0.004$, one negative cluster, Figure 24E).

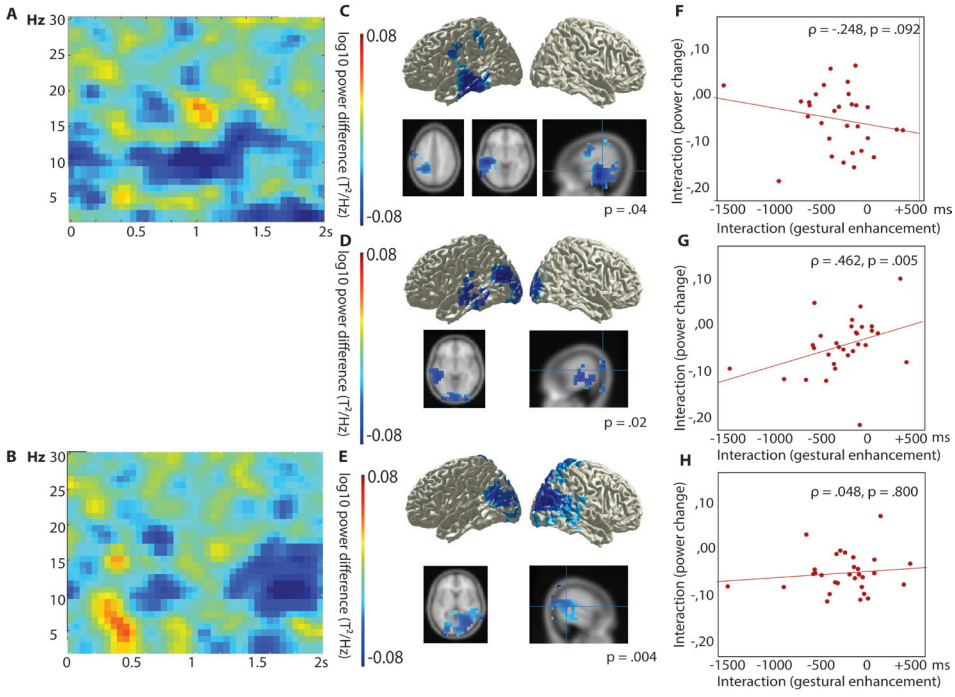


Figure 24 A: Time-frequency representation of power at representative left-temporal sensors, capturing both the alpha effect from the first time window (0.7-1.0s) as the second time window (1.0 - 1.4 s). B: Time-frequency representation of power at a representative cluster formed by channels from right-temporo-occipital regions, capturing the late alpha effect (1.4 - 2.0 s). C: Estimated source results of the first alpha cluster, masked by statistically significant clusters. D: Estimated source results of the second alpha cluster, masked by statistically significant clusters. E: Estimated source results of the third alpha cluster, masked by statistically significant clusters. F: Individual's alpha power modulation in the first time window as a function of individual's gestural enhancement in the 4-alternative forced choice identification task. G: Individual's alpha power modulation in the second time window as a function of individual's gestural enhancement in the 4-alternative forced choice identification task. H: Individual's alpha power modulation in the third time window as a function of an individual's gestural enhancement in the 4-alternative forced choice identification task.

7.4.3.2. Left-temporal individual alpha power modulations predict gestural benefit during degraded speech comprehension in non-native listeners

We correlated a participant's individual power modulations in the alpha band with the benefit from gestures participants experienced during the behavioral task. Here, we reasoned that although the accuracy on the 4-alternative forced choice identification task might attenuate behavioral scores due to the nature of the task, reaction times circumvent this problem. As the answers to our task were cued, participants may have just selected the right answer and understood the verb when the answers were presented. However, we expected that gestural enhancement should also speed up reaction times. Therefore, we calculated an individual's speeding or slowing caused by gestural enhancement by calculating the difference in reaction times between DMDO and CMCO and correlating it with an individual's power modulation. To investigate whether the power modulations that were estimated over specific regions were predictive of gestural enhancement of degraded speech comprehension on the behavioral task, we correlated an individual's behavioral scores with power modulations in the three different time windows. These analyses revealed that the more a listener's alpha power was suppressed in the second time window, the more a listener benefitted from gestural enhancement of degraded speech comprehension (1.0 - 1.4 s, Spearman's rho = .462, p = .005, one-tailed, see Figure 24F). Note that this is the time window where the meaningful part of the speech and the meaningful part of the gesture are unfolding. This correlation was not found in the early time window (0.7 - 1.0 s, Spearman's rho = -.248, p = .092 one-tailed, Figure 24G), nor in the late time window (1.4 - 2.0 s, Spearman's rho =.048, p = .80, Figure 24H). This indicates that the individual power modulations over left-temporal regions and occipital regions predict the behavioral benefit of a gesture a non-native listener experiences during degraded speech comprehension.

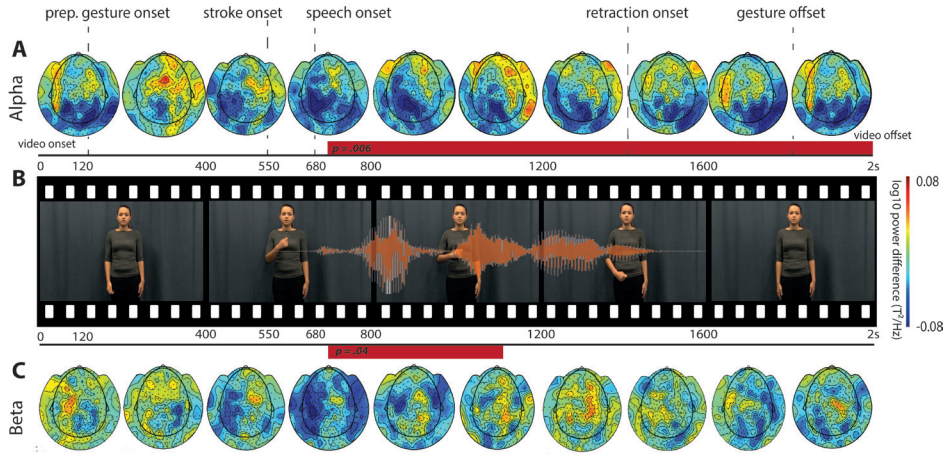


Figure 25 A: Topographical distribution of alpha power over the whole video interval for the gestural enhancement effect, binned per 200 ms. Red bar on timeline represents significant cluster in sensor-level analysis. B: Structure of the video. Orange waveform represents speech. C: Topographical distribution of beta power over the whole video interval for the gestural enhancement effect, binned per 200 ms. Red bar under timeline represents the significant cluster of the sensor analysis.

7.4.3.3. Beta power is more suppressed over LIFG and motor regions when gestures enhance degraded speech comprehension

Next we followed a similar procedure when we analyzed sensor-level differences in the beta band (14 - 22 Hz, range determined on grand-average TFR of all conditions combined, see Figure 26A). We studied the spatiotemporal course of the effect by plotting the topographical distribution of the gestural enhancement effect (Figure 25A). We there observed a left-lateralized effect in an early time window. Sensor-level analyses of the interaction effect indeed confirmed that beta power was more suppressed when speech was degraded and a gesture was present (one negative cluster, $p = 0.04$). This effect occurred when the stroke of the gesture and speech were unfolding (0.7 - 1.1 s).

We then used source-analysis to estimate the source of the gestural enhancement effect. These analyses demonstrated that the larger beta suppression could be localized to the LIFG, and left pre- and post-central gyrus (one negative cluster, $p = .03$, Figure 26B).

7.4.3.4.Non-native listener's individual beta power in motor cortex and LIFG predicts gestural benefit during degraded speech comprehension.

We correlated a listener's individual beta power with the amount of speeding/slowing a listener experienced during gestural enhancement of degraded speech comprehension in the 4-alternative forced choice identification task, and observed a correlation between the amount of beta suppression in the motor cortex/LIFG and the behavioral scores: the more a listener's beta power was suppressed over motor regions and LIFG, the more a listener could benefit from gestural enhancement of degraded speech comprehension (Spearman's rho = .438, p = 0.008, one-tailed, Figure 26C).

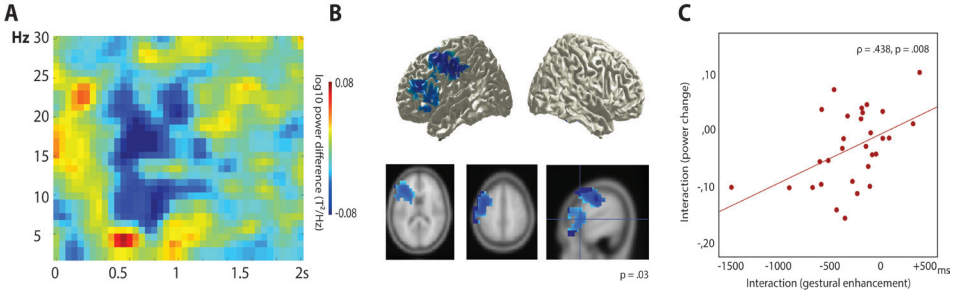


Figure 26 A: Time-frequency representation of power at representative left-frontal/left-motor sensors. B: Estimated source results of the beta cluster, masked by statistically significant clusters. C: Individual's beta power modulation as a function of individual's gestural enhancement in the 4-alternative forced choice identification task.

7.4.4. MEG results - Non-native listeners vs. native listeners (between-group)

7.4.4.1. Native listeners' alpha power is more suppressed in LIFG and ATL than in non-native listeners

We then compared the results of the non-native listeners to the results of the native listeners reported in Drijvers et al., (2018a) to test for between-group differences in the gestural enhancement effect. To this end, we first calculated sensor-level differences in the alpha band (8-12 Hz) between native and non-native listeners by comparing the gestural enhancement effect in the time window of interest (0.7-2.0 s). Here we observed a larger alpha power suppression reflecting the difference

in the gestural enhancement effect over left-frontal regions for native compared to non-native listeners over the entire time window (0.7 - 2.0 s, one negative cluster, p = .02, Figure 27A).

We then estimated the source of this difference in alpha power between the groups and observed a larger alpha power suppression for native than non-native listeners in LIFG and ATL (one negative cluster, p = .04, Figure 27B).

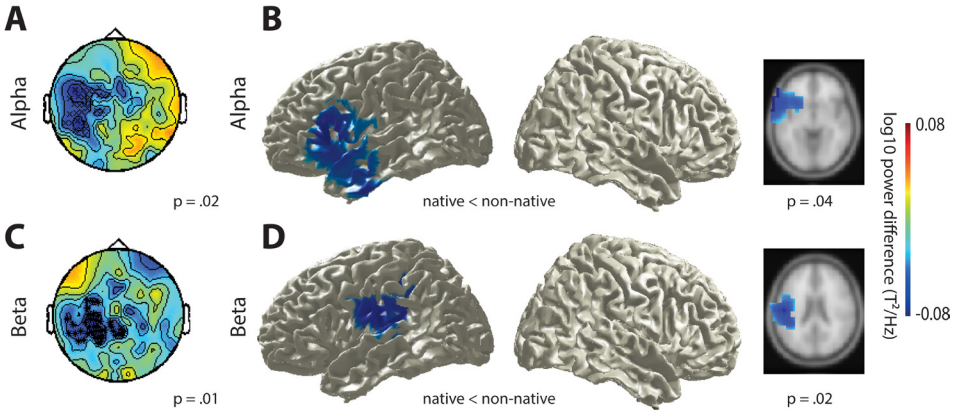


Figure 27 A: Topographical plot of difference in alpha power between non-native and native listeners on sensor level. B: Estimated source results of the alpha cluster, masked by statistically significant clusters. C: Topographical plot of difference in beta power between non-native and native listeners on sensor level. D: Estimated source results of the beta cluster, masked by statistically significant clusters.

7.4.4.2. Native listeners' beta power is more suppressed over primary somatosensory cortex than in non-native listeners.

On sensor-level, we observed a larger beta power suppression (14 - 22 Hz) for native compared to non-native listeners over the whole time window (0.7 - 2.0 s, one negative cluster, p = .01, 27C). The source of this effect was estimated over primary sensory cortex and left insula (one negative cluster, p = .02, Figure 27D).

7.5. Discussion

We set out to investigate the spatiotemporal oscillatory dynamics that support gestural enhancement of degraded speech comprehension in non-native listeners, and how these oscillatory modulations compare to earlier results observed in

native listeners (Drijvers et al., 2018a). Using this manipulation, we investigated how much benefit from gesture a listener has when resolving language in both externally induced (speech degradation) and internally induced (non-nativeness) adverse listening conditions.

Behaviorally, we observed a similar gestural enhancement effect for non-native listeners as for native listeners. Although the gestural enhancement effect was similar for both groups, non-native listeners were significantly slower in providing their answers to the 4-alternative forced choice identification task.

When gestures enhanced degraded speech comprehension for non-native listeners, we observed a larger alpha power suppression that commenced at central-parietal regions, and which over time was observable in left-temporal, occipital and right-temporal regions. We observed an early beta power suppression in LIFG and motor regions during gestural enhancement of degraded speech comprehension. We found distinct correlations for two successive time windows in the beta (0.7 - 1.1) and alpha band (1.0 - 1.4) (i.e., the time window in which the meaningful part of the gesture and speech were unfolding), that revealed that an individual's power modulations could predict an individual's gestural benefit when resolving language in adverse listening conditions.

Importantly, when comparing the gestural enhancement effect of non-native listeners to the gestural enhancement effect observed in native listeners, native listeners demonstrated more alpha suppression in LIFG and ATL, as well as a larger beta power suppression in primary somatosensory cortex and left insula. Below we discuss the putative role of these spatiotemporal effects during gestural enhancement of speech comprehension in internally and externally induced adverse listening conditions.

7.5.1. Non-native listeners (within-group)

7.5.1.1. Non-native listeners who more strongly engage motor regions and LIFG benefit more from gestures during degraded speech comprehension

In an early time window (0.7 - 1.1 s), we observed a larger beta power suppression in

LIFG and motor regions when gestures enhanced degraded speech comprehension in non-native listeners. This might suggest that the motor system is engaged to simulate the observed gesture more strongly when speech is degraded (Klepp et al., 2015; van Elk et al., 2010; Weiss & Mueller, 2012), possibly to extract meaningful information to aid ongoing degraded speech comprehension.

The observed beta power suppression extended to LIFG. Previous studies have argued that LIFG might be sensitive to semantic unification, such as the unification of information from different modalities, or lexical access operations (Hagoort, 2013). Other work demonstrated that LIFG is more engaged when the language processing system is faced with a higher unification load, for example because of semantic congruency (Wang, Jensen, et al., 2012; of gestures, see Drijvers et al., 2018a) or speech degradation (Hervais-Adelman, Carlyon, Johnsrude, & Davis, 2012; Obleser et al., 2007; Wild et al., 2012). In line with these findings, we suggest that LIFG is more engaged in this time window to unify gestural information with the degraded speech signal.

A listener's individual oscillatory power modulation correlated with a listener's individual behavioral benefit of gestural information during degraded speech comprehension: the more an individual listener's beta power was suppressed in this time window, the more gestural benefit a listener experienced in the 4-alternative forced choice identification task.

7.5.1.2. A left-lateralized network of motor regions, AG, pSTS/MTG and STG is engaged during gestural enhancement of degraded speech comprehension

In the same time window, we observed a stronger alpha power suppression (0.7 - 1.0 s) in pSTS/STG/MTG, left motor regions and left angular gyrus. Activation of the pSTS/MTG has been repeatedly found in studies on speech-gesture integration, and is thought to reflect an initial matching of the audiovisual stimuli (Dick et al., 2012, 2014; Holle et al., 2010; Willems et al., 2007, 2009). In line with these studies, we propose that the stronger alpha power suppression might reflect early engagement of the language system to perform an initial integration of lower-level characteristics of the audiovisual input.

Note that the abovementioned effects in the beta band (0.7 - 1.1 s) occur in a similar time window as the current effect in the alpha band (0.7 - 1.0 s), but that the beta, and not the alpha (0.7 - 1.0 s) effects correlate with gestural benefit during degraded speech comprehension. This confirms that the alpha band effect indeed might reflect an initial matching of, possibly lower-level, audiovisual information, that is similar for all non-native listeners and does not relate to the gestural enhancement that a listener experiences per se. The engagement of the AG, which is often seen as an association and supramodal integration hub (Binder, Desai, Graves, & Conant, 2009), and the motor system, which engages more strongly during gestural enhancement of degraded speech, might aid in this integration process.

7.5.1.3. Non-native listeners who more strongly engage visual regions and left-temporal regions, experience more gestural benefit during degraded speech comprehension

In contrast to the alpha effect in the first time window (0.7 - 1.0 s), an alpha effect in the subsequent time window (1.0 - 1.4s) over pSTS/STG/MTG *did* predict how much gestural benefit a non-native listener experiences during gestural enhancement of degraded speech comprehension. We suggest that a listener's power modulation in this time window is predictive of an individual's gestural benefit during degraded speech comprehension because the semantic information that is being unified with the gestures in an earlier time window, as was demonstrated by our beta effects (0.7 - 1.0 s), aids subsequent lexical access of the degraded input (Hagoort, 2013; Lau et al., 2008). Post-hoc power-power correlations between an individual's beta power in the early time window (0.7 - 1.0 s) and an individual's alpha power in the second time window (1.0 - 1.4 s) concur with this proposed interpretation: listeners who more strongly show beta suppression in the first time window, also show a larger alpha suppression in the second time window (Spearman's rho = .408, p = 0.013, one-tailed). Similarly, listeners who demonstrated a larger alpha suppression over visual regions might have allocated more visual attention to the gestures when speech is degraded to aid comprehension.

7.5.1.4. Right-temporal regions engage when more neural resources are recruited for comprehension.

We then observed a larger alpha power suppression in the alpha band (1.4 - 2.0s) over right-temporal and right-occipital regions. An individual's power modulation in this time window did not correlate with subsequent comprehension on the 4-alternative forced choice identification task. fMRI studies have suggested that right-lateralized regions are often recruited during non-native language processing (Higby, Kim, & Obler, 2013; Leonard et al., 2011). This might suggest that non-native listeners try to recruit more top-down information to facilitate comprehension and unification of the two input streams (Skipper et al., 2006, 2007). We suggest that these right-lateralized regions might be more engaged when non-native listeners require more neural resources for comprehension, especially when auditory cues are not reliable enough to map the semantic information from the gesture to. This is also in line with previous results where we observed right-lateralized effects when comparing matching and mismatching gestures (Drijvers & Ozyurek, 2018), where non-natives seemed to recruit additional resources to process mismatching semantic information.

7.5.2. Non-native listeners vs. native listeners (between-group)

In line with the behavioral results observed in native listeners (Drijvers et al., 2018a), we observed that for non-native listeners gestural enhancement was largest when speech was degraded. This gestural enhancement effect was similar for non-native and native listeners. However, we observed that non-native listeners were significantly slower in answering on the 4-alternative forced choice identification task. As participants were cued with four answering options, it might have been easier to recognize the degraded verb during this answering period than when the participants watched the video. This might have masked the actual comprehension difficulties that the listeners experienced in the video. However, this does not affect the reaction times. When a non-native listener for example might have experienced more difficulty in understanding the speech during the

video, it might take longer to find the correct answer in the 4-alternative forced choice identification task. These results thus indicate that although the gestural enhancement effect seemed similar for native and non-native listeners, non-native listeners possibly were sometimes hindered in processing and coupling the degraded auditory information to the semantic information conveyed by the gesture.

7.5.2.1. Non-native listeners might face more difficulty when retrieving gestural semantic information and unifying it with degraded auditory cues than native listeners

Finally, we compared the oscillatory modulations observed in non-native listeners to the modulations observed in native listeners during gestural enhancement of degraded speech comprehension. We observed a larger alpha power suppression for native listeners in LIFG and anterior temporal lobe (ATL) than for non-native listeners. As the ATL has been implicated as a domain-general semantic hub (Wong & Gallate, 2012) and we have found converging evidence for the engagement of the LIFG during the unification of degraded speech and gestures (Drijvers et al., 2018a, 2018b), this suggests that when speech is degraded, it might be more difficult for non-native listeners to access the semantic information from the gesture and unify it to the degraded auditory cues. This difficulty in semantic retrieval might be due to the fact that non-native listeners need more available auditory cues to facilitate access to the semantic cues conveyed by the gesture. In turn, this might cause these areas to be less engaged in non-native listeners than in native listeners, and might explain the slower reaction times in the 4-alternative forced choice identification task, despite the similar gestural enhancement effect that was observed.

7.5.2.2. Non-native listeners might face more difficulty utilizing phonological information that is conveyed by visible speech to aid degraded speech comprehension

We observed a larger beta power suppression over primary somatosensory cortex and left insula for native compared to non-native listeners during gestural

enhancement of degraded speech comprehension. Specifically, the lower part of the somatosensory cortex that is sensitive to information from visible speech (i.e., information conveyed by teeth, tongue and lip movements) was less engaged in non-native than native listeners, possibly because non-native listeners are less able than native listeners to simulate the information conveyed by visible speech when speech is degraded. The cluster overlapped with the left insula, which has been shown to be sensitive to the strength of cross-modal binding (Bushara et al., 2002) as well as being involved in phonological processing (Abdullaev & Melnichuk, 1997; Bamiou, Musiek, & Luxon, 2003; Booth, Wood, Lu, Houk, & Bitan, 2007; Tettamanti et al., 2005; Wild et al., 2012), suggesting that the observed effects are consistent with the idea that non-native listeners might face more difficulty using the phonological information that is conveyed by visible speech to aid degraded speech comprehension.

7.5.2.3. How does gestural enhancement of degraded speech comprehension differ for native and non-native listeners?

Our current results revealed a different spatiotemporal profile during gestural enhancement of degraded speech comprehension for listeners who are faced with internal ambiguity (non-native listeners) compared to listeners who are not faced with internal ambiguity (in Drijvers et al., (2018a) for native listeners). In the native listeners described in Drijvers et al., (2018a), we observed an alpha power suppression over right STS (0.7 - 1.0 s), followed by an alpha/beta power suppression in left-motor regions (1.0 - 1.6), left-temporal and occipital regions (1.6 - 2.0s). We observed a larger beta power suppression over the motor cortex and extended language network (1.3 - 2.0 s) and a larger gamma power increase over MTL (1.0 - 1.5s). These power modulations correlated with an individual's behavioral benefit in the 4-alternative forced choice identification task and were suggested to support general unification, integration and lexical access processes during language comprehension, as well as simulation of and increased visual attention to iconic gestures over time.

The observed oscillatory modulations in non-natives suggest similar core processes that support gestural enhancement of degraded speech comprehension

as were observed in native listeners in Drijvers et al., (2018a). However, the different spatiotemporal time course of the effects observed in non-native compared to native listeners might suggest that the two listener groups employ different processing strategies. For example, non-native listeners seem to immediately engage motor cortex and LIFG to extract semantic information from the gesture, and might attempt to immediately unify this information with the signal to aid retrieval of the degraded input. Subsequently, when this integration is hindered, non-native listeners engage additional resources by engaging right-temporal regions to aid in comprehension of ambiguous information. Alternatively, native listeners seem to have more access to the degraded phonological information than non-native listeners and might therefore be less hindered in using the semantic information from the gestures to resolve the degraded input. They can therefore already optimize their processing strategy in an early time window, whereas non-native listeners are not able to do this as it is more difficult for them to access the degraded input to map the semantic information from the gesture to. This is in line with unimodal, behavioral studies that investigated the effects of auditory semantic context on non-native degraded speech comprehension (Bradlow & Alexander, 2007; Golestani, Rosen, & Scott, 2009; Hazan et al., 2006; Mayo, Florentine, & Buus, 1997; Oliver, Gullberg, Hellwig, Mitterer, & Indefrey, 2012; Zhang et al., 2016), and fits with our previous behavioral study (Drijvers & Ozyurek, under review) and EEG results (Drijvers & Ozyurek, 2018).

7.6. Conclusion

Our data revealed that spatiotemporal oscillatory dynamics can predict how much a listener benefits from semantic information conveyed by gestures when speech comprehension is challenged by internally (e.g. non-nativeness) and externally (e.g., speech degradation) induced adverse listening conditions. Our behavioral results suggested that although native and non-native listeners revealed a similar gestural enhancement effect in the 4-alternative forced choice identification task, non-native listeners were significantly slower than native listeners when indicating which verb they heard in the video. This suggests that non-native listeners possibly faced more difficulty unifying the degraded auditory cues with

the semantic information conveyed by gestures. In line with this interpretation, the observed oscillatory modulations in both non-native and native listeners suggest similar core processes that support unification and lexical access processes, as we well as simulation of the gesture and increased visual attention to gestures to aid degraded speech comprehension. However, compared to native listeners, non-native listeners might have less access to the phonological information in the degraded signal, as demonstrated by less engagement of the mouth area of the primary somatosensory cortex and left insula. Moreover, non-native listeners might experience more difficulty unifying the semantic information conveyed by the gesture with the speech signal, causing areas that are involved in unification and retrieval (i.e., LIFG and ATL) to be less engaged.

7.7. Acknowledgements

This work was supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from Netherlands Organization for Scientific Research. OJ was supported by James S. McDonnell Foundation Understanding Human Cognition Collaborative Award [220020448] and the Royal Society Wolfson Research Merit Award. We are very grateful to Nick Wood (†), for helping us in editing the video stimuli, and to Gina Ginos, for being the actress in the videos.

Chapter 8

Visual attention to gestures reflects processing differences in native and non-native listeners during degraded speech comprehension

8.1. Abstract

Visual information conveyed by iconic hand gestures and visible speech can enhance speech comprehension under adverse listening conditions for both native and non-native listeners. However, listeners often mostly gaze at a speaker's face, but rarely at their gestures. We used eye-tracking to investigate whether and how native- and non-native listeners of Dutch allocated overt visual attention to these visual articulators during clear and degraded speech comprehension. Participants watched video clips of an actress uttering clear or degraded (6-band noise-vocoded) action verbs while performing a gesture or not, and were asked to indicate the word they heard in a 4-alternative forced choice identification task. Gestural enhancement was largest when speech was degraded for all listeners, but more strongly for native listeners. Both native and non-native listeners mostly gaze at the face during comprehension, but non-native listeners gazed more often at gestures than native listeners, possibly to extract semantic information to aid speech comprehension. However, only a native listener's gestural benefit during degraded speech comprehension could be predicted by an individual's gaze allocation to gestures. We conclude that it might be more challenging for non-native listeners to resolve the degraded auditory cues and couple those cues to phonological information that is conveyed by visible speech. This diminished phonological knowledge might hinder and delay the use of semantic information that is conveyed by gestures for non-native listeners. As native listeners are more able to utilize and resolve degraded auditory cues than natives, they can map more visual information to the speech signal, resulting in better speech comprehension, especially when speech is degraded.

This chapter is based on Drijvers, L., Vaitonyte, J., Ozyurek, A. (in revision). Visual attention to gestures reflects processing differences in native and non-native listeners during degraded speech comprehension

8.2. Introduction

In everyday conversational contexts, we often communicate in challenging or adverse listening conditions. These listener-related challenges can emerge because of external factors, such as noise (Peelle, 2018), but also because of internal factors, such as when communicating in a non-native language (Lecumberri et al., 2010). Everyday conversational contexts are often multimodal and can include auditory inputs, such as speech, but also visual input, such as visible speech and gestures. These visual inputs can aid challenges that listeners face during speech comprehension. Visible speech, which consists of movements of tongue, teeth and lip movements, has been shown to enhance clear and degraded speech comprehension for both native and non-native listeners (Erber, 1975; Munhall, 1998; Navarra & Soto-Faraco, 2007; L. a Ross et al., 2007; Sumby & Pollock, 1954), as it can provide phonological information about the speech signal that can enhance comprehension. Next to visible speech, iconic hand gestures, that can convey semantic information about object attributes, actions and space (McNeill, 1992), can also enhance clear and degraded speech comprehension (Drijvers, Özyürek, & Jensen, 2018; Holle, Obleser, Rueschemeyer, & Gunter, 2010), especially in a joint context with visible speech (Drijvers & Özyürek, 2017), and for both native and non-native speakers (Drijvers & Ozyurek, in press). However, little is known about how native and non-native listeners allocate their visual attention to benefit from these inputs during speech comprehension in a joint context, especially in adverse listening conditions, such as when speech is degraded.

Understanding speech in the presence of noise has been demonstrated to be more difficult for non-native than for native listeners (Bradlow & Alexander, 2007; Brouwer, Van Engen, Calandruccio, & Bradlow, 2012; Kilman, Zekveld, Hällgren, & Rönnberg, 2014; Mayo et al., 1997; Scharenborg, Coumans, & van Hout, 2018), even when non-native listeners are highly proficient (Cutler, Garcia Lecumberri, & Cooke, 2008). As noise decreases the available acoustic information in the speech signal, it might be more difficult for non-native listeners to make a phonological mapping between the speech signal and perceptual/linguistic representations, as these might have not been fully tuned to the non-native language (Flege, 1992; Iverson et al., 2003; Lecumberri et al., 2010). Specifically in such situations, visual

phonological information that is conveyed by visible speech has been shown to enhance non-native language learning and comprehension (Hannah et al., 2017; Jongman et al., 2003; Kawase et al., 2014; Kim, Sonic, & Davis, 2011; Wang et al., 2008). In native listeners, it has been suggested that visual attention is more often directed to the mouth of a talker to extract more information from visible speech when speech is degraded (Buchan, Paré, & Munhall, 2007; Król, 2018; Munhall, 1998; Rennig, Wegner-Clemens, & Beauchamp, 2018). However, it has not been studied how non-native listeners might allocate their visual attention to benefit from visible speech information in adverse listening conditions.

Visible speech is not the only visual information source that can aid comprehension. Listeners often perceive visual input that not only consists of visible speech, but also iconic hand gestures, which can convey semantic information about the speech signal. Previous work has demonstrated that listeners integrate this semantic information with speech (Kelly, Creigh, et al., 2010) and that both native and non-native listeners can benefit from semantic information that is conveyed by gestures (Dahl & Ludvigsen, 2014; Sueyoshi & Hardison, 2014), especially when speech is degraded (Drijvers & Özyürek, 2017; Drijvers et al., 2018; Drijvers & Özyürek, 2018).

To date, there is no work that investigated whether listeners allocate overt visual attention to gestures when speech is degraded to aid comprehension. In clear speech, native speakers tend to fixate on the speaker's face during multimodal language comprehension for 90-95% of the time (Argyle & Cook, 1976; Argyle & Graham, 1976; Gullberg & Kita, 2009), and therefore tend to not gaze to gestures (Gullberg & Holmqvist, 1999; 2002; 2006; Gullberg & Kita, 2009), except when speakers look at their own gestures (Gullberg & Holmqvist, 2006), a gesture is produced in the periphery of gesture space (McNeill, 1992), or when a gesture moves into a hold before moving on (Gullberg & Kita, 2009). Although gestures thus convey meaningful information, listeners seem to be able to abstract this information without directly fixating on them (Gullberg & Kita, 2009), which is in line with findings from the sign language domain, where signers fixate on the face more than other visual cues, such as the hands (Agrafiotis, Canagarajah, Bull, & Dye, 2003; Emmorey, Thompson, & Colvin, 2009; Muir & Richardson, 2005).

However, this could be different for non-native listeners, especially when speech is degraded. As non-native listeners are more hindered by noise compared to native listeners (Mayo et al., 1997; Bradlow & Alexander, 2007), they might rely more strongly on visual semantic information conveyed by gestures, especially when their phonological knowledge of a language is not sufficient (Hazan et al., 2006). For example, previous EEG work suggested that non-native listeners might focus more on gestures when speech is clear than native listeners (Drijvers & Özyürek, 2018). In this study, we investigated modulations of the N400 component during speech-gesture integration in clear and degraded speech, in a native and non-native listener group by using a violation paradigm where gestures either matched or mismatched the speech signal. The N400 component is thought to be sensitive to semantic unification operations (Kutas & Federmeier, 2014). We observed an N400 effect when comparing mismatching to matching gestures in both clear and degraded speech for native listeners. However, we only observed an N400 effect in clear speech for non-native listeners, but not in degraded speech. This N400 effect in clear speech was larger for non-native listeners than native listeners, which might indicate that they focus more strongly on gestures than native listeners, to extract semantic information to aid comprehension. Similarly, previous neuroimaging work has indicated that both native and non-native listeners engage their visual cortex more when speech is degraded and a gesture is present than when speech is clear or no gesture is present, possibly to allocate more visual attention to gestures and increase information uptake (Drijvers, Ozyurek & Jensen, 2018; Drijvers, Van der Plas, Ozyurek & Jensen, in revision). Non-native listeners however engage areas involved in semantic retrieval and semantic unification less than native listeners during gestural enhancement of degraded speech, suggesting that non-native listeners might be hindered in integrating the degraded phonological cues with the semantic information conveyed by the gesture. However, so far, it is unknown if this is in any way reflected in the overt visual attention that listeners allocate to visible speech or to gestures, and if overt visual attention to gestures correlates with an enhancement in speech comprehension.

8.2.1. The present study

The current study investigates how native and non-native listeners allocate visual attention to visible speech and gestures in clear and degraded speech. More specifically, we aim to gain insight into whether allocating gaze towards gestures when speech is degraded can predict how much a listener benefits from gestural information during comprehension, and whether and how this may differ for native compared to non-native listeners. To investigate this, we used eye-tracking to record eye movements with a high temporal resolution. Eye-tracking provides an excellent method to study how signal degradation affects online processes of native and non-native word comprehension (e.g., Brouwer & Bradlow, 2016; McQueen & Huettig, 2012; see for a review: Van Engen & McLaughlin, 2018), and to study how allocating attention is reflected in gaze behavior (Posner, 2016).

To investigate these questions, we presented native and non-native participants with videos in which an actress uttered a verb in clear or degraded speech, and while making a gesture or not. All participants completed a behavioral 4-alternative forced choice identification task after each item that asked which verb they had heard in the videos. We were interested in their accuracy results and reaction times, as well as gaze allocation to the face, the mouth, and the body during clear and degraded speech comprehension, as measured by the proportion of fixations to these areas of interest.

8.2.1.1. Behavioral hypotheses

On the behavioral-4-alternative forced choice identification task, we expected that both listener groups would benefit more from gestures when speech is degraded than when speech is clear. We predicted that this benefit would be larger for native listeners than for non-native listeners. This would then be reflected by a higher accuracy level and faster reaction times, as well as a larger gestural enhancement effect during degraded speech comprehension.

8.2.1.2. Eye-tracking hypotheses - face & mouth

In the eye-tracking results, we expected that within the two listener groups, both native and non-native listeners would look more at the face and mouth when speech was degraded than when speech was clear, irrespective of whether a gesture was present or not (Buchan et al., 2007; Król, 2018; Munhall, 1998; Rennig et al., 2018).

However, previous literature suggested that non-native speakers might have difficulties in resolving phonological information in the speech signal when speech is degraded (Cutler, Weber, Smits, & Cooper, 2004; Krizman, Bradlow, Lam, & Kraus, 2016) and are aided by extra visual information to resolve the degraded phonological input (Hazan et al., 2006). We therefore expected that when comparing the two groups, non-native listeners would look more to the face and mouth than native listeners, especially when speech was degraded. However, in the presence of gesture, non-native listeners' gaze allocation to visible speech might be less pronounced, as will be outlined below.

8.2.1.3. Eye-tracking hypotheses - gesture

We had similar expectations for gaze allocation to gesture. Although previous literature suggested that listeners do not often gaze overtly to gesture in natural communication, as they are able to extract visual semantic information peripherally (Gullberg & Holmqvist, 1999, 2002, 2006; Gullberg & Kita, 2009), we did expect that within the two listener groups both native and non-native listeners gaze more at gestures when speech was degraded than when speech was clear. This would be in line with previous research that suggested that more visual attention is allocated to gestures to increase the uptake of gestural information when speech is degraded (Drijvers, Ozyurek & Jensen, 2018; Drijvers, Van der Plas, Ozyurek & Jensen, in revision).

When comparing the two groups, we expected that non-native listeners would gaze more at gestures than native listeners in both clear and degraded speech. As non-native listeners might find it difficult to couple the phonological information conveyed by visible speech to the speech signal, they might try to

increase their visual semantic information uptake to aid comprehension. This might also result in sustained visual attention to gestures, especially when the phonological information conveyed by visible speech is difficult to resolve. We explored this option by using cluster-based permutation tests to analyze the exact moment when gaze allocation patterns might diverge within and between groups during comprehension, which is not possible by using conventional eye-tracking analysis methods. Finally, we were interested in whether gaze allocation to gestures could predict the gestural enhancement the listeners experienced during comprehension. We expected that gaze allocation to gestures during degraded speech comprehension would predict the benefit a listener experiences from the gesture during the 4-alternative forced choice identification task in both groups.

8.3. Methods

8.3.1. Participants

Twenty Dutch participants (mean age = 26.0, SD = 7.58) and twenty-one German advanced learners of Dutch (mean age = 23.05, SD = 2.62) with no neurological, language, hearing or motor disorders participated in the experiment. All participants were right-handed and obtained education at a University level.

The non-native German advanced learners of Dutch were recruited on the basis of the following inclusion criteria: (i) having lived/studied in the Netherlands for at least a year, (ii) having used Dutch for at least once a week, (iii) acquired Dutch after age 12. On average, the German participants acquired Dutch between 12 and 22 years (mean age = 18.25, SD = 2.8) as part of their preparation for a Dutch educational program. One of the German participants had to be excluded because the inclusion criteria were not met. All participants gave informed written consent before participating and received a financial compensation for participation.

8.3.1.1. LexTALE assessment

We used the Dutch version of the Lexical Test for Advanced Learners of English (LexTALE), a non-speeded visual lexical decision test (Lemhöfer & Broersma, 2012), to ensure our German participants were indeed highly-proficient in Dutch.

In this test, participants were presented with 40 Dutch words and 20 nonwords. The nonwords are formed by either changing letters in an existing words, or a recombination of existing morphemes, and the task for participants was to note down whether a word was an existing word in Dutch or not. Participants had to score above 60% in on the LexTALE test to participate in the experiment. A score of 60% and higher is predicted to correlate with an upper intermediate level (B2 level or higher). Native listeners were asked to also fill out the LexTALE test as a control. After the main experiment, we administered an adapted version of the LexTALE test (LexTALE 2) which contained 40 verbs that were used in the experiment to ensure that participants were familiar with them, and 20 non-words that were created in a similar manner as the non-words in the first LexTALE test.

8.3.2. Stimulus materials

The stimuli that were used in this experiment were the same stimuli as described in Drijvers, Ozyurek & Jensen (2018). Participants were presented with 160 video clips in which a non-professional actress uttered a highly-frequent Dutch action verbs in clear or degraded speech, while she performed an iconic gesture or not. The actress in the video was always visible from her knees up, in front of a neutrally-colored background, and wearing neutrally-colored clothing. Previous work confirmed that both native and non-native listeners were familiar with the verbs that were used (see e.g., Drijvers & Ozyurek, 2017; 2018; 2019), and in every video a different verb was used.

All videos were 2 seconds long. In the videos that contained a gesture, gesture preparation (i.e., the first frame in which the actress moved her hand up) started 120 ms after video onset (see for video structure Figure 28). On average, the stroke of the gesture started at 550 ms, followed by gesture retraction at 1380 ms, and gesture offset at 1780 ms. Speech onset on average started at 680 ms, which maximized the overlap between the stroke and gesture segment so that the iconic gesture and speech could benefit mutual comprehension (see Habets, Kita, Shao, Ozyurek, & Hagoort, 2011).

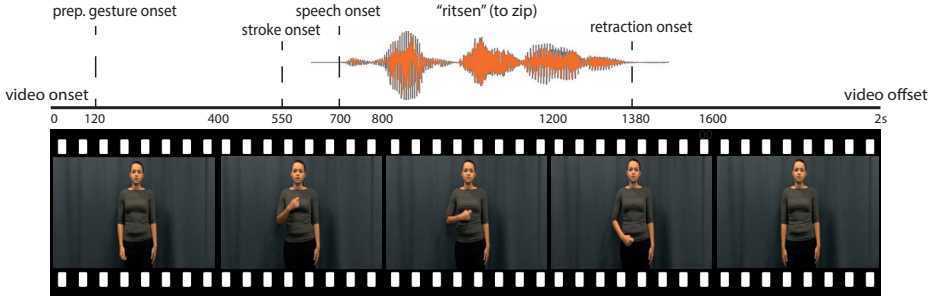


Figure 28 Schematic overview of video file. Videos were 2000 ms long. Preparation of a gesture started at 120 ms., stroke at 550 ms., speech at 680 ms., retraction onset was at 1380 ms., gesture offset was at 1780 ms.

All the gestures that were used were iconic, but were potentially ambiguous without speech, as they are in the case of naturally occurring co-speech gestures (Krauss et al., 1991). In a pretest that was conducted as part of Drijvers & Ozyurek (2017), we presented participants with the gesture videos without any sound. We asked participants to write down which verb they thought was depicted in the video, and then presented them with the verb we associated with the gesture in the video and asked them to indicate on a 7-point scale how iconic they found that verb of the gesture in the video. Iconic gestures that did not reach 5 points on this scale were discarded. Overall, our gestures had a mean recognition rate of 59%, which indicates that they are potentially ambiguous in the absence of speech (e.g., a 'rowing'-gesture, that could fit with the verbs 'sweeping' or 'rowing', and thus needs speech to be disambiguated).

The sound files that were played in the videos were extracted, intensity-scaled to 70 dB, denoised in *Praat* (Boersma & Weenink, 2015), and then recombined with their corresponding video files. All sound files were cleaned, and from these clean versions, a 6-band noise-vocoded version was created using a custom script in *Praat*. Noise-vocoding degrades the spectral content of the audio file (Shannon et al., 1995) while the temporal envelope of the sound preservers. This causes the sound to still be intelligible to some extent, with the more bands being present in the signal, the more intelligible the signal becomes (e.g., 6-band noise-vocoding is more intelligible than 2-band noise-vocoding). For both native and non-native listeners, 6-band noise-vocoding is the noise-vocoding level where listeners

benefit most from the semantic information that is conveyed by the gesture. Therefore, we used 6-band noise-vocoding as the degradation level for the current experiment.

We included four conditions in our experiment, divided over 2 within-subject factors (Noise-vocoding (degraded/clear) and Gesture (present/absent): clear speech only (CO), degraded speech only (DO), clear speech + gesture (CGCG), degraded speech + gesture (DG). The differences between these conditions were assessed between the two listener groups; native and non-native listeners. We were particularly interested in the 'gestural enhancement effect', which can be calculated by taking the interaction between Noise-vocoding (present/absent) and Gesture (present/absent). All conditions contained 40 unique verbs and videos. Every participant thus saw 160 different videos in total.

8.3.3. Procedure

Non-native listeners were asked to fill out the LexTALE assessment online prior to coming to the lab, to ensure their proficiency level was high enough to participate. If their score was above 60%, they learned Dutch after or at age 12 and they used Dutch on a regular basis, they were invited to participate in the study. Native listeners filled in the LexTALE test on paper upon arrival.

The participants were then instructed on the task and set up with the eye tracker in a dimly-lit soundproof booth. They were asked to watch and listen to videos and fill out what verb they heard in the videos in a subsequent 4-alternative forced choice identification task. All participants were seated approximately 70 cm from the computer screen and held a four-button box to submit their answers. The experimental stimuli were presented on a 1650 x 1080 monitor using Experiment Builder (SR Research), while eye-movements were monitored at a sampling rate of 1 kHz with an SR Research EyeLink 1000 eye-tracker.

After being instructed on the task, participants underwent a 9-point calibration and validation procedure. This procedure was repeated until the average discrepancy between the calibration point and the participants gaze was $<0.75^{\circ}$. Each trial started with a fixation cross (1000 ms), which was followed by

an experimental video (2000 ms), a blank screen (1000 ms) and finally, the four answering options of the 4-alternative forced choice identification task where the participant had to indicate what verb they heard (5000 ms) (see Figure 29). The order of the stimuli was pseudo-randomized for all participants. In total, the experiment lasted 20 minutes. After the main experiment, the participant had to fill out the second LexTALE task.

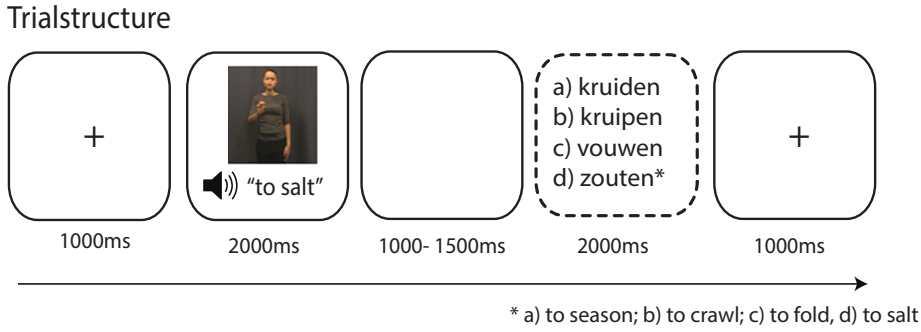


Figure 29 Trial overview. Participants encountered a fixation cross (1000 ms), listened and watched to the video (2000 ms), followed by a short delay (1000 ms) and the 4-alternative forced choice identification task (5000 ms max., screen disappeared after answering).

8.3.4. Eye-tracking analysis

We segmented the data in epochs of 2 seconds, corresponding to the length of the video. All trials that had a trackloss of 25% and higher were removed. All data points were automatically coded as fixations, saccades or blinks using the Eyelink algorithm, and were loaded and processed in MATLAB, partly using the FieldTrip Toolbox (Oostenveld et al., 2011). The timing of fixations was always relative to the onset of the video.

We defined three areas of interest: face, mouth & body. Initially we only included 'face' and 'body' in our analyses, but we specified a subregion in the 'face' area of interest to investigate whether specific effects were attributable to the mouth region (see Figure 30). The area of interest that comprised the 'body' was made on the basis of the x-y coordinates that corresponded to the furthest points in which a gesture was seen to occur in a video, and thus comprised the whole gesture space. Fixations that fell outside of these areas of interest were not analyzed

further. We calculated the proportions of fixations at each area of interest, and this formed the dependent variable in the subsequent analyses.

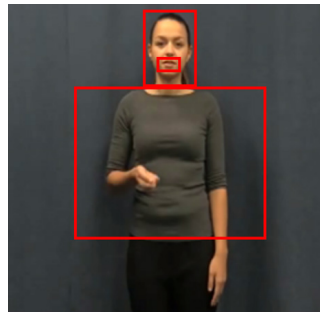


Figure 30 Areas of interest: mouth, face and gesture.

8.3.5. Cluster-based permutation tests

We used non-parametric cluster-based permutation tests to test for differences between conditions and groups and to control for multiple comparisons (Maris & Oostenveld, 2007). Our time-window of interest was the entire epoch, from video onset (0 ms) to video offset (2000 ms), and clustering was done along the temporal dimension. We computed the difference between two paired conditions or unpaired groups and created a distribution of these difference values. The observed values were thresholded with the 95th percentile of this distribution, and these clusters formed the cluster candidates. These values were then randomly reassigned over the conditions (1000 permutations) to form the permutation distribution. Every permutation the cluster candidate with the highest sum of the difference was added to the permutation distribution. The observed values were then compared to the permutation distribution. Clusters that fell in the highest or lowest 2.5 percentile of the distribution were considered significant. The calculation of interaction effects followed a similar procedure, but compared two differences to each other, or the difference of these differences per group.

8.3.6. Correlational analyses

One of our main interests is to investigate whether looking at a gesture when speech is degraded can predict comprehension during the 4-alternative forced choice identification task. We therefore extracted the mean fixation proportion of each participant in the time window where the meaningful part of the gesture is unfolding (from stroke onset, 550 ms, to retraction onset, 1380 ms) in the DG condition, and correlated this with the gestural enhancement effect in the behavioral task ((DG - CG) - (DO - CO)) per participant, using Spearman correlations.

8.4. Results

8.4.1. Behavioral results – LexTALE

We used the LexTALE test to assess the Dutch proficiency level of all participants. Native listeners scored significantly better on the first LexTALE test than non-native listeners ($t(38) = 5.587$, $p < .001$, Native listeners: MeanScore = 91.9, SD = 6.1, Non-native listeners: MeanScore = 78.1, SD = 9.2), as well as on the second adapted LexTALE test ($t(22.6) = 8.561$, $p < .001$, Native listeners: MeanScore = 96.3, SD = 2.7, Non-native listeners: MeanScore = 78.6, SD = 8.8).

8.4.2. Behavioral results – 4-alternative forced choice identification task (accuracy and reaction times)

We then tested for differences in accuracy and reaction times by conducting two mixed repeated-measures ANOVA's with ListenerGroup (native/non-native) as a between-subjects factor, and Noise-Vocoding (clear/degraded) and Gesture (present/not present) as within-subjects factors. All results and individual data points are displayed in Figure 31 (accuracy) and Figure 32 (reaction times), and raincloud plots were created by using code by Allen, Poggiali, Whitaker, Marshall, & Kievit (2018).

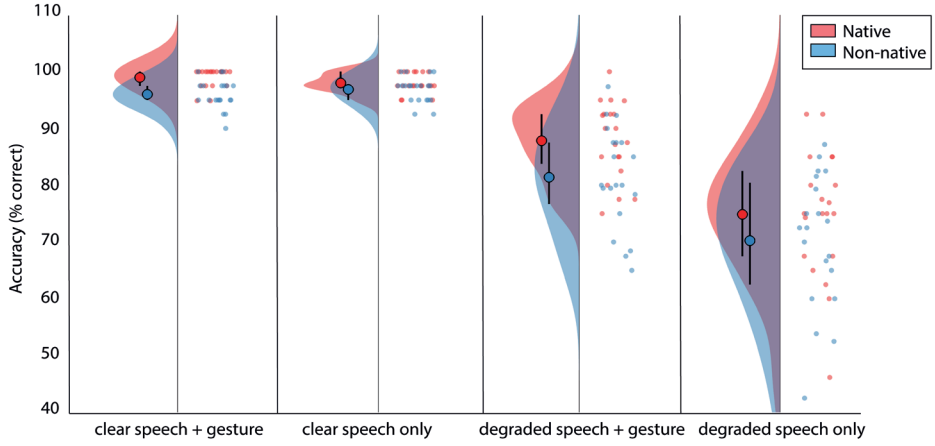


Figure 31 Raincloud plots of accuracy scores in the 4-alternative forced choice identification task per condition. Native listeners are displayed in red, non-native listeners in blue. Per condition (x-axis), two distributions are displayed. Left: Width of the distributions represent the density in order of 10-3. Large dot represents mean, lines represent quantiles of boxplot. Right: Individual dots on the right halves of each plot per condition represent individual data points.

As can be observed from Figure 31, both native and non-native listeners answered more accurately when speech was clear than speech was degraded ($F(1,38) = 193.90$, $p < .001$, partial $\eta^2 = .836$) and when a gesture was present as opposed to not present ($F(1,38) = 63.91$, $p < .001$, partial $\eta^2 = .627$). These effects did not differ per listener group, as was demonstrated by a lack of an interaction effect between ListenerGroup and Noise-Vocoding ($F(1,38) = 1.65$, $p = .20$, partial $\eta^2 = .042$) and a lack of an interaction effect between ListenerGroup and Gesture ($F(1,38) = 1.47$, $p = .233$, partial $\eta^2 = .037$). We observed an interaction between Noise-Vocoding and Gesture ($F(1,38) = 70.815$, $p < .001$, partial $\eta^2 = .651$), indicating that both groups experienced a larger gain in accuracy caused by gesture in degraded than in clear speech. Contrary to our hypotheses, this gestural enhancement (Native(DGDO vs CGCO) vs Non-native(DGDO vs CGCO)) effect did not differ between the native and non-native listener groups ($F(1,38) = < .001$, $p < .987$, partial $\eta^2 = < .001$), indicating that the gestural enhancement effect was not larger for natives than non-natives. However, native listeners were in general more accurate than non-native listeners ($F(1,38) = 6.42$, $p = .016$, partial $\eta^2 =$

.145), even though the magnitude of the gestural enhancement effect was similar for native and non-native listeners.

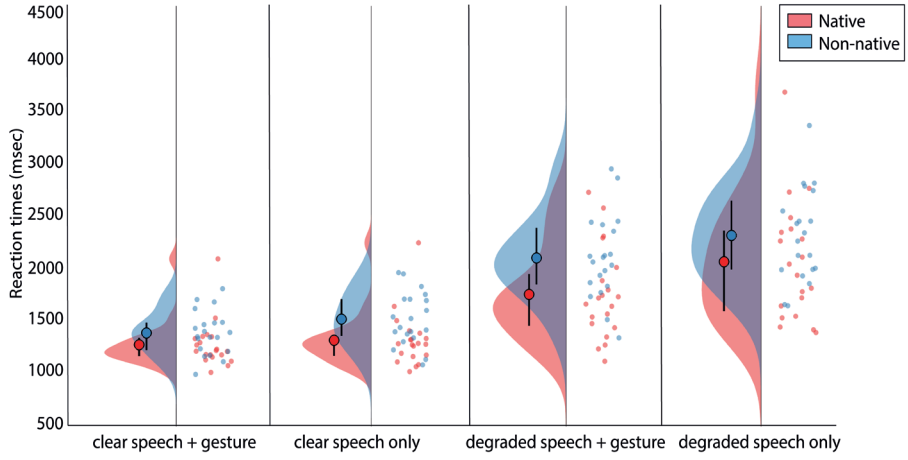


Figure 32 Raincloud plots of reaction times in the 4-alternative forced choice identification task per condition. Native listeners are displayed in red, non-native listeners in blue. Per condition (x-axis), two distributions are displayed. Left: Width of the distributions represent the density in order of 10-3. Large dot represents mean, lines represent quantiles of boxplot. Right: Individual dots on the right halves of each plot per condition represent individual data points.

In Figure 32, the reaction times on the 4-alternative forced choice identification task are displayed for native and non-native listeners. Both native and non-native listeners were quicker in answering on the 4-alternative forced choice identification task when speech in a video was clear than when it was degraded ($F(1,38)$ = 213.83, $p < .001$, partial $\eta^2 = .849$), and when a gesture was present as compared to not present ($F(1,38) = 62.723$, $p < .001$, partial $\eta^2 = .623$). Both of these effects did not differ per listener group, as demonstrated by the lack of an interaction between ListenerGroup and Noise-Vocoding ($F(1,38) = 2.323$, $p = .136$, partial $\eta^2 = .058$) and the lack of an interaction between ListenerGroup and Gesture ($F(1,38) = 0.006$, $p = .94$, partial $\eta^2 = < .001$). Native and non-native listeners both experienced the largest speed up in reaction times by gestures in degraded speech compared to clear speech, as indicated by an interaction effect between Noise-Vocoding and Gesture ($F(1,38) = 24.731$, $p < .001$, partial $\eta^2 = .394$). In contrast to the results in accuracy, and in line with our predictions, we observed a three-

way interaction between Noise-Vocoding, Gesture and ListenerGroup ($F(1,38) = 6.965$, $p < .012$, partial $\eta^2 = .155$), indicating that the speed up in reaction times caused by the gestural enhancement effect was larger for native listeners than non-native listeners. Finally, overall, native listeners were quicker to answer than non-native listeners ($F(1,38) = 4.798$, $p < .035$, partial $\eta^2 = .112$)

8.4.3. Eye-tracking results - face and mouth

8.4.3.1. Native listeners (within-group)

We first analyzed the eye-tracking results of both groups separately to uncover specific gaze allocation patterns per listener group, per area of interest, before comparing both groups to test for between-group differences. In line with our predictions, native listeners significantly looked more at the face when speech was degraded than when speech was clear (DO>CO, $p <$. 001, 1019-2000 ms) (see Figure 33A). When a gesture was present, native speakers also looked more at the face when speech was degraded than when speech was clear (DG>CG, $p <$ 0.001, 1216 – 1936 ms). Native listeners fixated most on the face when speech was degraded and no gesture was present (DGCG< DOCO, $p < .002$, 1600 – 2000 ms).

To specify whether gaze was allocated to visible speech or other parts of the face, we then created an extra area of interest that covered the area of the mouth. Here, we again observed that native listeners significantly gazed more at the mouth when speech was degraded than when speech was clear (DO>CO, $p <.001$, 1199 – 1990 ms), and when a gesture was present as compared to not present (DG>CG, $p < .001$, 320 – 1826 ms) (see Figure 34A). Native listeners fixated most on the mouth when speech was degraded and a gesture was present (DGCG > DOCO, $p = .02$, 649 – 760 ms.)

8.4.3.2. Non-native listeners (within-group)

We then analyzed gaze allocation to the face and mouth within the non-native listener group. In line with our hypotheses, non-native listeners looked more at the face when speech was degraded as compared to clear (DO>CO, $p < .001$, 1243 – 2000 ms), and when a gesture was presented in degraded than in clear speech

(DG<CG, p < .008, 1500 – 1883 ms) (see Figure 33B). When speech was degraded and no gesture was present, non-native listeners tended to gaze at the face more often (DGCG < DOCO, p < .001, 1464 – 2001). Single comparisons between gesture and no-gesture conditions in both speech conditions can be found in Supplementary Materials S5.

We zoomed into these effects by distilling how often non-native listeners fixated on the mouth of the actress (see Figure 34B). Non-native listeners looked at the mouth more when speech was degraded than when speech was clear (DO>CO, p < .001, 1351 – 2000 ms). However, there was no difference observed between the degraded and clear conditions that contained a gesture (DG = CG, p = .39). When speech was degraded and no gesture was present, non-native listeners looked at the mouth longer than when a gesture was present or speech was clear (DGCG < DOCO, p = .001, 1611 – 2000 ms.)

8.4.3.3. Native vs. non-native listeners (between-group)

No differences between native and non-native listeners were observed when comparing the differences between the proportions of fixations to the face in the DG-CG (p = .13) condition. However, the difference between fixations to the face in degraded speech vs. clear speech conditions was larger for native than non-native listeners (Native(DOCO) > Non-native(DOCO), p = .03, 264 – 425 ms).

Similarly, we did not observe differences between native and non-native listeners when comparing the differences in fixations to the mouth in the DG-CG condition (p = .16), or DO-CO condition (p = .003). However, we observed a larger difference in CG-CO for non-native than native listeners (p = .045, 858 – 992 ms.) indicating that non-native listeners might focus more on other visual articulators than the mouth when speech is clear and a gesture is present than native listeners. We followed up on this by running a post-hoc test and comparing looks to gesture (described below) between the two groups in the CG condition, and confirmed that non-native listeners look more at gestures in clear speech than native listeners (p = .03, 856 - 933 ms).

8.4.4. Eye-tracking results - gesture

8.4.4.1. Native listeners (within-group)

Our second visual articulator of interest was gesture. Native listeners looked more at the torso when speech was clear than when speech was degraded (DO<CO, p < .001, 769 – 2001 ms) and when a gesture was present in clear speech than when a gesture was present in degraded speech (DG<CG, p = .002, 1353 – 1884 ms) (see Figure 33C). The difference in fixation proportions to conditions that contained a gesture was smaller than when no gesture was present (DGCG < DOCO, p < .001, 1648 – 2000 ms). This possibly indicates that when native listeners process degraded speech in a gesture and no gesture context, visual attention is probably allocated to visual articulators for a longer time than when speech is clear. Single comparisons between gesture and no-gesture conditions in both speech conditions can be found in supplementary materials.

8.4.4.2. Non-native listeners (within-group)

Similarly to native listeners, we then investigated gaze allocation to gestures within the non-native listener group. Non-native listeners initially look more at the body when speech is clear than when speech is degraded (DO<CO, p < .001, 259 – 513 ms), but then look more at the body when speech is degraded as compared to clear (DO>CO, p < .001, 1278 – 2000 ms) (see Figure 33B). When a gesture is present, non-native listeners look more at the body than when speech is clear than when speech is degraded (DG<CG, p = .013, 1521 – 1759 ms). The difference between fixation proportions to conditions that contained a gesture was smaller than the difference between fixation proportions to the body in conditions that did not have a gesture (DGCG < DOCO, p < .001, 1641 – 2000 ms), indicating that similar to native listeners, when non-native listeners process degraded speech in a gesture and no gesture context, visual attention is probably allocated to informational resources for a longer time than when speech is clear. Single comparisons between gesture and no-gesture conditions in both speech conditions can be found in supplementary materials.

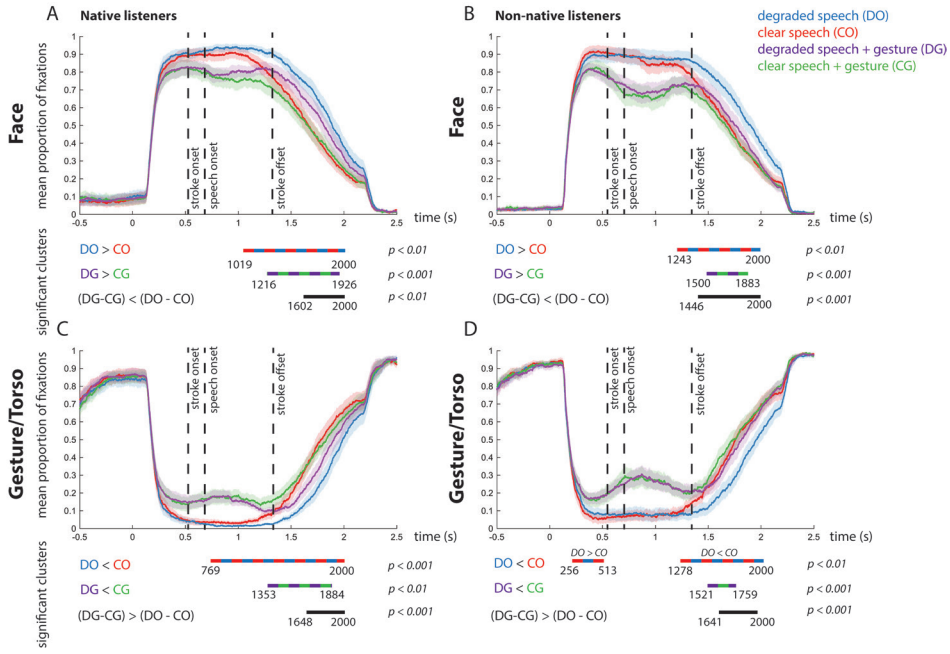


Figure 33 Mean proportion of fixation over time on the face for native listeners (33A) and non-native listeners (33B). Mean proportion of fixations over time on the body ('Gesture/Torso') for native (33C) and non-native listeners (33D). In all graphs, each color represents a condition (DO = blue, CO = red, DG = purple, CG = green), and shaded color bars around the mean proportion lines represent standard error. Below each graph, the difference between the conditions and the direction of the effects is specified per comparison. The colored dashed lines represent the differences between the conditions in the significant time-intervals.

8.4.4.3. Native vs. non-native listeners (between-group)

No differences were observed when comparing gaze to the body when comparing the differences between DO-CO (p = .41) and DG-CG (p = .56) between the native and non-native group. We observed a difference between the groups when comparing the difference DG-DO, which was larger for non-native listeners than native listeners (Non-nativeDGDO > NativeDGDO, p = 0.029, 743 - 1087 ms.) and for CG-CO, which was larger for non-native listeners than native listeners (Non-nativeCGCO > NativeCGCO, p = .008, 731 - 952 ms), which might indicate that non-native listeners focus more on gestures during both clear and degraded speech as compared to native listeners. In addition to the post-hoc comparison of

fixations to CG described in section 8.4.3.3. on the between-group comparison for gaze allocation to the face and mouth, we ran another post-hoc test to confirm a similar pattern in degraded speech: non-native listeners indeed significantly look more to gestures in degraded speech than native listeners (p = .01, 981 - 1151 ms).

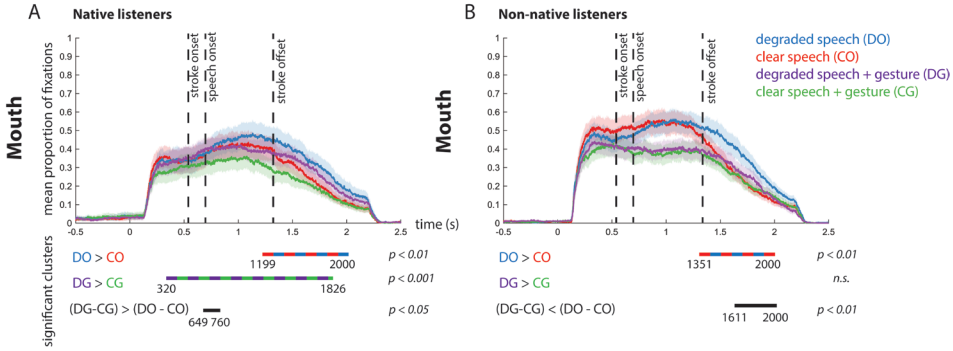


Figure 34 Mean proportion of fixation over time on the mouth for native listeners (34A) and non-native listeners (34B). In all graphs, each color represents a condition (DO = blue, CO = red, DG = purple, CG = green), and shaded color bars around the mean proportion lines represent standard error. Below each graph, the difference between the conditions and the direction of the effects is specified per comparison. The colored dashed lines represent the differences between the conditions in the significant time-intervals. 'n.s.' denotes 'not significant'.

8.4.4.4. Correlational analyses

Native listeners who look more at a gesture when speech is degraded, experience a larger gestural benefit during degraded speech comprehension on accuracy (r = .521, p = 0.019), but this was not reflected in a speeding up of their reaction times (r = -.136, p = 0.567). Non-native listeners who look more at a gesture when speech is degraded, do not experience a larger benefit during degraded speech comprehension on accuracy (r = .056, p = 0.813) nor on reaction times (r = -.065, p = 0.787).

8.5. Discussion

We investigated whether and how native and non-native listeners allocate overt visual attention to visible speech and gestures during clear and degraded speech comprehension, and whether gaze allocation to these visual articulators

could predict the gestural benefit listeners experience during degraded speech comprehension. On the 4-alternative forced choice identification task, both native and non-native listeners demonstrated a clear gestural enhancement effect during degraded speech comprehension. The behavioral data revealed a larger speeding up in reaction times during gestural enhancement of degraded speech comprehension for natives than non-natives, but this was not reflected in their accuracy scores. Our eye-tracking results revealed that overall, non-native listeners seem to gaze more at gestural information than native listeners. However, gaze allocation to gesture only predicted how much a listener benefits from gestural information during degraded speech comprehension in native, but not non-native listeners. Below we interpret these results in more detail.

8.5.1. Gestural enhancement of degraded speech comprehension is larger for native than non-native listeners

In line with our hypotheses, both native and non-native listeners demonstrated a clear gestural enhancement effect on the 4-alternative forced choice identification task (following Drijvers & Ozyurek, 2017; Drijvers, Ozyurek & Jensen, 2018). This gestural enhancement effect was largest when speech was degraded, and was similar for native and non-native listeners on the accuracy, but not the reaction time measurements. This means that gestural semantic information causes a speeding up in reaction times when speech is degraded for native listeners, but less for non-native listeners.

8.5.2. Native and non-native listeners look more to the face and mouth when speech is degraded than when speech is clear

In general, and in line with previous literature, the face formed the locus of attention during comprehension for both native and non-native listeners (Baron-Cohen et al., 2001; Rogers et al., 2018). As predicted, native and non-native listeners significantly looked more often to the face and mouth when speech was degraded than when speech was clear, irrespective of whether a gesture was present or not. This effect started after speech onset and lasted until the end of the video. As can be observed from Figure 33A/B, the difference between the

degraded and clear speech conditions is mostly evident at the end of the video, where there is a more rapid decline in proportions of fixations to the face in clear speech than in degraded speech for both native and non-native listeners. This suggests that when speech is degraded, listeners demonstrate sustained visual attention to the face, possibly to extract more information to aid comprehension when speech is degraded (Ross et al., 2007; Vatikiotis-Bateson et al., 1998), and the lack of phonological information might delay or hinder comprehension. This is also demonstrated by the fact that both for native and non-native listeners this effect occurs late in the videos, and is largest in the no-gesture conditions, where there was no additional information that could aid comprehension.

During our analyses we then made the post-hoc decision to add the mouth area as a more specific area of interest instead of the whole face, to disentangle whether these effects were specific to visible speech, or the face as a whole, as comprehension can also be aided by other facial articulators (e.g. prosody, which can be distributed over the face (Yehia et al., 2002). For the conditions without a gesture (DO vs. CO), we again observed that both native and non-native listeners gazed more at the mouth in a relatively late time-window, which again might reflect sustained visual attention to the mouth to extract phonological information to aid comprehension. When a gesture was present (DG vs. CG), we observed a difference between degraded and clear conditions in native listeners, but not non-native listeners (see Figure 33A/B). For native listeners, this effect started early in the video (320 ms). It is unclear why these conditions already differed before speech onset. However, as there is no difference observed for DG vs. CG for non-native speakers over the whole time window, we postulate that this effect reflects that native listeners are less likely to look at the mouth when speech is clear to aid comprehension than non-native listeners. Non-native listeners might aim to extract phonological information for comprehension, possibly because their knowledge of the phonology of the L2 is not as strong as in native listeners (Mayo et al., 1997; Cutler et al., 2004; Bradlow & Alexander, 2007).

Finally, when comparing the proportion of gaze allocation to the mouth in conditions that contain a gesture and conditions that do not contain a gesture, we observed a larger difference between the conditions that contain a gesture (DG

vs. CG) than conditions that do not contain a gesture (DO vs. CO). For native listeners, this effect occurred in an early time window (649 - 760 ms) around speech onset. The trajectories of proportions of fixations to the mouth in the DG and CG conditions continue to differ for native listeners throughout the video, but not for non-native listeners. This effect thus suggests that when speech starts, native listeners might try to immediately incorporate the phonological information that is conveyed by the mouth, especially when a gesture is present and speech is degraded. This would allow native listeners to optimally benefit from both visual articulators in a joint context (Drijvers & Ozyurek, 2017).

A similar approach might however be too taxing for non-native listeners. As can be observed from Figure 34B, the DG and CG conditions continue to attract a similar proportion of fixations, whereas the DO and CO conditions differ, with the DO conditions attracting prolonged attention to the mouth. This indicates that non-native listeners need a similar amount of phonological cues that are conveyed by visible speech in clear and degraded speech when a gesture is present, and it might be more difficult to benefit from both visual articulators in a joint context. Instead, non-native listeners might be attracted to gestural information more strongly to aid comprehension.

8.5.3. Native and non-native listeners demonstrate sustained visual attention to visible speech, but not gestures when phonological cues are hard to disambiguate.

We predicted that listeners would gaze more at gestures when speech was degraded than when speech was clear. When no gestures were present (DO vs. CO), native and non-native listeners looked more at the body when speech was clear than when speech is degraded. For native listeners, this difference commenced immediately after speech onset (769 ms), indicating that when speech was degraded, native listeners directed less gaze to the body. Instead, native listeners might look more at the mouth when phonological cues are hard to disambiguate. For non-native listeners however, this effect only occurred slightly before speech offset (1278 ms), which might indicate that when comprehension of the degraded signal was challenging, sustained attention was allocated to the face/mouth to resolve these

cues. This might delay the uptake of gestural information for non-native listeners. This again confirms that understanding and processing degraded speech is more detrimental for non-native listeners as compared to native listeners (Bradlow & Alexander, 2007; Brouwer et al., 2012; Kilman et al., 2014; Mayo et al., 1997; Scharenborg et al., 2018).

In contrast to our predictions, both native and non-native listeners demonstrated a larger portion of fixations to gestures in clear speech, which commenced close to stroke and speech offset (native listeners: 1353 ms, non-native listeners: 1521 ms). We do not believe this is a direct effect that is caused by gesture, as the meaningful part of the gesture has (almost) unfolded. Instead, we believe the difference between these conditions emerges because in the DG condition, participants fixate more strongly on the head and mouth than in the clear conditions. Possibly, the semantic information of the gesture is already extracted by the listeners at that timepoint (1351 ms and later, see Figure 33C/D), and the degraded speech signal cause a delayed use of the semantic information as compared to clear speech, as there is sustained visual attention to other visual articulators to extract cues for disambiguation.

8.5.4. Non-native listeners look more at gestures during clear and degraded speech than native listeners

In both degraded speech and in clear speech, we observed a larger proportion of fixations to gesture for non-native compared to native listeners in the time interval that the meaningful part of the gesture was unfolding (in line with Drijvers & Ozyurek, 2018). For clear speech, this effect occurred earlier than for degraded speech (clear speech: 856 - 933 ms; degraded speech: 981 - 1151 ms). This again confirms our hypothesis that the degraded speech signal might delay the use of semantic information. Non-native listeners might find it more difficult to resolve the phonological cues in the speech signal and couple them to the phonological information conveyed by visible speech than native listeners, and may therefore try to focus more on the semantic information that is conveyed by the gesture than native listeners. Native listeners might be able to benefit from both visual articulators in a joint context quicker than non-native listeners due to their native

listener status. These results concur with earlier work where we observed a larger N400 effect during speech-gesture integration in clear speech for non-native than native listeners (Drijvers & Ozyurek, 2018), as well as our previous MEG results, that suggested that coupling the semantic information of the gesture to the degraded speech signal might be hindered for non-native listeners (Drijvers, Van der Plas, Ozyurek & Jensen, in revision).

8.5.5. Gaze allocation to gestures predicts gestural enhancement during degraded speech for native but not non-native listeners

We hypothesized that in both groups, gaze allocation to gestures could predict the gestural enhancement a listener experiences during degraded speech comprehension. Whereas this was not the case for non-native listeners, we demonstrated that a native listener's gaze to gestures during degraded speech could predict the increase in accuracy a native listener experiences during gestural enhancement of degraded speech comprehension. This finding again suggests that native listeners can jointly benefit from both visual articulators during comprehension, but non-natives might be more hindered in this process. Although non-native listeners allocate more gaze to gestures, they cannot resolve enough cues in the degraded speech signal so that the gesture can predict comprehension. Possibly, non-native listeners need more phonological cues, partly conveyed by visible speech, to aid comprehension.

8.6. Conclusion

We demonstrated that both native and non-native listeners look more at the face and mouth when speech is degraded than when speech is clear. Native and non-native listeners both looked more at gestures when speech was clear than when speech was degraded. This is possibly due to the fact that native and non-native listeners both demonstrated sustained visual attention to the face and mouth when phonological cues in speech were hard to disambiguate. Non-native listeners allocated more gaze to gestures than native listeners, but as disambiguating the degraded auditory cues was more challenging, the use of semantic information might be more delayed and hindered than for native listeners. As native listeners

are more able to utilize and resolve degraded auditory cues than non-natives, native listeners can map more visual information to the speech signal, resulting in better speech comprehension, especially when speech is degraded.

8.7. Acknowledgements

This research was supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from Netherlands Organization for Scientific Research. We are very grateful to Nick Wood, for helping us in editing the video stimuli, and to Gina Ginos, for being the actress in the videos.

Chapter 9

Speech-gesture integration studied by rapid frequency-tagging

9.1. Abstract

During communication in real-life settings, the brain integrates information from auditory and visual modalities to form a unified percept of our environment. In the current MEG study, we used rapid invisible frequency tagging (RIFT) to generate steady-state evoked fields and investigate the integration of and interactions between these different modalities in a multimodal, semantic context. Until now, technical limitations have prevented the use of higher frequencies in frequency tagging studies. Here, we used a new projector with a 1440 Hz refresh rate to present participants with videos of an actress uttering action verbs (tagged at 61Hz) accompanied by a gesture (tagged at 68Hz). Integration ease was manipulated by auditory factors (clear/degraded speech) and visual factors (congruent/incongruent gesture). We reliably identified an enhanced intermodulation frequency of the auditory and visually tagged signals at 7 Hz (f2-f1) when integration was easiest (i.e., when speech was clear and accompanied by a congruent gesture). This signature of non-linear audiovisual integration was strongest in left inferior frontal gyrus and left-temporal regions, areas known to be involved in speech-gesture integration. We suggest that this enhanced power at the intermodulation frequency reflects the ease of integration and that speech-gesture information interacts in higher-order areas. Furthermore, we provide a proof-of-principle of the use of RIFT to study the integration of and interactions between audiovisual stimuli in a semantic context.

This work is based on Drijvers, L., Spaak, E., Herring, J., Ozyurek, A., Jensen, O. (in preparation). Speech-gesture integration studied by rapid frequency tagging.

9.2. Introduction

During communication in real-life settings, our brain needs to integrate auditory input with visual input to form a unified percept of the environment. Several magneto- and electroencephalography (M/EEG) studies have demonstrated that integration of non-semantic audiovisual inputs can occur as early as 50-100 ms after stimulus onset (e.g., Giard & Peronnet, 1997; Molholm et al., 2002; Talsma, Senkowski, Soto-faraco, & Woldorff, 2010), and encompasses a widespread network of primary sensory and higher-order regions (e.g., Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Calvert, 2001; Werner & Noppeney, 2010).

The integration of these audiovisual inputs has been studied in more detail by using frequency-tagging (Giani et al., 2012; Regan et al., 1995b). Here, an auditory or visual stimulus can be periodically modulated at a specific frequency, for example by modulating the luminance of a visual stimulus or the amplitude of an auditory stimulus. This produces steady-state evoked potentials (SSEPs, SSEFs for MEG) with strong power at the tagged frequency (Norcia et al., 2015; Picton et al., 2003; Vialatte et al., 2010). Using this technique is especially interesting in the context of studying audiovisual integration, because it enables the tagging of an auditory stimulus and a visual stimulus at two different frequencies (f_1 and f_2) in order to study whether and how these two inputs interact in the brain. Previous work has suggested that when the auditory and visual signals interact, this could result in power at the intermodulation frequencies of the two stimuli (e.g., f_2-f_1 or f_2+f_1) (Regan & Regan, 1989). Such intermodulation frequencies are thought to only arise from a non-linear interaction of the two oscillatory signals; in case of audio-visual integration the intermodulation is likely to reflect neuronal activity that non-linearly combines the signals of the two inputs (Regan & Regan, 1988; Zemon & Ratliff, 1984).

However, previous work has reported inconclusive results on the occurrence of such intermodulation frequencies as a signature of non-linear audiovisual integration. Furthermore, this integration has so far only been studied in non-semantic contexts (e.g., the integration of tones and gratings). For example, whereas Regan et al., (1995) identified intermodulation frequencies in an area close

to the auditory cortex, Giani et al., (2012) identified intermodulation frequencies within (i.e., as a result of tagging two frequencies in the visual domain), but not between modalities (i.e., as a result of tagging two frequencies in the auditory and visual domain). In both of these previous studies, frequency-tagging was applied at lower frequencies (< 30Hz for visual stimuli, <40Hz for auditory stimuli) to identify these intermodulation frequencies (Giani et al., 2012; Regan et al., 1995b). This might be problematic when considering that spontaneous neuronal oscillations at lower frequencies (e.g., alpha and beta oscillations) are also likely entrained by frequency tagging (Keitel, Quigley, & Ruhnau, 2014; Spaak, de Lange, & Jensen, 2014). In the current study, we use novel projector technology to perform frequency-tagging at high frequencies (rapid invisible frequency tagging; RIFT), and in a semantic context. Previous work has demonstrated that neuronal responses to a rapidly flickering LED can be driven and measured up to 100 Hz (Herrmann, 2001), and can successfully be used to study sensory processing in the brain (Herring, 2017). We here leverage these rapid neural responses in order to circumvent the issue of endogenous rhythms interacting with low-frequency tagging signals.

Previous behavioral and neuroimaging studies have demonstrated that listeners process and integrate speech and gestures at a semantic level, and that this integration relies on a network involving left inferior frontal gyrus (LIFG), left-temporal regions (STS/MTG), motor and visual cortex (Dick, Mok, Raja Beharelle, Goldin-Meadow, & Small, 2014; Drijvers, Ozyürek, & Jensen, 2018a, 2018b; Holle, Gunter, Ruschemeyer, Hennenlotter, & Iacoboni, 2008; Holle, Obleser, Rueschemeyer, & Gunter, 2010; Kircher et al., 2009; Straube, Green, Weis, & Kircher, 2012; Willems, Özyürek, & Hagoort, 2007, 2009; Zhao, Riggs, Schindler, & Holle, 2018, Drijvers, Ozyurek & Jensen, 2018a, 2018b, Josse et al., 2012; see for an overview: Ozyurek, 2014). Using frequency-tagging in such a context to study whether intermodulation frequencies can be identified as a signature of non-linear audiovisual integration would provide a proof-of-principle for the use of such a technique to study the integration of multiple inputs during complex dynamic settings, such as multimodal language comprehension.

In the present study, we set out to explore whether RIFT can be used to identify

intermodulation frequencies as a result of the interaction between a visually and auditory tagged signal in a semantic context. Participants watched videos of an actress uttering action verbs (tagged at f_1 = 61 Hz) accompanied by a gesture (tagged at f_2 = 68 Hz). Integration ease of these inputs was modulated by auditory factors (clear/degraded speech) and visual factors (congruent/incongruent gesture). For the visually tagged input, we expected that power would be strongest at 68Hz in occipital regions. For the auditory tagged input, we expected that power would be strongest at 61 Hz in auditory regions. We expected that interactions between the visually tagged and auditory tagged signal would be non-linear in nature, resulting in spectral peaks at the intermodulation frequencies of f_1 and f_2 (i.e., f_2+f_1 and f_2-f_1). On the basis of previous work, the locus of the intermodulation frequencies was expected to occur in LIFG and left-temporal regions such as pSTS/MTG, areas known to be involved in speech-gesture integration.

9.3. Methods

9.3.1. Participants

Twenty-nine right-handed native Dutch-speaking adults (age range = 19 - 40, mean age = 23,68, SD = 4.57, 18 female) took part in the experiment. All participants reported normal hearing, normal or corrected-to-normal vision, no neurophysiological disorders and no language disorders. All participants were recruited via the Max Planck Institute for Psycholinguistics participant database and the Radboud University participant database, and gave their informed consent preceding the experiment. Three participants (2 females) were excluded from the experiment due to unreported metal in dental work (1) or excessive motion artifacts (>75% of trials affected) (2). The final data set included the data of 26 participants.

9.3.2. Stimulus materials

Participants were presented with 240 video clips that contained an actress uttering a highly-frequent action verb accompanied by a matching or a mismatching iconic gesture (see for a detailed description of pre-tests on recognizability and iconicity

of the gestures, Drijvers & Ozyürek, 2017). All gestures used in the videos were rated as being potentially ambiguous when they would be viewed without speech, which allowed for mutual disambiguation of speech and gesture (Habets et al., 2011).

In all videos, the actress would be standing in front of a neutrally colored background, in neutrally colored clothes. We predefined the verbs that would form the 'mismatching gesture', in the sense that we asked the actress to utter the action verb, and depict the other verb in her gesture. This approach was chosen because we included the face and lips of the actress in the videos, and we did not want to recombine a mismatching audio track to a video to create the mismatch condition. All videos were on average 2000 ms long (SD = 21.3 ms). After 120 ms, the preparation (i.e., the first frame in which the hands of the actress moved) of the gesture started. On average, at 550 ms (SD = 74.4 ms), the meaningful part of the gesture (i.e., the stroke) started, followed by speech onset at 680 ms (SD = 112.54 ms). None of these timings differed between conditions. None of the iconic gestures were prescripted. All gestures were performed by the actress on the fly.

All audio files were intensity-scaled to 70 dB and denoised using *Praat* (Boersma & Weenink, 2015), before they were recombined with their corresponding video files using Adobe Premiere Pro. For 80 of the 160 sound files, we created noise-vocoded versions using *Praat*. Noise-vocoding pertains the temporal envelope of the audio signal, but degrades the spectral content (Shannon et al., 1995). We used 6-band noise-vocoding, as we demonstrated in previous work that this is the noise-vocoding level where the auditory signal is reliable enough for listeners to still be able to use the gestural information for comprehension (Drijvers & Ozyürek, 2017). To achieve this, we band-passed the sound files between 50 and 8000 Hz in 6 logarithmically spaced frequency bands with cut-off frequencies at 50, 116.5, 271.4, 632.5, 1473.6, 3433.5 and 8000 Hz. These frequencies were used to filter white noise and obtain six noise bands. We extracted the amplitude envelope of each band using half-wave rectification and multiplied the amplitude envelope with the noise bands. These bands were then recombined. Sound was presented to participants using MEG-compatible air tubes.

In summary, we manipulated integration strength in the videos by auditory

(clear/degraded) and visual (congruent/incongruent) factors (see Figure 35). This resulted in six conditions: clear speech + matching gesture (CM), clear speech + mismatching gesture (CMM), degraded speech + matching gesture (DM) and degraded speech + mismatching gesture (DMM), and two conditions not containing gestures (clear speech only and degraded speech only). All of the conditions contained 40 videos. Note that we did not include the two no-gesture conditions in any of the analyses in this study, as no movement was observable in the area that was visually frequency-tagged. All verbs and gestures were only presented once. Participants were asked to pay attention to the videos and identify what verb they heard in the videos in a 4-alternative forced choice identification task.

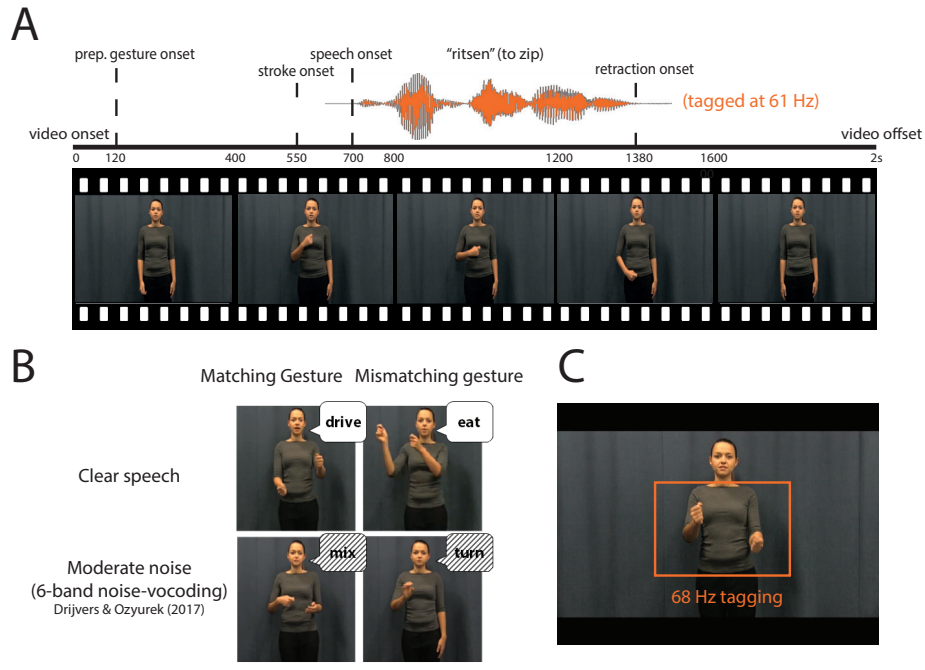


Figure 35 A. Illustration of the structure of the videos. Speech was amplitude-modulated at 61Hz. B. Illustration of the different conditions. C. Area that is used for visual frequency-tagging at 68Hz.

9.3.3. Procedure

Participants were tested in a dimly-lit magnetically shielded room and seated 70

cm from the projection screen. All stimuli were presented using MATLAB 2016b (Mathworks Inc, Natrick, USA) and the Psychophysics Toolbox, version 3.0.11 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). To achieve rapid invisible frequency tagging, we used a GeForce GTX960 2GB graphics card with a refresh rate of 120Hz, in combination with a PROPixx DLP LED projector (VPixx Technologies Inc., Saint-Bruno-de-Montarville, Canada), which can achieve a presentation rate up to 1440 Hz. This high presentation rate is achieved by the projector interpreting the four quadrants and three colour channels of the GPU screen buffer as individual smaller, grayscale frames, which it then projects in rapid succession, leading to an increase of a factor 12 (4 quadrants * 3 colour channels * 120 Hz = 1440 Hz) (User Manual for ProPixx, VPixx Technologies Inc., Saint-Bruno-de-Montarville, Canada).

9.3.3.1. Frequency-tagging

The area of the video that would be frequency-tagged was defined by the rectangle in which all gestures occurred. The pixels within that area were always tagged at 68 Hz. This was achieved by multiplying the luminance of the pixels within that square with a 68 Hz sinusoid (modulation depth = 100%; modulation signal equal to 0.5 at sine wave zero-crossing, in order to preserve the mean luminance of the video), phase-locked across trials (see Figure 36). For the auditory stimuli, frequency-tagging was achieved by multiplying the amplitude of the signal with a 61 Hz sinusoid, with a modulation depth of 100% (following Lamminmäki, Parkkonen, & Hari, 2014). In a pretest, we presented 11 native Dutch speakers with half of the stimuli containing the amplitude modulation, and half of the stimuli not containing the amplitude modulation in both clear and degraded speech. Participants were still able to correctly identify the amplitude modulated stimuli in clear speech (mean % correct without amplitude modulation: 99.54, with amplitude modulation: 99.31) and in degraded speech (mean % correct without amplitude modulation: 72.74, with amplitude modulation: 70.2) and did not significantly suffer more compared to when the signal was not amplitude modulated.

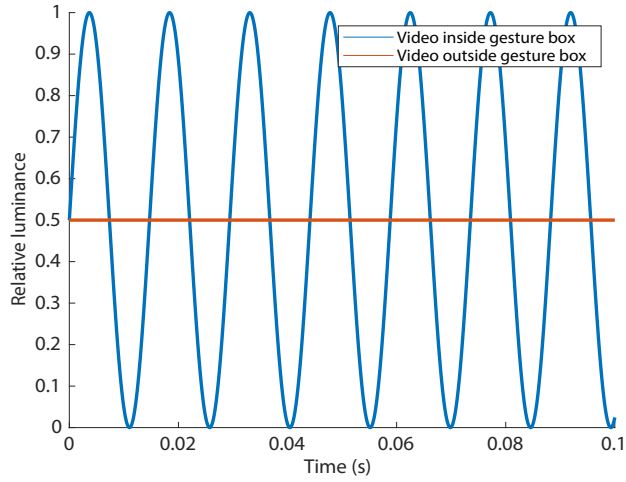


Figure 36 Illustration of luminance manipulation for visual-frequency tagging. 68 Hz frequency-tagging was achieved by multiplying the luminance of the pixels with a 68Hz sinusoid. Modulation signal was equal to 0.5 at sine wave zero-crossing to perserve the mean luminance of the video, phase-locked across trials.

Participants were asked to attentively watch and listen to the videos. Every trial started with a fixation cross (1000 ms), followed by the video (2000 ms), a short delay period (1500 ms), and a 4-alternative forced choice identification task (3000 ms). In the 4-alternative forced choice identification task, participants were presented with four options, and had to identify which verb they heard in the video by pressing one of 4 buttons on an MEG-compatible button box. This task ensured that participants were attentively watching the videos, and to check whether the verbs were behaviorally resolved. After a button press, a new trial would start after 1000 ms. Participants were instructed not to blink during video presentation.

Throughout the experiment, we presented all screens at a 1440 Hz presentation rate. Brain activity was measured using MEG, and was recorded throughout the experiment. The whole experiment lasted approximately 30 minutes, and participants were allowed to take a self-paced break after every block. All stimuli were presented in a randomized order per participant.

9.3.4. MEG data acquisition

MEG was recorded using a 275-channel axial gradiometer CTF MEG system (CTF MEG systems, Coquitlam, Canada). We used an online low-pass filter at 300 Hz and digitized the data at 1200 Hz. All participants' eye gaze was recorded by an SR Research Eyelink 1000 eye tracker for artifact rejection purposes. The head position of the participants was tracked in real time by recording markers on the nasion, and left and right periauricular points (Stolk et al., 2013). This allowed readjusting the head position of participants relative to their original starting position when the deviation was larger than 5 mm. After the experiment, T1-weighted structural magnetic resonance images (MRI) were collected from 24 out of 26 participants using a 3T MAGNETOM Skyra system.

9.3.5. MEG data analysis

9.3.5.1. Preprocessing

All MEG data were analyzed using the FieldTrip toolbox (Oostenveld et al., 2011) running in a Matlab environment. All data were segmented into trials starting 1s before and ending 3s after the onset of the video. The data were demeaned and line noise was attenuated using a discrete Fourier transform approach at 50, 100 and 150 Hz. All trials that contained jump artifacts or muscle artifacts were rejected using a semi-automatic routine. The data were then down-sampled to 400Hz. Independent component analysis (Bell & Sejnowski, 1995; Jung et al., 2000) was used to remove eye movements and cardiac-related activity. All data were then inspected on a trial-by-trial basis to remove artifacts that were not identified using these rejection procedures. These procedures resulted in rejection of 8.3% of the trials.

9.3.5.2. Frequency-tagging analyses - Sensor-level

To investigate the response in auditory and visual regions to the frequency-tagged signal, we first calculated time-locked averages of the event-related fields by

averaging gradiometer data over trials, over conditions, and over participants. All tagged stimuli were presented phase-locked over trials. We used an approximation of planar gradiometer data to facilitate interpretation of the MEG data, as planar gradient maxima are thought to be located above the neuronal sources that may underlie them (Bastiaansen & Knösche, 2000). This was achieved by converting the axial gradiometer data to orthogonal planar gradiometer pairs, which were combined by using root-mean-square (RMS) for the ERFs. For the power analyses, we computed the power separately for the two planar gradient directions, and combined the power data by averaging the two. To visualize the responses per tagging frequency (Figure 37), we used a notch filter between 60 and 62 Hz to display the ERF at 68 Hz, and a notch filter between 67 and 69 Hz to display the ERF at 61 Hz.

We then performed a spectral analysis on an individual's ERF data pooled over conditions, in the time-window in which both the auditory and visual stimulus unfolded (0.5 - 1.5 s), and a post-stimulus baseline (2.0 - 3.0s). We chose this post-stimulus time-window because, contrary to the pre-stimulus time-window, it is not affected by the button press of the 4-alternative forced choice identification task. This allowed for a spectral decomposition of an individual's ERF data. We computed power spectra of 1 - 130 Hz over both the baseline and stimulus window using fast Fourier transform and a single Hanning taper of the 1s segments. This data was then averaged over conditions, and the stimulus window was compared to the baseline window.

9.3.5.3. Frequency-tagging analyses - Source-level

To reconstruct activity at the tagging frequencies, we calculated coherence coefficients between a dummy modulation signal at either 61Hz or 68Hz and the observed MEG signal at those frequencies. Although the phase of the tagging was designed to be identical over trials, the projector that we used occasionally experienced a delay in presenting the video material (in 16 of the 26 participants). We corrected for this by translating any observed delays between video onset and offset markers into a phase-difference, which was then subtracted from the signal. Note that this correction only uses information in the stimulus marker channel

and the length of the original video files, and does not rely on any information in the measured MEG signal.

We then performed source analysis to localize the neuronal sources that were coherent with the modulation signal at either 61Hz or 68 Hz, and compared the difference in coherence in the stimulus and post-stimulus window to localize potential regions of interest (ROI) for the auditory and visual tagged frequencies. This was done pooled over conditions. Source analyses on coherence values (for 61 and 68 Hz) and power values (for the intermodulation frequency at 7Hz, see results), was performed using dynamic imaging of coherent sources (DICS; Gross et al., (2001)) as a beamforming approach. We computed a common spatial filter per subject from the lead field matrix and the cross-spectral density matrix (CSD) that was the same for all conditions. An individual's leadfield was obtained by spatially co-registering an individual's anatomical MRI to the MEG data by the anatomical markers at the nasion and left and right periaucular points. Then, for each participant, a single-shell headmodel was constructed on the basis of the MRI (Nolte, 2003). A source model was created for each participant by warping a 10mm-spaced grid defined in MNI space to the individual participant's segmented MRI. The MNI template brain was also used for those participants (2/26) that did not have an individual MRI scan.

The ROIs for the auditory and visually tagged signals were defined by taking the grid points that showed the highest 20% of coherence difference values between stimulus and baseline. For these ROIs, coherence difference values were extracted per condition. The ROI for the intermodulation frequency at 7 Hz was defined by taking those grid points showing the 20% highest power difference values between stimulus and baseline. For this ROI, power difference values were extracted per condition.

9.3.5.4. Statistical comparisons

As we predefined our frequencies of interest and have specific regions of interest for analysis, we compared the differences between values per conditions using 2x2 repeated measures ANOVA's, with the factors Speech (clear/degraded) and Gesture (matching/mismatching). Correction for multiple comparisons in the

post-hoc tests was applied by using a Bonferroni correction.

9.4. Results

Participants watched videos of an actress uttering action verbs in clear or degraded speech, accompanied by a matching or mismatching gesture. After the video, participants were asked to identify the verb they heard in a 4-alternative forced choice identification task. Video presentation was manipulated by tagging the gesture space in the video by 68 Hz flicker, while the sound in the videos was tagged by 61 Hz amplitude modulation.

9.4.1. Behavioral results

In our behavioral task we replicated previous results (see Drijvers, Ozyürek, & Jensen, 2018; Drijvers & Özyürek, 2018) by demonstrating that when the speech signal was clear, response accuracy was higher than when speech was degraded ($F(1, 25) = 301.60$, $p < .001$, partial $\eta^2 = .92$) (mean scores and SDs: CM: 94.7% (SD = 4.0%), CMM: 90.2% (SD = 5.6%), DM: 85.0% (SD = 8.2%), DMM: 66.5% (SD = 7.8%)). Similarly, response accuracy was higher when a gesture matched compared to mismatched the speech signal ($F(1, 25) = 184.29$, $p < .001$, partial $\eta^2 = .88$). The difference in response accuracy was larger in degraded speech than in clear speech ($F(1, 25) = 4.87$, $p < .001$, partial $\eta^2 = .66$).

Similar results were found in the reaction times. Participants were faster to identify the verbs when speech was clear as compared to when speech was degraded ($F(1, 25) = 198.06$, $p < .001$, partial $\eta^2 = .89$) (mean RTs and SDs: CM: 1086.3 ms, SD = 177.1 ms, CMM: 1127.92 ms, SD = 153.84 ms, DM: 1276.96 ms, SD = 230.13 ms, DMM: 1675.77 ms, SD = 246.69 ms). Participants were faster to identify the verbs when the gesture matched the speech signal as compared to when the gesture mismatched the speech signal ($F(1, 25) = 105{,}42$, $p < .001$, partial $\eta^2 = .81$). This difference in reaction times was larger in degraded speech than in clear speech ($F(1, 25) = 187{,}78$, $p < .001$, partial $\eta^2 = .88$).

9.4.2. MEG results - Frequency-tagging

9.4.2.1. Both visual and auditory frequency tagging produce a clear steady-state response that is larger than baseline

As a first step, we calculated the time-locked averages of the event-related fields pooled over conditions. Auditory frequency tagging produced an auditory steady-state response over left and right-temporal regions (see Figure 37A), and visual frequency tagging at 68Hz produced a clear visual steady-state response at occipital regions (see Figure 37B).

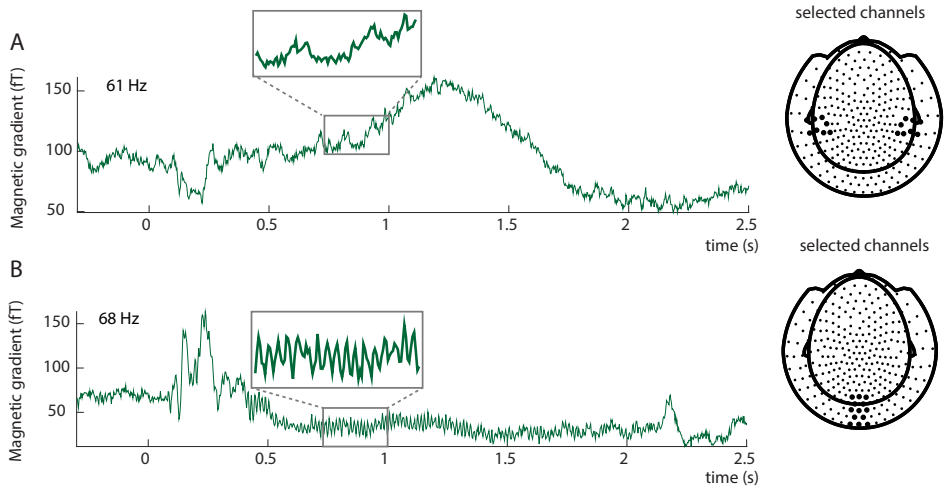


Figure 37: Event-related fields show clear responses at both tagged frequencies. Auditory input was tagged by 61 Hz amplitude modulation (A), Visual input was tagged by 68 Hz flicker (B). The highlighted parts reflect an enlarged part of the signal to clearly demonstrate the effect of the tagging on the event-related fields. Tagging was phase-locked over trials. A: Average ERF from single subject for 61 Hz at selected sensors overlying the left and right temporal lobe. The highlighted sensors in the right plot reflect the sensors for which the ERF is plotted. B: Average ERF for 68 Hz from single subject at selected channels overlying occipital cortex. The highlighted channels in the right plot reflect the sensors for which the ERF is plotted. ERFs are plotted using planar gradient data (combined using root mean squared) for visualization purposes.

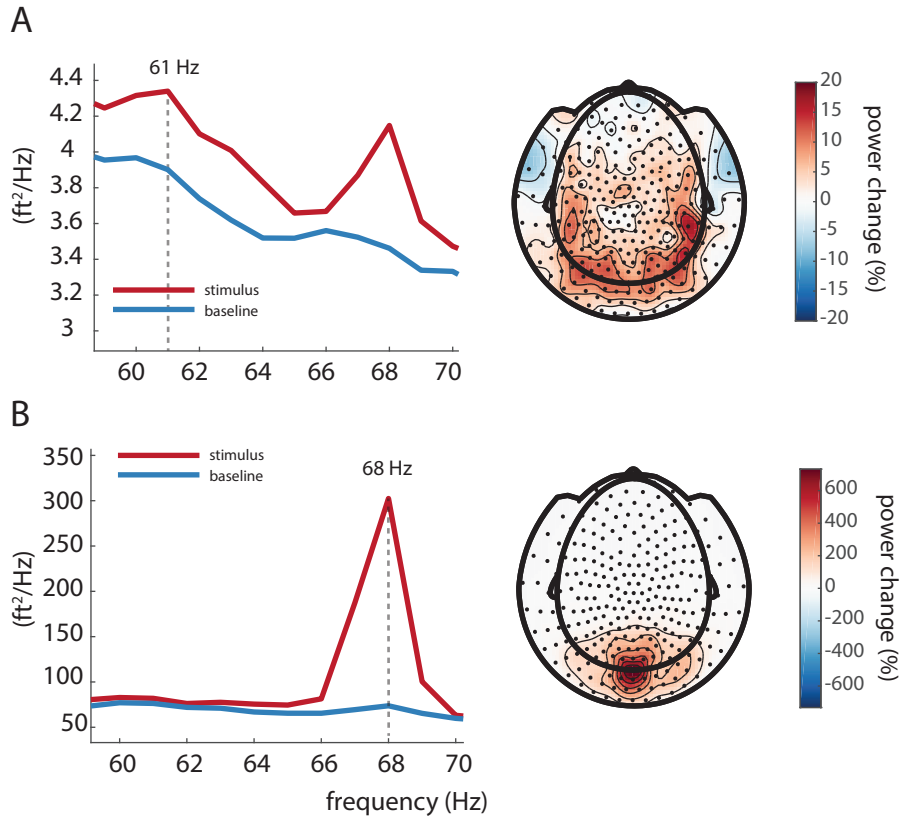


Figure 38: A: Power is largest at the tagged frequency of the auditory stimulus (61 Hz). Note the 68Hz tagged signal of the visual stimulus is still visible at left- and right-temporal sensors. 61Hz power is stronger in the stimulus time window than in the baseline time window, and is widely spread over posterior regions, with maxima at right-temporal regions. B: A massive power increase is observed at the tagged frequency (68 Hz) for the visual stimuli. 68Hz power is larger in the stimulus than in the baseline window and is strongest over occipital regions.

Both visual and auditory responses at the tagged frequency were larger in the stimulus time window than in the baseline time window. To compare these two intervals, we used similar sensors as we used for the ERF analyses described above, and plotted the difference in spectral power calculated from the ERF power between the stimulus time window (0.5 - 1.5 s) and a post-stimulus baseline (2.0 - 3.0s). The stimulus time window was based on the time interval where speech and gestures were both unfolding. A post-stimulus instead of a pre-stimulus baseline was chosen as the pre-stimulus baseline was contaminated by the button press of

the 4-alternative forced choice identification task. Note that the visually tagged signal at 68Hz seems to be more focal and strong than the auditory tagged signal at 61Hz (see Figure 38). These analyses confirm that we were able to induce high-frequency steady-state responses simultaneously for both auditory and visual stimulation.

9.4.2.2. Coherence is strongest at occipital regions for the visually tagged signal (68 Hz)

To compare experimental conditions and to identify the neural generators of the tagged signals, we adopted an ROI approach. We computed source-level coherence coefficients for all conditions pooled together to identify the visual ROI. This was done by computing coherence between a visual dummy 68Hz modulation signal and the observed MEG data. The relative coherence increase between stimulus and baseline was largest in occipital regions (see Figure 39A). We then formed our visual ROI by selecting those grid points exceeding the 80th percentile of coherence increase. For each participant, the percentage of change in coherence between stimulus and baseline was computed in that ROI per condition and compared in a 2x2 (Speech: clear/degraded, Gesture: matching/mismatching) RM-ANOVA (see Figure 39B). Coherence change was larger for videos containing clear speech than videos containing degraded speech ($F(1, 25) = 17.14$, $p < .001$, partial $\eta^2 = .41$), but did not differ between matching or mismatching trials ($F(1, 25) = 0.025$, $p = .87$). We observed a significant interaction between Speech and Gesture ($F(1, 25) = 26.87$, $p < .001$). Pairwise comparisons revealed a stronger coherence change in videos containing clear speech and a matching gesture (CM) than clear speech and a mismatching gesture (CMM) ($t(25) = 3.26$, $p_{bon} = .015$), and a stronger coherence change in videos containing degraded speech and a mismatching gesture (DMM) than in videos containing degraded speech and a matching gesture (DM) ($t(25) = -4.03$, $p_{bon} < .001$). Coherence change was larger in CM than in DM ($t(26) = 6.59$, $p_{bon} < .001$), but not larger in CM than in DMM ($t(26) = 2.02$, $p_{bon} = .27$), and not larger in CMM compared to DMM ($t(26) = -1.74$, $p_{bon} = .48$). These results thus indicate that visual regions processed the frequency-tagged gestural signal more strongly when speech was clear than when

speech was degraded. This suggests that when speech is clear, participants allocate more visual attention to gestures than when speech is degraded.

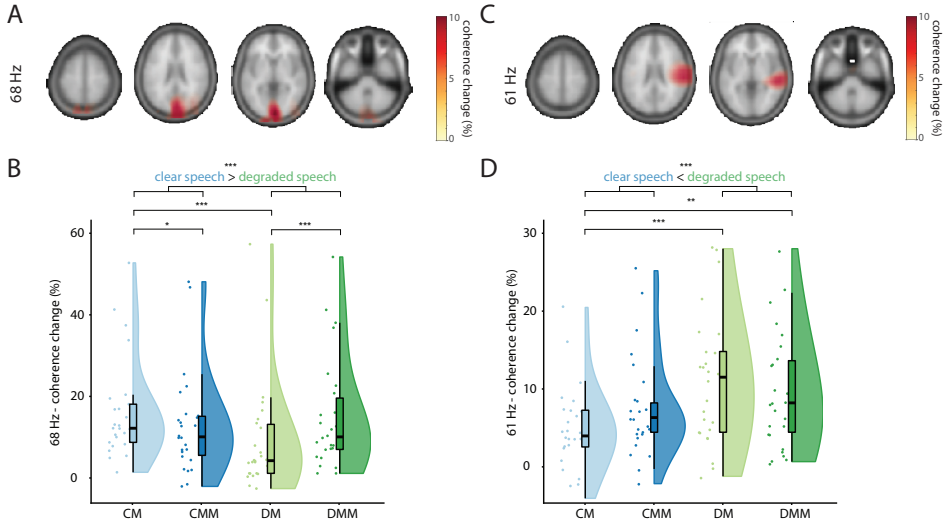


Figure 39 Sources of the visually tagged signal at 68Hz (A/B) and sources of the auditory tagged signal at 61 Hz (C/D), and individual scores in the respective ROI per condition (CM/CMM/DM/DMM). A: Coherence change in percentage when comparing coherence values in the stimulus window to a post-stimulus baseline for 68Hz (the frequency of the visual tagging), pooled over conditions. Only positive coherence values are plotted (>80% of maximum). Coherence change is largest over occipital regions for the visually tagged signal. B: Coherence change values in percentage extracted from the 68Hz ROI with the 20% highest coherence values per condition. Raincloud plots reveal raw data, density and boxplots for coherence change. C: Coherence change in percentage when comparing coherence values in the stimulus window to a post-stimulus baseline for 61Hz (the frequency of the auditory tagging), pooled over conditions. Only positive coherence values are plotted (>80% of maximum). Coherence change is largest over right-temporal regions. D: Coherence change values in percentage extracted from the 61 Hz ROI with the 20% highest coherence values per condition. Raincloud plots reveal raw data, density and boxplots for coherence change.

9.4.2.3. Coherence is strongest at right-temporal regions for the auditory tagged signal (61 Hz)

Similar to the visually tagged signal, we first computed coherence coefficients for all conditions pooled together to identify the auditory tagged ROI. This was done by computing source-level coherence between a dummy 61Hz modulation signal

(reflecting the auditory tagging drive) and the observed MEG data. The coherence difference between stimulus and baseline peaked at right temporal regions (Figure 39C). We then formed the auditory ROI by selecting those grid points exceeding the 80th percentile of coherence change. Again, coherence change values per condition and per participant were compared in a 2x2 RM-ANOVA (see Figure 39D). Coherence was larger in degraded speech conditions than in clear speech conditions ($F(1, 25) = 12.87$, $p = .001$, partial $\eta^2 = .34$), but did not differ between mismatching and matching conditions ($F(1, 25) = 0.09$, $p = .77$, partial $\eta^2 = .04$). No interaction effect was observed ($F(1, 25) = 3.13$, $p = .089$, partial $\eta^2 = .11$). Pairwise comparisons revealed that there was no difference in coherence when comparing CM and CMM ($t(25) = -1.44$, $p_{bon} = .81$), or between DM and DMM ($t(25) = 1.38$, $p_{bon} = .90$). Coherence was larger in DM than in CM ($t(25) = -4.24$, $p_{bon} < .001$), and in DMM than in CM ($t(25) = -3.90$, $p_{bon} < .01$) but not when comparing CMM to DMM ($t(25) = -1.40$, $p_{bon} = .87$). These results thus indicate that right-lateralized auditory regions processed the frequency-tagged auditory signal more strongly when speech was degraded than when speech was clear. This suggests that when speech is degraded, participants allocate more auditory attention to speech than when speech is clear.

9.4.2.4. An intermodulation frequency was observed at 7Hz (f_2-f_1) but not 129 Hz (f_2+f_1)

To test whether intermodulation frequencies (f_2-f_1, f_2+f_1) could be observed, we then calculated power spectra of the ERFs in the stimulus time window and the post-stimulus time window at 7Hz and 129 Hz. Only for 7Hz a difference between stimulus and baseline was observed at left frontal and left temporal channels (Figure 40A). No reliable differences were observed for 129 Hz (Figure 40B).

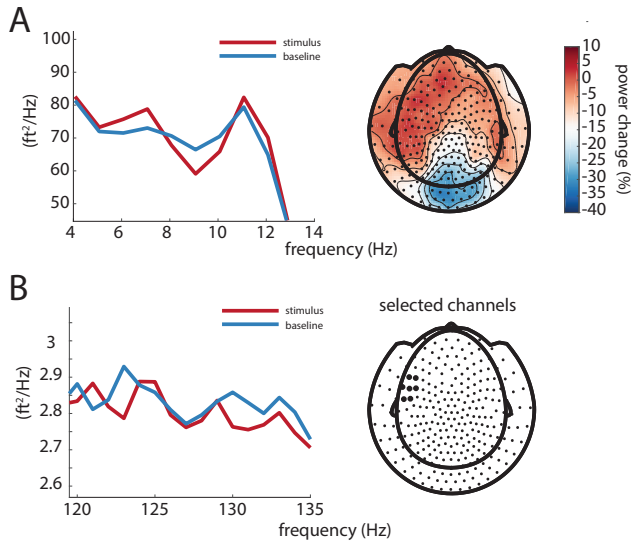


Figure 40 An intermodulation frequency could be observed at 7Hz (f_2-f_1) (A) but not 129 Hz (f_2+f_1) (B). A: 7Hz power in the stimulus window is larger than baseline over left-temporal and left-frontal sensors. B: No difference could be observed at 129 Hz between stimulus and baseline. The black highlighted channels represent the sensors at which the power spectra of the ERFs was calculated.

As a next step, we then took a similar approach as for the visually and auditory tagged stimuli and calculated the coherence difference between stimulus and baseline at 7Hz, pooled over conditions (see Figure 41A). This was done by computing source-level coherence between a dummy 7Hz modulation signal (the intermodulation of our 61 and 68 Hz tagging signals) and the observed MEG data, to identify an ROI for further analyses between the condition differences. The coherence analysis did not reveal any differences between stimulus and baseline. It should be noted here that our frequency-tagged signals at f_1 and f_2 were exactly phase-consistent across trials, since the phase was uniquely determined by the stimuli themselves. However, it is possible that the phase of the intermodulation signal has a much weaker phase consistency across trials, since it depends not only on the stimuli but also on the nature of the non-linear neural interaction. If this is the case, we might still observe an effect for the power at the intermodulation frequency, rather than the coherence. We therefore performed source analysis on the power of the combined conditions versus baseline. Here, we observed a power

change at 7Hz in left frontal and temporal regions that mirrored the effect we observed at sensor level (Figure 41B). We then took the grid points corresponding to the power values that exceeded the 80th percentile to compare the strength of this 7Hz signal between conditions by using a 2x2 RM-ANOVA (Figure 41C). Power was larger in clear speech conditions than in degraded speech conditions ($F(1, 25) = 10.26$, $p = .004$, partial $\eta^2 = .29$), but did not differ between matching and mismatching trials ($F(1, 25) = 0.01$, $p = .91$). Pairwise comparisons revealed that 7Hz power was not different for CM compared to CMM ($t(25) = 1.14$, $p_{bon} = 1$), and not different for DM compared to DMM ($t(25) = -.67$, $p_{bon} = 1$). However, 7Hz power was larger in CM than in DM ($t(25) = 3.01$, $p_{bon} = .025$), and larger in CM than in DMM ($t(25) = 2.82$, $p_{bon} = .045$). No difference was observed between CMM and DMM ($t(25) = 1.61$, $p_{bon} = .6$). These results thus demonstrate that left-frontotemporal regions processed the intermodulation signal (i.e., the interaction between the auditory and visually tagged signal) more strongly when speech was clear than when speech was degraded, and most strongly when speech was clear and a matching gesture was present. This suggests that the interaction between the auditory and visually tagged signal is strongest when integration is easiest.

9.5. Discussion

In the current MEG study we provide a proof-of-principle that rapid invisible frequency tagging (RIFT) can be used to tag visual and auditory input at high frequencies, as well as differences in the power at the tagged frequencies per condition. Power of the visually-tagged input was strongest over occipital regions, and strongest when speech was clear. Power of the auditory-tagged input was strongest over right-temporal regions and strongest when speech was degraded. Second, we were able to identify an intermodulation frequency at 7hz (f_2-f_1) as a result of the interaction between a visually frequency-tagged signal (gesture; 68 Hz) and an auditory frequency-tagged signal (speech; 61 Hz). In line with our hypotheses, this intermodulation frequency was strongest in LIFG and left-temporal regions, (pSTS/MTG), especially when integration of these audiovisual inputs was easier (i.e., when speech was clear and a gesture matched the speech signal). Below we provide putative interpretations of these results.

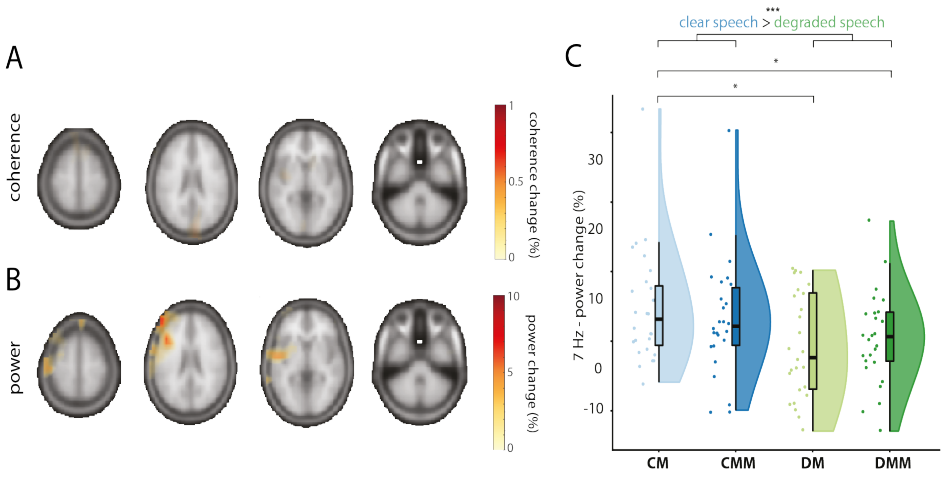


Figure 41 Sources of the intermodulation frequency (f_2-f_1) at 7Hz and individual scores in the left-frontotemporal ROI per condition (CM/CMM/DM/DMM). A: Coherence change in percentage when comparing coherence values in the stimulus window to a post-stimulus baseline for 7Hz (intermodulation frequency, f_2-f_1), pooled over conditions. Only positive coherence values are plotted (>80% of maximum). No differences could be observed. B: Power change in percentage when comparing coherence values in the stimulus window to a post-stimulus baseline for 7Hz, pooled over conditions. Power changes were largest in left-frontal and left-temporal regions. Only positive coherence values are plotted (>80% of maximum). C: Power change values in percentage extracted from the 7Hz ROI with the 20%highest coherence values per condition. Raincloud plots reveal raw data, density and boxplots for power change.

9.5.1. Clear speech enhances visual attention to gestural information

In occipital regions, we observed a stronger drive by the 68 Hz visual modulation signal when speech was clear than when speech was degraded. This is in line with previous eye-tracking work that demonstrated that when speech is degraded, listeners gaze more often to the face and mouth than to gestures to extract phonological information to aid comprehension (Drijvers, Vaitonyte, Ozyurek, in revision), as well as previous work that revealed that the amplitude of SSVEPs was enhanced by visual attention, irrespective of whether the stimuli were task-relevant (Morgan, Hansen, Hillyard, & Posner, 1996; Müller et al., 2006). Note that gestural information is often processed in the periphery of a listener's visual field (Gullberg & Holmqvist, 1999, 2002, 2006; Gullberg & Kita, 2009). As listeners do not necessarily need to extract the phonological information conveyed by the

lips when speech is clear, overt visual attention might be directed to a 'resting' position in the middle of the screen during clear speech processing, resulting in stronger coherence when speech is clear than when speech is degraded. Pairwise comparisons of the conditions revealed that in clear speech, coherence was larger when the gesture matched than mismatched with the signal. In line with the interpretation above, a listener might have reconsidered the auditory input when noticing that the gesture mismatched the perceived auditory input, and might have directed their attention to the face/lips of the actress, which, in turn, reduces visual attention to the gesture.

However, we observed a reverse direction of this effect when speech was degraded, and observed stronger coherence when the gesture mismatched than matched the degraded speech signal. We speculate that when speech is degraded and a gesture matches the signal, a listener might more strongly allocate visual attention to the information conveyed by the face/lips, so that information conveyed by the lips and the information conveyed by the gesture can jointly aid in disambiguating the degraded speech signal (Drijvers & Ozyürek, 2017). However, when speech is degraded and a gesture mismatches the signal, the uncertainty of both inputs may result in a visual reconsideration of both inputs, and thus a less fixed locus of attention. These interpretations are rather speculative, and further work is needed to disambiguate different interpretations. For example, future work could consider tagging the lip-region to further investigate how a listener allocates visual attention to these two visual articulators during comprehension.

9.5.2. Degraded speech enhances auditory attention to speech information

In line with our hypotheses, we observed stronger drive by the 61 Hz amplitude modulation signal in temporal areas overlapping with auditory cortex when speech was degraded than when speech was clear. This response was strongest at right-temporal regions, which is in line with previous work that demonstrated that for speech stimuli, the ASSR is often localized to right-lateralized sources (Lamminmäki et al., 2014; Ross, Herdman, & Pantev, 2005). Although both left- and right-lateralized process speech, right-lateralized sources are often observed because right-lateralized regions are sensitive to spectral changes and prosodic

information, and processing of low-level auditory cues (Zatorre & Gandour, 2008; Scott et al., 2000).

Previous work has reported enhanced ASSR responses to amplitude-modulated multi-speech babble when attention to this input increased (Keitel, Schro, & Mu, 2011; Ross, Picton, Herdman, & Pantev, 2004; Saupe, Widmann, Bendixen, & Mu, 2009; Talsma et al., 2010; Tiitinen et al., 1993). The enhanced ASSR which we observed in the degraded compared to clear speech conditions could thus reflect an increase in attention to the speech signal when speech is degraded. Note that no differences in coherence were observed when comparing matching and mismatching gestures in both clear and degraded speech. As the gesture congruency manipulation is a visual manipulation, this could mean that modulation of the ASSR is modality-specific (Parks, Hilimire, & Corballis, 2011; Rees, Frith, & Lavie, 2001). However, an alternative interpretation is that our tagging frequency in the auditory modality was too high to induce strong enough steady-state responses to observe condition effects, as previous work has demonstrated that ASSRs are strongest around 40 Hz (Galambos, Makei, & Talmachoff, 1981). Future work could consider using a lower frequency as the tagging frequency for auditory stimuli when studying multimodal integration.

9.5.3. The auditory tagged speech signal and visually tagged gesture signal interact in left-frontotemporal regions

We set out to study whether intermodulation frequencies could be identified in a multimodal, semantic context as a result of the interaction of the visual and auditory tagged signals. In contrast to previous work by Giani et al., (2012) using lower frequencies, we did observe an intermodulation frequency when the two inputs interacted at 7Hz (f_2-f_1), but not at 129 Hz (f_2+f_1). As responses in lower frequencies tend to be stronger than in higher frequencies, the higher-frequency intermodulation frequency might not have been identifiable due to signal-to-noise ratio. This might partially be caused by the use of an auditory tagging frequency of 61 Hz, which induced considerably smaller responses than the visually-tagged stimuli.

Note that although we observed a stronger 7Hz peak at sensor level in the stimulus window than the baseline window, we did not observe stronger coherence between a dummy signal at 7 Hz and the observed MEG data at source level. This indicates that the phase of the intermodulation signal is not as consistent over trials as the f_1 and f_2 signals, which in turn might imply that the time point of interaction of the two signals differs across trials. This is reasonable to assume, when considering that degraded words might take longer to recognize and integrate with gestural information than clear words. This could explain why we observed a clear difference between stimulus and baseline when we reconstructed the sources of the intermodulation frequency on the basis of power, but not coherence.

In line with our hypotheses, the source of the intermodulation frequency was localized in LIFG and left-temporal (pSTS/MTG) regions. These areas are thought to be involved in the integration of speech and gestures (Dick, Mok, Raja Beharelle, Goldin-Meadow, & Small, 2014; Drijvers, Ozyürek, & Jensen, 2018a, 2018b; Holle, Gunter, Ruschemeyer, Hennenlotter, & Iacoboni, 2008; Holle, Obleser, Rueschemeyer, & Gunter, 2010; Kircher et al., 2009; Straube, Green, Weis, & Kircher, 2012; Willems, Özyürek, & Hagoort, 2007, 2009; Zhao, Riggs, Schindler, & Holle, 2018, see for an overview: Ozyurek, 2014). We thus propose that the observed intermodulation signal is a direct reflection of the non-linear integration of speech and gesture information in LIFG and left-temporal regions (pSTS/MTG).

However, although the power at this intermodulation frequency was stronger in clear speech conditions than in degraded speech conditions, we did not observe an effect of gesture congruency. This could reflect that the interaction of the auditory and visual signal solely conveys the ease of lower-level integration of the two inputs, and how strongly the combined information interacts in certain regions. This interaction is then expected to be easier when speech is clear than when speech is degraded. As the congruency manipulation targets a higher level of integration than the auditory manipulation, this higher-level integration of semantic information might be mediated by other mechanisms than what is reflected by the intermodulation frequency that is observed here.

9.5.4. Proof of principle: using RIFT to study the integration of complex and dynamic audiovisual stimuli in a semantic context.

The current MEG study provides a proof of principle of the use of rapid invisible frequency tagging (RIFT) to study the integration of audiovisual stimuli, and is the first study to identify intermodulation frequencies as a result of the interaction between auditory and visual stimuli in a semantic context. Note that although previous work has reported the occurrence of intermodulation frequencies in a non-semantic context (Regan et al., 1995b), other work failed to identify between-modality intermodulation frequencies (Giani et al., 2012). This could be due to the fact that lower frequencies were used for tagging. Another possibility is that this might have been due to the nature of the stimuli that have been used in these studies. As Giani et al., (2012) suggest, the occurrence of intermodulation frequencies resulting from audiovisual integration of non-semantic inputs such as tones and gratings might reflect low-level spatiotemporal coincidence detection that is prominent for transient stimuli, but less for sustained steady-state responses. Similarly, previous fMRI work that investigated the difference between transient and sustained BOLD responses revealed that primary auditory and visual regions were only involved in the integration of rapid transient stimuli at stimulus onset. However, integration for sustained responses did involve higher-order areas (Werner & Noppeney, 2011). The observed 7Hz intermodulation frequency in response to our semantic audiovisual stimuli was also localized to higher-order areas, rather than early sensory regions. This again underlines the possibility that the observed intermodulation frequency in the current study reflects the ease of integration of these audiovisual stimuli in certain higher-order regions. However, note that many biologically plausible neural models for implementing non-linear neuronal operations have been reported (Kouh & Poggio, 2008). At this point, it thus remains unclear whether the presence of intermodulation frequencies reflects specific computational or neuronal processes.

An important advantage of using RIFT is that spontaneous neuronal oscillations in lower frequencies were not entrained by our tagging frequencies. This might explain why a clear intermodulation frequency was observed in the current study, but was less easy to identify in previous work. Future studies might thus consider

exploiting this feature and using RIFT to study the role of these lower frequency oscillations in sensory processing.

9.6. Conclusion

We provided a proof of principle that RIFT can be used to tag visual and auditory inputs at high frequencies. Second, we demonstrated that RIFT can be used identify intermodulation frequencies in a multimodal, semantic context. The observed intermodulation frequency was the result of the interaction between visually and auditory tagged stimuli, and was localized in LIFG and pSTS/MTG, areas known to be involved in speech-gesture integration. The strength of this intermodulation frequency was strongest when integration between speech and gestures was easiest. In conclusion, we thus propose that the strength of this intermodulation frequency reflects the ease of semantic audiovisual integration and that the combined input interacts in down-stream higher order areas.

9.7. Acknowledgements

This work was supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from Netherlands Organization for Scientific Research. OJ was supported by James S. McDonnell Foundation Understanding Human Cognition Collaborative Award [220020448] and the Royal Society Wolfson Research Merit Award. LD notes that ES coined the acronym RIFT. We are very grateful to Nick Wood (†), for helping us in editing the video stimuli, and to Gina Ginos, for being the actress in the videos.

Chapter 10

General discussion & Conclusion

Imagine you're in that crowded cafe again to have another drink with that friend. When you enter the noise-filled seating area, you can hear your friend shouting something while her hand mimics a glass that is being brought up to the mouth. Although it is difficult to hear her due to the grinding coffee machines in the background, you probably would understand what she means due to that drinking-gesture.

Integrating audiovisual cues is an important strategy to reduce uncertainty under adverse conditions, and has been thoroughly studied in non-semantic audiovisual contexts (e.g., Rohe & Noppeney, 2016, 2018; Saldern & Noppeney, 2013). In this thesis, I reported eight experimental studies that investigated the behavioral and neural integration of audiovisual integration at a semantic level, by investigating how gestures, such as that drinking-gesture, enhance language comprehension under adverse listening conditions. These adverse listening conditions were induced by external factors, such as speech degradation, and induced by internal factors, such as when you are a non-native listener of a language.

In this thesis I have demonstrated on a behavioral and neural level that gestures enhance language comprehension in both externally and internally induced adverse listening conditions. Mechanistically, the studies reported in this thesis were grounded in a theoretical framework that assumes that increases and decreases of oscillatory power in different frequency bands are involved in enabling integration of information from different modalities and engaging relevant brain regions in this process over time (Jensen & Mazaheri, 2010; Varela et al., 2001). I demonstrated that oscillatory power decreases in the alpha (8-12 Hz) and beta (13-30 Hz) band reflected the engagement of the extended language network (LIFG/pSTS/MTG), motor regions and visual regions during speech-gesture integration. I provided evidence that these alpha and beta power modulations seemed to be general for both types of adverse listening conditions and that they support similar core processes involved in unification, simulation and lexical access to aid comprehension. However, the distinct spatiotemporal time courses of these oscillatory modulations suggest different processing strategies of semantic audiovisual information dependent on the type of adverse

listening condition. Next, I will provide summaries of the core findings of each chapter. Then, I will compare how native and non-native listeners benefit from visual information in clear and adverse listening conditions, and how these results inform current models of the neural integration of speech and gestures. Finally, I propose ideas for future investigations to understand how a listener weighs and integrates auditory and visual semantic information in adverse listening conditions to enhance comprehension.

10.1. Summary of core findings

Chapters 2-4 investigated how native listeners integrate speech and gestures in externally induced adverse listening conditions. In **chapter 2** I investigated to what extent iconic gestures, on top of visible speech, enhanced degraded speech comprehension for native listeners when speech was severely noise-vocoded, moderately noise-vocoded and when speech was clear. I demonstrated that native listeners benefitted most from having both iconic gestures and visible speech present, as compared to having just visible speech present, or having only auditory information present. This enhancement effect was mostly prevalent at a moderate noise-vocoding level, where the auditory signal was reliable enough for native listeners to be able to integrate the information from both visible speech and gestures to aid in comprehension.

The results in **chapter 2** formed the basis of the MEG experiment described in **chapter 3**. Here, I characterized modulations of neuronal oscillations to investigate the spatiotemporal dynamics underlying gestural enhancement of moderately degraded speech comprehension. I provided evidence that alpha (8-12 Hz) and beta (13 - 30 Hz) oscillatory power decreases and gamma (65 - 80 Hz) power increases reflected active processing and engagement of the hand-area of the motor cortex, visual regions and the extended language network (LIFG, STS, STG/MTG) when gestures enhanced degraded speech comprehension. I concluded that these modulations support general unification and lexical access processes and suggested that listeners simulate gestures as well as allocate more visual attention to these gestures to aid comprehension of degraded speech. Strikingly, the observed individual low- and high-frequency power modulations

during gestural enhancement of degraded speech comprehension were also predictive of how much benefit a listener experienced from this gesture during degraded speech comprehension.

I probed speech-gesture integration in a different manner by using a violation paradigm in **chapter 4**. Here, I asked whether and how oscillatory modulations support the semantic integration of speech and gestures when integration load was manipulated by auditory (e.g., degraded speech) and visual (e.g., gesture congruency) factors. I showed that when visual factors manipulated integration load and a mismatching gesture accompanied speech, a stronger alpha/beta power suppression reflected a larger engagement of LIFG, motor and visual regions than when a matching gesture accompanied speech. However, when auditory factors increased integration load and speech was degraded, pSTS/MTG and medial temporal regions were less engaged, possibly reflecting the hindered integration of gestures and the degraded signal when the gesture does not aid lexical access. These results were, spatially and temporally, similar to the observed oscillatory power modulations in **chapter 3.** This also demonstrated that both enhancement (**chapter 3**) and congruency (**chapter 4**) manipulations tapped into similar processes underlying speech-gesture integration.

In **chapters 5-8** I reasoned that not only externally factors, such as speech degradation, can impact language comprehension, but also internal factors, such as being a non-native listener (following Mattys, Davis, Bradlow, & Scott, 2012). In these chapters I thus switched the focus from studying only externally induced adverse listening conditions to studying both externally and internally adverse listening conditions, to test whether the patterns that were observed in externally adverse listening conditions were specific to that type of adverse listening condition, or were more general.

To test this, I conducted a similar behavioral experiment with non-native listeners in **chapter 5** as the behavioral experiment with native listeners that I described in **chapter 2**. In **chapter 5**, I investigated whether and to what extent visible speech enhances degraded speech comprehension for non-native listeners, and whether non-native listeners also experienced an additive benefit from gestures on top of visible speech in severely degraded speech, moderately degraded speech

and clear speech. Additionally, these results were directly compared to the results in **chapter 2** to describe differences between native and non-native listeners. In this experiment I demonstrated that non-native listeners showed a similar yet smaller gestural enhancement effect in moderately degraded speech than native listeners. Furthermore, in contrast to native listeners, non-native listeners did not experience any benefit from visible speech when speech was severely degraded. I concluded that non-native listeners thus might need more auditory information than native listeners to benefit from both visible speech and gestures in a joint context.

To further investigate possible differences in speech-gesture integration in externally and internally induced adverse listening conditions, I then used EEG in **chapter 6** to study modulations of the N400 component during speech-gesture integration in clear and degraded speech, and whether this differed for native and non-native listeners. The N400 is an ERP component known to be sensitive to semantic unification operations. I showed that native listeners demonstrated a larger N400 in response to mismatching gestures than matching gestures when speech was clear, and that the N400 amplitude was even more negative when speech was degraded. For non-native listeners I observed a difference in N400 amplitude between mismatching and matching gestures in clear speech, but not in degraded speech. Interestingly, the N400 effect found in clear speech was larger for non-native listeners than for native listeners. I thus concluded, in line with the results from **chapter 5**, that native and non-native listeners differ in the extent to which the semantic information from the gesture is coupled to the degraded speech signal on a neural level. Non-native listeners might require more auditory cues to optimally make use of gestural information during comprehension. However, when speech was clear, non-native listeners might focus more on gestures than native listeners, as they might be less confident about their comprehension of the speech signal.

In **chapter 7**, I conducted a similar MEG experiment with non-native listeners as the MEG experiment with native listeners described in **chapter 3** (i.e., using gestural enhancement of moderately degraded speech), and directly compared the results of these two studies to each other to investigate whether differences

between internally and externally induced adverse listening conditions would affect the oscillatory dynamics associated with processing multimodal language in these contexts. Additionally, I investigated whether these oscillatory modulations were predictive of how much a listener benefitted from gestural information during language comprehension in adverse listening conditions.

Similar to native listeners, non-native listeners demonstrated an alpha/beta power suppression over the extended language network (LIFG/pSTS/MTG), motor and visual regions when gestures enhanced degraded speech comprehension. These oscillatory modulations suggest similar core processes supporting lexical access and unification as were observed in native listeners. I also demonstrated that a listener's individual spatiotemporal oscillatory modulations could predict how much a listener benefits from the semantic information conveyed by gestures in both externally and internally induced adverse listening conditions. However, when comparing the two listener groups, non-native listeners engaged the mouth area of the primary somatosensory cortex, LIFG and anterior temporal lobe (ATL) less than native listeners. In line with **chapter 5 and 6**, these results thus again suggested that non-native listeners might be hindered in processing and coupling degraded auditory cues to the semantic information conveyed by the gesture.

In **chapter 8**, I used eye-tracking as a means to measure overt visual attention to gestures during speech comprehension in externally and internally induced adverse listening conditions. The results of **chapters 3, 4, 6 and 7** suggested that when speech is degraded, listeners might allocate more visual attention to gestures to aid comprehension. However, studying ERP components and oscillatory modulations does not provide direct evidence for this claim. Using eye-tracking, I demonstrated that both native and non-native listeners gaze more at the face than at gestures or the torso during comprehension (Rogers, Speelman, Guidetti, & Longmuir, 2018), especially when speech was degraded. Both native and non-native listeners also allocated more visual attention to gestures when speech was clear than when speech was degraded. This is probably due to the fact that native and non-native listeners both showed sustained visual attention to the face and mouth rather than to gesture when phonological cues were hard to disambiguate. In general, non-native listeners however allocated more visual

attention to gestures than native listeners, but as disambiguating the degraded auditory cues was more challenging for them, the use of the semantic information conveyed by the gesture might have been hindered. This probably caused non-native listeners to benefit less from gestures than native listeners. Finally, visual attention to gestures during degraded speech comprehension was only predictive of how much a listener benefitted from a gesture in native but not non-native listeners. Again, these results thus showed that for non-native listeners it might be more challenging to process the degraded auditory cues and couple them to the phonological information conveyed by visible speech, which hinders the use of gestures to disambiguate the degraded speech signal. Native listeners on the other hand are more able to resolve the degraded auditory cues than non-native listeners, which allowed them to immediately benefit from both visual articulators in a joint context during comprehension.

Finally, in **chapter 9**, I provided a proof of principle that rapid invisible frequency tagging (RIFT) can be used to tag visual and auditory inputs at high frequencies, as well as to study the lower-level interaction of auditory and visual information during speech-gesture integration. In previous literature, frequency-tagging has been only applied at low frequencies. This might be problematic when considering that spontaneous neuronal oscillations at lower frequencies (e.g., alpha/beta oscillations) are also likely entrained by frequency-tagging (Keitel et al., 2014; Spaak et al., 2014). I used novel projector technology with a 1440Hz refresh rate to apply frequency tagging at higher frequencies to circumvent this issue. Here, participants were presented with videos of an actress uttering action verbs (tagged at 61Hz amplitude modulation) accompanied by a gesture (tagged at 68 Hz flicker). Integration ease of speech and gestures was manipulated by auditory (clear/degraded speech) and visual factors (congruent/incongruent gesture). I demonstrated a stronger drive by the 61Hz amplitude modulation signal in right-temporal areas overlapping with auditory cortex when speech was degraded than when speech was clear. I speculate that this reflects that degraded speech enhances auditory attention to speech information. In occipital regions, I observed a stronger drive by the 68 Hz visual modulation signal when speech was clear than when speech was degraded. This might reflect enhanced attention

to gestures when speech is clear. Importantly, I demonstrated that RIFT can be used to identify the intermodulation frequency of the auditory tagged (61 Hz) and visually tagged (68Hz) signal (7Hz, as a result of f_2-f_1), which was strongest over LIFG and left-temporal regions, and strongest when speech was clear and a gesture matched the speech signal. I propose that this is a direct reflection of the non-linear integration of speech and gestures in these regions. Additionally, I propose that the strength of this intermodulation frequency reflects the ease of speech-gesture integration.

10.2. Discussion and implications

10.2.1. How does speech-gesture integration compare in externally and internally induced adverse listening conditions?

Taken together, this thesis reports the first set of studies that has uncovered the spatiotemporal neuronal dynamics and the network associated with semantic audiovisual integration in a multimodal context. The observed alpha and beta power modulations suggest a mechanistic role of brain oscillations in enabling the integration of speech and gestures and engaging a network of left-frontal (LIFG), left-temporal (pSTS/MTG, ATL), visual and motor regions in this process. This mechanism seems to be general and applicable to speech-gesture integration in both clear and adverse listening conditions. Importantly, I also observed a clear relationship between an individual's oscillatory modulations and the behavioral benefit a listener experienced during comprehension.

However, although similar core mechanisms between native and non-native listeners were observed, the distinct spatiotemporal time courses of the observed oscillatory modulations, as well as differences in eye-tracking patterns and behavioral responses in native compared to non-native listeners, might suggest that these two listener groups also employ different processing strategies and differentially weigh and integrate auditory and visual information during comprehension. The behavioral work in **chapter 2** and **chapter 5**, as well the EEG and eye-tracking work described in **chapter 6** and **chapter 8** also suggest distinct differences between adverse listening conditions that are solely challenged by

external factors, as compared to adverse listening conditions that are challenged by both internal and external factors. These differences, and their implications on understanding multimodal language in clear adverse listening conditions, will be reviewed and discussed in this section. First, the oscillatory results from **chapter 3, 4 and 7** will be discussed. Then, the behavioral, EEG and eye-tracking results from **chapters 2, 5, 6 and 8** will be discussed and all findings will be interpreted in a broader context.

In **chapter 3**, which focused on how oscillatory dynamics support speech-gesture integration in externally induced adverse listening conditions, I observed that native listeners immediately engaged rSTS when speech was degraded and a gesture was present, possibly to optimally process the word by increasing the uptake of gestural information. This was followed by engagement of the (hand area) of the motor cortex, reflecting simulation of the gesture to aid comprehension, and engagement of the LIFG and left-temporal regions, reflecting general unification and lexical access processes. Simultaneously, a gamma power increase over MTL reflected that the semantic information conveyed by gestures can facilitate a matching process with lexical memory traces that aids retrieval of the degraded input. Similar spatiotemporal results were obtained in **chapter 4**, where we used a violation paradigm to probe semantic integration.

In **chapter 7**, I observed similar oscillatory patterns supporting general unification and lexical access processes in native as in non-native listeners, as well as similar brain regions that were engaged during speech-gesture integration. However, the time course of how these different brain regions were engaged in integrating speech and gestures slightly differed between native and non-native listeners. This might suggest that the non-native listeners use a different strategy to weigh and integrate the auditory and visual inputs.

Non-native listeners demonstrated an immediate engagement of motor regions and LIFG at speech onset, suggesting that non-native listeners might focus immediately on gestures to aid comprehension. The strength of this engagement also correlated with a listener's observed gestural benefit in the subsequent 4-alternative forced choice identification task (as was observed in **chapter 3**). In tandem with these effects, I observed that in the same time window, an individual's

engagement of motor regions, AG, pSTS/MTG and STG did not correlate with the benefit a listener might experience during comprehension. This could confirm that these effects reflect a more low-level integration of the auditory input. In a subsequent time-window, the strength of engagement of left-temporal regions and visual regions *did* predict the benefit a listener experienced of a gesture during comprehension. A power-power correlation between the early effect in LIFG and motor regions and this subsequent left-temporal and visual effect suggested that listeners who immediately simulate gestures when speech is degraded and try to unify this with the degraded input, benefitted more from the gesture during retrieval of the degraded word. However, when integration is hindered, non-native listeners might engage additional resources in right-temporal regions to aid comprehension.

Putting all of these oscillatory results together, the core differences between the two groups thus seem to lie at the level of access to the phonological information in the degraded speech signal. The behavioral, EEG and eye-tracking results from **chapters 2, 5, 6 and 8** concur with this explanation. For native listeners, whose prior knowledge of the language is not diminished, it is easier to resolve the degraded auditory cues than for non-native listeners, who have less prior knowledge of the target language. This allows native listeners to optimally benefit from both visual articulators in a joint context, as was observed in the behavioral results of **chapter 2**. Similar results were obtained from the behavioral study described in **chapter 5**, where I observed that the benefit that non-native listeners experience from visible speech information is particularly smaller than the benefit that native listeners experience, especially when speech quality suffers. In line with this, **chapter 6** revealed that when speech is degraded, non-native listeners might be hindered in their ability to optimally integrate the semantic information that is conveyed by the gestures with the speech signal. Interestingly, these results are completely in line with the differences in oscillatory patterns that were observed between the two groups in **chapter 7**. Native listeners more strongly engaged the mouth region of the primary sensory cortex, suggesting that they find it easier to simulate and use the information conveyed by visible speech for comprehension. Second, native listeners more strongly engaged regions involved in semantic retrieval and

unification than non-native listeners, suggesting it was easier for native listeners to access this information and use it for comprehension. This allowed native listeners to thus optimally benefit from both visual articulators in a joint context.

The eye-tracking results from **chapter 8** however suggest that although non-native listeners look more at gestures than native listeners, both native and non-native listeners do not look more at gestures when speech is degraded than when speech is clear. This seems counter-intuitive when considering that the oscillatory patterns observed in **chapters 3, 4, and 7** suggested that both native and non-native listeners allocate more visual attention to gestures when speech is degraded. As described in previous work by Gullberg and colleagues (Gullberg & Holmqvist, 1999, 2002, 2006; Gullberg & Kita, 2009), I observed that listeners mostly gaze at the face during comprehension. Listeners thus seem to be able to extract semantic information without directly fixating on gestures, and even when eye gaze is not fixated on gestures, visual areas of the brain are more engaged.

Instead, the difference in gaze behavior between the two groups corroborates the results from **chapters 2-7**: non-native listeners seem more hindered by externally induced adverse listening conditions than native listeners (Bradlow & Alexander, 2007; Brouwer et al., 2012; Kilman et al., 2014; Mayo et al., 1997; Scharenborg et al., 2018). Specifically, non-native listeners seem to experience more difficulties when processing the degraded auditory cues and coupling them to the visible speech signals. Therefore, they might rely more strongly on visual semantic information from gestures during comprehension than native listeners, as was indicated by more fixations to gestures by non-native compared to native listeners, especially because their knowledge of the language is somewhat diminished (in line with Hazan et al., 2006). This diminished knowledge of the target language however limits them to benefit from both visual articulators in a joint context in a similar manner as native listeners do.

10.2.1.1. Do different types adverse listening conditions impact speech-gesture integration in a similar manner?

The next question is whether the different types of adverse listening conditions impact speech-gesture integration in a similar or different matter. Based on

the results from **chapters 2-8** it becomes evidently clear that in both types of adverse listening conditions, the semantic information conveyed by gestures aids comprehension, and that similar oscillatory dynamics support this process. However, I argue that how a gesture aids comprehension might depend on the type of adverse listening condition.

Let's first zoom into the observed effects of both types of adverse listening conditions in a bit more detail. As argued above, both internally and externally induced adverse listening conditions cause a listener to be less able to utilize the phonological information in the speech signal for comprehension. This effect seems to be additive when testing how externally and internally induced adverse listening interact in a joint context (as observed in **chapters 5, 6, 7 and 8**). I postulate that this additive effect causes a turning point in the weighing of auditory and visual information: when there are not enough reliable auditory cues in the signal anymore, listeners might be hindered in using the semantic information that is conveyed by the gesture to aid in comprehension.

However, on the basis of the behavioral results observed in **chapters 5, 6, 7 and 8**, I believe that externally induced adverse listening conditions might impact speech-gesture integration more strongly than internally induced listening conditions. For the effect on comprehension to be equal between the two types of adverse listening conditions, one would expect that the results for non-native listeners who integrate gestures in clear speech would be equal to the results of native listeners who integrate gestures in degraded speech. However, the effect on the ability to benefit from a gesture during comprehension seems to be larger in the latter case. This can be due to the fact that the non-native listeners who were tested in this thesis all were highly proficient language users. The benefit from gestures could for example be larger at a lower level of proficiency (Dahl & Ludvigsen, 2014; Sueyoshi & Hardison, 2005). This possibility is however difficult to test in the context of iconic gestures that are potentially ambiguous in the absence of speech. When future work would for example include less proficient listeners than the highly-proficient listeners that are used here, these listeners might not be able to recognize or understand the verbs. In that case, a bidirectional, mutual enhancement of speech and gestures cannot occur.

10.2.2. A comprehensive model of the neural integration of speech and gestures in clear and adverse listening conditions

The similarities and differences in oscillatory modulations that were observed in the two listener groups can inform current models of the neural integration of speech and gestures. In this section I aim to first discuss the observed differences in the neural integration of speech and gestures in both types of adverse listening conditions. Then, I will discuss how these results contribute to more general current debates on the role of LIFG and pSTS/MTG in the neural integration of speech and gestures.

Native and non-native listeners demonstrated different time courses of engagement of similar task-relevant brain regions (i.e., LIFG, pSTS/MTG, motor regions, visual regions) during gestural enhancement of degraded speech comprehension. As has become clear in the discussion about the differences between the two listener groups, non-native listeners might be more hindered in integrating the speech and gestures, especially when speech is degraded. For example, for the non-native listeners that were described in **chapter 7**, the results suggested that when speech is degraded and a gesture is present, non-native listeners immediately engage motor regions to simulate the gestural information and attempt to unify this information with the speech in LIFG to help retrieval of the degraded input. As described in **chapter 3**, native listeners however were able to already optimize their processing strategy right at speech onset (as reflected by the engagement of rSTS), and were therefore less hindered in integrating the two inputs. This might suggest that the speech-gesture integration process is thus more complex for non-native listeners, especially when speech is degraded.

I speculate that as soon the weighing of auditory and visual information shifts due to the additive load of internally and externally induced adverse listening conditions, non-native listeners' speech-gesture integration encompasses a more complex processing cycle that engages (multiple iterations of) a feedforward/feedback loop between LIFG and left-temporal regions (see for a visualisation, Figure 42).

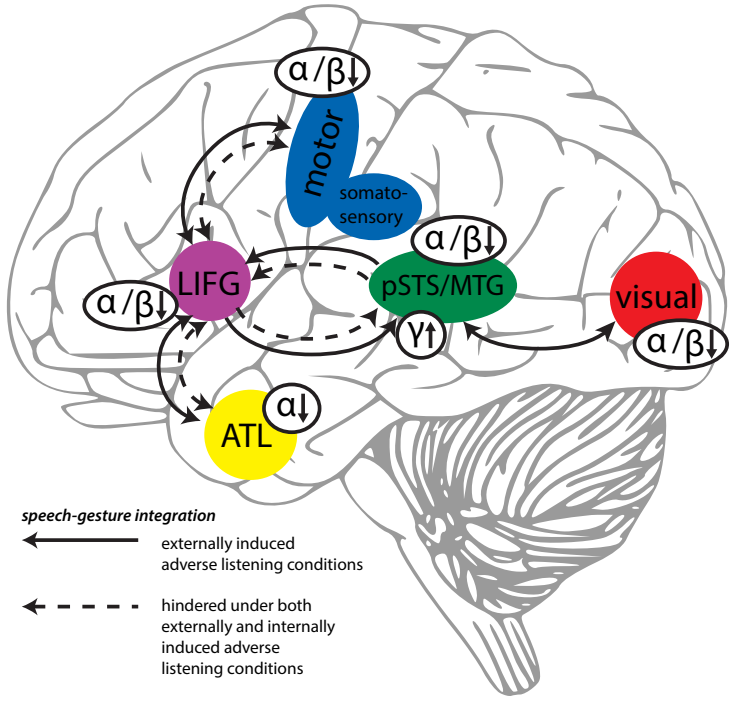


Figure 42. Schematic overview of the MEG results from chapters 3, 4, 7 and 9. Solid circles reflect results regarding oscillatory modulations in the alpha, beta and gamma band observed in chapters 3, 4 and 7. Solid arrows represent the processing cycle of speech-gesture integration under externally induced adverse listening conditions that involves visual regions, pSTS/MTG, motor regions (hand area), somatosensory regions (lower part that is sensitive to visible speech information), LIFG and anterior temporal regions such as ATL. Dashed arrows represent hindered processing cycle under both externally induced and internally induced listening conditions. The brain regions that connect to these arrows are less engaged in non-native listeners (challenged by internally induced adverse listening conditions) when understanding speech and gestures in externally induced adverse listening conditions. Note that all arrows have bidirectional arrowheads, as without measures of directed connectivity, it is unclear what the direction of these effects is.

Specifically, I propose that initially, like in native listeners, a lower level matching of audiovisual information takes place in pSTS/MTG, which is then forwarded to LIFG. Here, higher-level semantic information that a listener can extract while simulating a gesture by engaging the motor regions might be unified with the speech signal, as well as phonological information that is provided by visible speech. This information might then be fed back to left-temporal regions, where

lexical access/retrieval might be attempted while utilizing the unified information that is fed back from LIFG (in line with the MUC model, Hagoort, 2013; Lau, Phillips, & Poeppel, 2008). When integration is hindered, this processing cycle might include more iterations or a longer processing time, especially when the semantic information from the gesture or phonological information from visible speech is less available.

Note however that these thoughts are purely speculative, especially when considering directionality of these effects. This proposed processing cycle could thus only be tested by taking into account the direction of information flow during comprehension. This is something that is not tested in the current work. Future work could try to involve measures of directed connectivity, or might consider how rhythmic neuronal synchronization in different frequency bands might reflect the direction of information flow (following Schoffelen et al., 2017).

10.2.2.1. The role of LIFG and pSTS/MTG in the neural integration of speech and gestures

An important discussion in current literature on the neural integration of speech and gestures is on the involvement of LIFG and pSTS/MTG in this process. The results observed in **chapters 3, 4, and 7** have direct implications for this discussion. These implications will be discussed below.

In previous work, some studies have argued that LIFG is involved in the semantic unification of speech and iconic gestures and pSTS/MTG in the unification of speech and gestures who have a stable common object representation (e.g. pantomimes, Willems, Özyürek, & Hagoort, 2007, 2009). Others have argued that pSTS/MTG forms the primary integration region, and LIFG has more a modulatory or revising role in the process (Green et al., 2009; Holle et al., 2008, 2010). Finally, some studies have suggested that both LIFG and pSTS/MTG are involved in integrating gestures and speech (Dick et al., 2014).

The MEG experiments described in **chapters 3, 4 and 7** were the first studies that investigated the spatiotemporal neural dynamics underlying speech-gesture integration and the engagement of LIFG/pSTS/MTG over time. In all of these

studies, I observed engagement of the LIFG in all speech/gesture conditions compared to baseline. This rules out the possibility that LIFG would solely have a modulatory or revising role in speech-gesture integration (Green et al., 2009; Holle et al., 2008, 2010). Instead, I propose that the observed alpha/beta power decreases in LIFG are associated with the semantic unification of speech and gestures.

It should be noted however that the strength of engagement of LIFG is dependent on the context in which speech and gesture is integrated. For example, as was observed from **chapter 4**, this engagement was stronger when a gesture mismatched the speech signal, and similarly, was stronger for native listeners than for non-native listeners in **chapter 7**. This suggests that the LIFG possibly has a unifying function of the different inputs irrespective of congruency but that an increased integration load also increases engagement of the LIFG to unify the inputs. However, when integration is even more strongly hindered, such as when speech is degraded and a mismatching gesture is present (**chapter 4**), or when a non-native listener integrates a degraded speech signal with a gesture (**chapter 7**), this engagement might be less when the inputs cannot be resolved.

The results that were observed in pSTS/MTG throughout **chapters 3, 4 and 7** fit with this explanation. pSTS/MTG was demonstrated to be more engaged when speech was degraded and a gesture enhanced comprehension, possibly reflecting facilitated lexical access of the degraded input due to the gesture (**chapter 3**). This engagement was less when integration was hindered due to a mismatching gesture (**chapter 4**), and was similar for native and non-native listeners (**chapter 7**). This suggests that pSTS/MTG is sensitive to hindered audiovisual integration and lexical access processes when the visual semantic information cannot help to retrieve or disambiguate the degraded lexical item. In conclusion, I thus tentatively propose that LIFG and pSTS/MTG are both involved in the integration of speech and gestures and that engagement of LIFG can be associated with the semantic unification of speech and gestures, whereas engagement of pSTS/MTG can be associated with lower-level integration of the two inputs, and subsequent success of lexical access.

10.2.3. A general oscillatory mechanism supports the engagement of task-relevant brain regions during speech-gesture integration

In **chapters 3, 4 and 7** I have demonstrated that alpha/beta power decreases support the engagement of task-relevant brain regions during speech gesture integration in both external and internally induced adverse listening conditions.

Until now, such alpha/beta power decreases, which are thought to reflect engagement of task-relevant brain regions, and alpha/beta power increases, which are thought to reflect functional inhibition of task-irrelevant brain regions, were mostly observed in sensory regions (e.g., visual/motor regions, see Jensen & Mazaheri, 2010; Klimesch, Sauseng, & Hanslmayr, 2007; Payne & Sekuler, 2014; Pfurtscheller & Lopes da Silva, 1999; see for a few findings in language-related regions; Piai, Roelofs, Rommers, & Maris, 2015; Wang, Hagoort, & Jensen, 2018), and in response to non-semantic audiovisual integration (Hipp et al., 2011).

In the current thesis, I was able to prove that these oscillatory modulations are not specific to the sensory, visual or motor domain, but can also robustly be observed as part of a language-processing network in higher-order downstream areas. Moreover, I was able to demonstrate that these oscillatory modulations are predictive of audiovisual integration in a semantic context. This thus suggests a more general 'basic' neural mechanism of local engagement/inhibition, which is not per se specific to a certain cognitive domain, but which might be part of a 'set' of mechanisms that are relevant for cognitive information processing.

10.3. Future directions

10.3.1. Going beyond the word level: studying multimodal language comprehension in a larger context

The current thesis used word-level stimuli to study the semantic audiovisual integration of speech and gestures in an isolated context. Rightly so, recent work has also raised the question whether 'the data from these simple experiments are relevant for understanding everyday language processing' (Hasson, Egidi, Marelli, & Willems, 2018). Although the current interpretations might be indeed

limited to only the word-level, these results provide an important and necessary foundation to understand the interplay between these different types of adverse listening conditions, and how these adverse listening conditions impact semantic audiovisual integration in a joint context.

This interplay becomes extremely relevant when we consider that in daily face-to-face communication, the context in which speech and gestures occur in is part of an even richer environment, in which gestures are made at the sentence or discourse level, and in dialogue or interaction. It is for example highly imaginable that the different cues that are conveyed by both visible speech and gestures influence comprehension in a different manner depending on how a certain sentence is unfolding, or depending on the context of a conversation they occur in. For example, at some point phonological information that is conveyed by visible speech movements might be more relevant and informative for comprehension than the semantic information that is conveyed by gestural information, or the other way around. How these different inputs influence comprehension over time in a sentence-, interactional, or discourse-level context remains poorly understood. The current work however provides an important step towards more ecologically valid stimuli, in which multimodal inputs and adverse listening conditions are the norm, and not the exception. Studying similar questions in a sentence-, interactional or discourse-level context can help understand as well as test whether the mechanisms observed at this word-level extend to these richer contexts during language comprehension. To achieve this, future work could study how a listener allocates attention to different kinds visual input (i.e., lip information, or meaningful gestural information), while watching social interactions, or while communicating with a (virtual) conversational partner, as well as how this differs under adverse listening conditions. As the current thesis demonstrated that visual cues can aid comprehension, future work could investigate whether this disambiguation occurs because visible speech information or gestural information improves predictions about upcoming words. For example, studies could use gating paradigms to see whether gestural information can help to predict upcoming words, and could investigate whether similar neural mechanisms are observed in such a context as at the word-level.

This would also be highly relevant in the light of current word recognition models (e.g., Shortlist B, Norris & McQueen, 2008), as it could inform us about how gestures could sharpen predictions of multiple lexical hypotheses that are evaluated in parallel.

Moving from the word-level to the sentence-, interactional or discourse-level is, however, in my opinion not the only question that deserves attention when expanding the contextual horizon. One must consider that different *types* of context can also influence semantic audiovisual integration in different manners, or at different levels. For example, a listener can consider linguistic types of context, such as varieties of prosody, syntax, semantics, phonology or pragmatics during audiovisual integration. These types of context seem closer related to the speech signal than for example visual contexts, such as visible speech, gestures or facial expressions. How these different types of context interact or influence each other during comprehension remains unknown. This however seems an important avenue for future research, as it might inform us on how the mechanisms that are thought to underlie multimodal language comprehension manifest itself depending on the context that is available.

10.3.2. Using neuronal oscillations to study multimodal language comprehension

Using neuronal oscillations as a tool to study the influences of such contexts during language comprehension might be especially fruitful when one wants to investigate how a certain network might dynamically adapt to incorporate influences from the current focus of contextual attention. By studying power, coherence, and cross-frequency coupling, some oscillatory patterns, as well as the potential (hierarchical) relations between frequency bands, have been related to higher-level language operations in a unimodal, speech context (see Meyer, 2017). The question however remains whether the oscillatory mechanisms that are proposed to underlie (multimodal) speech processing are ubiquitous throughout the brain, and also apply to other types of (non-linguistic) stimuli. Moreover, to understand whether and how these mechanisms underlie multimodal language comprehension, I think it is important to also investigate the relationship between

an individual's observed oscillatory modulations and their behavioral responses. As was demonstrated in **chapter 7**, these relations can be crucial for understanding processing mechanisms.

Although I observed clear brain-behavior relationships between an individual's oscillatory modulations and their behavioral measures, a debate still exists on whether neuronal oscillations are an epiphenomenon or causally important. This also applies to the domain of (multimodal) language comprehension. Future work could address this question by using brain stimulation measures to intervene in language-related operations, or by using RIFT to entrain neuronal oscillations that are thought to be crucial for language-related operations. Moreover, single unit intracranial recordings in epileptic patients could be informative on how neuronal oscillations serve to coordinate the neuronal firing supporting multimodal language comprehension.

Importantly, studying oscillatory modulations allows us to move towards a network-based understanding of multimodal language comprehension. As mentioned earlier in this Discussion, using directed connectivity measures (e.g., Granger causality analysis) could prove to be useful to identify how different regions that are involved in multimodal language comprehension interact with each other within this context, and to consider how rhythmic neuronal synchronization in different frequency bands might reflect the direction of information flow during integration or comprehension.

Related to this, I think it is important to point out that I doubt that solely identifying *where* in the brain the integration of speech and gestures takes place is the most interesting question to ask. Instead, I believe we should ask *how* the brain performs this speech-gesture integration in both space and time, and, most importantly, how this network of regions that is involved in this process dynamically adapts to different types of contexts in which integration could occur (be it clear or degraded speech, or as a native or non-native listener). Discovering how the brain supports the cooperation of different brain regions or networks that are involved in this process, as well as how the weight of these cooperating networks might differ depending on the context in which multimodal language occurs in, might allow to define a basic 'set' of mechanisms that can possibly also

be observed in other domains. Some future directions on how to investigate these ideas are laid out below.

10.3.3. Using rapid invisible frequency tagging as a tool to study signatures of non-linear (audiovisual) integration

Rapid invisible frequency tagging (RIFT) is a promising tool to study the non-linear interactions between auditory and visual inputs in the brain. In **chapter 9** I used RIFT to generate steady-state evoked responses at high frequencies (> 60 Hz) as a tool to investigate the integration of audiovisual information in a semantic context. By tagging the auditory speech signal at 61 Hz, and the visual semantic signal at 68 Hz, I was able to identify power at an intermodulation frequency at 7Hz as a result of a non-linear integration of the two inputs. This response was enhanced when integration was easiest (i.e., when speech was clear and a gesture was present).

I believe that the use of rapid invisible frequency tagging opens up exciting opportunities for future studies to investigate the time course of selective attention distribution during (semantic) audiovisual information processing. This could be directly applicable to the experimental ideas laid out in section 10.3.1 and 10.3.2. The use of this rapid invisible tagging technique does not make participants aware of the goal of an experiment, and could allow tracking attention to different inputs (i.e., visible speech information or gestural information) over time. This makes the use of this technique especially relevant for the study of multimodal language comprehension in interactional settings, as well as during comprehension of sentence-level input. Future studies could for example consider incorporating tagging visible speech as well as gestural information to investigate how listeners divide attention to different inputs over time.

However, studying the time course of certain effects might also be challenged by the use of complex, dynamic stimuli. In **chapter 9**, I presented the first study that used complex, dynamic stimuli and high frequency-tagging to study interactions between auditory and visual inputs. Here, we demonstrated a clear signature of non-linear audiovisual integration when studying the *power* of the intermodulation frequency, but when studying the *phase coherence* in the signal,

we did not see a similar signature of non-linear audiovisual integration. Currently, it is unclear whether this is because of the complex nature of the stimuli that are used, or whether this is informative about the actual mechanism underlying the integration of the auditory and visually tagged stimuli. For example, we could speculate that this might show that the integration of auditory and visual information takes place at different points in time during comprehension, which indicates that the phase of the intermodulation signal is not consistent over time. This could imply that words in degraded speech might take longer to be recognized than words in clear speech. Alternatively, gestural input could facilitate speed up comprehension when it is congruent with the speech signal, and possibly reduce lexical competition when speech is degraded. However, at this point many explanations for the lack of a phase coherence effect in **chapter 9** could be possible. For example, the phase consistency per condition, or the angle of the phase coherence could depend on the strength of integration.

Other exciting avenues for using RIFT as a tool to study multimodal integration lie in investigating the role of low-frequency oscillations in sensory processing. Current studies using frequency tagging often use tagging frequencies that are in the range of lower frequencies (e.g., alpha), which is problematic because this is likely to entrain spontaneous neural oscillations as well (Keitel et al., 2014; Spaak et al., 2014). Future analyses or studies could use RIFT to investigate whether the integration and interaction of audiovisual information might be established by low-frequency phase synchronization between regions. This, for example, could serve as a temporal reference frame for high-frequency activity (Bonnefond, Kastner, & Jensen, 2017). Second, the amplitude of the frequency-tagged signal in downstream areas might be modulated by the phase of low-frequency activity. This could reflect the increase in information that is transferred from primary sensory areas to downstream areas, depending on the strength or success of integration. Third, RIFT may be used to manipulate high-frequency oscillations during language comprehension, for example by entrainment of gamma band oscillations. Finally, studies could investigate whether the phase of low-frequency oscillations is predictive of integration ease. It is imaginable that a specific phase angle might reflect an optimal excitability window where integration could

maximally occur. Using RIFT might thus be the ideal technique to study how a listener allocates attention to different inputs over time during interaction (see section 10.3.1).

10.4. Limitations of the current work

Although some of the limitations of the current work have been laid out above in their respective discussions, some additional limitations of the current work have remained implicit. First, it should be pointed out that all studies in this thesis made use of a variation of the same paradigm, where participants watched a video clip and were then asked to recall the word in the video. Although **chapter 2** and **chapter 5** used an open-set identification task, the 4-alternative forced choice identification task used in the other chapters might have masked the actual comprehension participants experienced during the video. Luckily, the reaction times of these responses are not affected by this. Future work could experiment with a similar paradigm, but in which a vocal response is given after the videos.

Second, the benefit of having used a similar behavioral task in all studies in this thesis is that the gestural enhancement effect could be replicated in every study. However, **chapter 8** suggested that the gestural enhancement effect might be larger for native than for non-native listeners. Although this result feels intuitively plausible, we did not observe a difference between the two groups on the behavioral tasks in **chapter 5** and **chapter 7**. Several reasons could explain this difference. In **chapter 5**, we for example did not measure reaction times and solely found similar effects as in **chapter 8** that pertained to the accuracy scores. Moreover, in **chapter 7**, we collected mismatching gestures as well as part of a larger paradigm, which might alter the task strategy for both groups during comprehension. Future replication of the difference in gestural enhancement as described in **chapter 8** is thus needed to determine whether this difference between the gestural enhancement effect between natives and non-natives persists.

Third, another limitation of the current work could be that we only used highly-proficient non-native listeners, and not an additional group of lower proficiency. Including an extra group could have allowed us to determine whether the observed effects in internally induced adverse listening conditions would be

stronger dependent on language proficiency (e.g., Hazan et al., 2006).

10.5. Conclusion

Integrating audiovisual cues in a semantic context can help to reduce uncertainty during multimodal language comprehension. In this thesis I have demonstrated that semantic visual information conveyed by gestures enhances language comprehension in an externally induced listening condition, such as in a crowded cafe, and in internally induced listening conditions, such as when you are a non-native listener. The brain might mechanistically achieve this by suppressing oscillatory alpha and beta power to engage the extended language network, visual and motor regions in this process. This oscillatory mechanism seems to be general, and engaging these brain regions supports general unification, simulation and lexical access processes which aid comprehension when speech is degraded, and when you are a non-native listener. However, distinct spatiotemporal time courses of the engagement of these regions as well as differential eye gaze patterns suggested that a listener's processing strategies might differ depending on the context in which a listener understands multimodal language. Non-native listeners for example might have less access to the phonological information in the degraded speech signal, and might experience more difficulty unifying the semantic information conveyed by gestures with the speech signal. This causes areas that are involved in unification, lexical access and simulation to be less engaged during comprehension. This means that there is an additive effect on integration when both types of adverse listening conditions occur. Finally, I provided evidence that RIFT can be used to study the integration of audiovisual information in a semantic context. This opens up exciting opportunities to investigate how listeners distribute their attention to different visual articulators to aid comprehension over time, as well as to study the role of low-frequency oscillations in such processes.

REFERENCES

Abdullaev, Y. G., & Melnichuk, K. V. (1997). Cognitive operations in the human caudate nucleus. *Neuroscience Letters*, *234*(2–3), 151–155. https://doi.org/10.1016/S0304-3940(97)00680-0

Adank, P., & Devlin, J. T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. *NeuroImage*, *49*(1), 1124–1132. https://doi.org/10.1016/j.neuroimage.2009.07.032

Agrafiotis, D., Canagarajah, N., Bull, D. R., & Dye, M. (2003). Perceptually optimized sign language video coding based on eye tracking analysis. *Electronics Letters*, *39*(24), 1703–1705. https://doi.org/10.1049/el

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. (2018). Raincloud plots: a multi-platform tool for robust data visualization. *PeerJ Preprints*. https://doi.org/10.7287/peerj.preprints.27137v1

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Oxford, England: Cambridge U Press.

Argyle, M., & Graham, J. A. (1976). The central Europe experiment: Looking at persons and looking at objects. *Environmental Psychology and Nonverbal Behavior*, *1*(1), 6–16. https://doi.org/10.1007/BF01115461

Aydelott, J., Dick, F., & Mills, D. L. (2006). Effects of acoustic distortion and semantic context on event-related potentials to spoken words. *Psychophysiology*, *43*(5), 454–464. https://doi.org/10.1111/j.1469-8986.2006.00448.x

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, *26*(9), 1338–1367. https://doi.org/10.1080/01690965.2010.542671

Bamiou, D. E., Musiek, F. E., & Luxon, L. M. (2003). The insula (Island of Reil) and its role in auditory processing: Literature review. *Brain Research Reviews*, *42*(2), 143–154. https://doi.org/10.1016/S0165-0173(03)00172-3

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(August), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bastiaansen, M. C. M., & Knösche, T. R. (2000). Tangential derivative mapping of axial MEG applied to event-related desynchronization research. *Clinical Neurophysiology*, *111*(7), 1300–1305. https://doi.org/10.1016/S1388-2457(00)00272-8

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica*, *1–2*(123), 1–30. Retrieved from http://philpapers.org/rec/BEADIH

Beattie, G., & Shovelton, H. (1999). Mapping the Range of Information Contained in the Iconic Ha nd Gestures that Accompany Spontaneous Speech. *Journal of Language and Social Psychology*, *18*(4), 438–462. https://doi.org/10.1177/0261927X99018004005

Beattie, G., & Shovelton, H. (2002). An experimental investigation of some properties of individual iconic gestures that mediate their communicative power. *British Journal*

of Psychology, *93*(2), 179–192. https://doi.org/10.1348/000712602162526

Beauchamp, M. S. (2005). See me, hear me, touch me: Multisensory integration in lateral occipital-temporal cortex. *Current Opinion in Neurobiology*, *15*(2), 145–153. https://doi.org/10.1016/j.conb.2005.03.011

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. https://doi.org/10.1038/nn1333

Beauchamp, M. S., Lee, K., Argall, B., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*, 809–823. https://doi.org/10.1016/S0896-6273(04)00070-4

Becker, R., Pefkou, M., Michel, C. M., & Hervais-Adelman, A. G. (2013). Left temporal alpha-band activity reflects single word intelligibility. *Frontiers in Systems Neuroscience*, *7*(December), 121. https://doi.org/10.3389/fnsys.2013.00121

Bell, A. J., & Sejnowski, T. J. (1995). An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, *7*(6), 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, *124*(2), 143–152. https://doi.org/10.1016/j.bandl.2012.10.008

Biau, E., & Soto-Faraco, S. (2015). Synchronization by the hand: the sight of gestures modulates low-frequency activity in brain responses to continuous speech. *Frontiers in Human Neuroscience*, *9*(September), 527. https://doi.org/10.3389/fnhum.2015.00527

Biau, E., Torralba, M., Fuentemilla, L., de Diego Balaguer, R., & Soto-Faraco, S. (2015). Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. *Cortex*, *68*, 76–85. https://doi.org/10.1016/j.cortex.2014.11.018

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer.

Bonnefond, M., Kastner, S., & Jensen, O. (2017). Communication between Brain Areas Based on Nested Oscillations Brain communication based on nested oscillations. *ENeuro*. https://doi.org/10.1523/ENEURO.0153-16.2017

Booth, J. R., Wood, L., Lu, D., Houk, J. C., & Bitan, T. (2007). The role of the basal ganglia and cerebellum in language processing. *Brain Research*, *1133*(1), 136–144. https://doi.org/10.1016/j.brainres.2006.11.074

Boulenger, V., Hoen, M., Jacquier, C., & Meunier, F. (2011). Interplay between acoustic/phonetic and semantic processes during spoken sentence comprehension: An ERP study. *Brain*

and Language, *116*(2), 51–63. https://doi.org/10.1016/j.bandl.2010.09.011

Bradlow, A. R., & Alexander, J. a. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, *121*(4), 2339–2349. https://doi.org/10.1121/1.2642103

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, *112*(1), 272–284. https://doi.org/10.1121/1.1487837

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, (10), 433–436.

Broersma, M., & Cutler, A. (2011). Competition dynamics of second-language listening. *Quarterly Journal of Experimental Psychology*, *64*(1), 74–95. https://doi.org/10.1080/17470218.2010.499174

Brouwer, S., & Bradlow, A. R. (2016). The Temporal Dynamics of Spoken Word Recognition in Adverse Listening Conditions. *Journal of Psycholinguistic Research*, *45*(5), 1151–1160. https://doi.org/10.1007/s10936-015-9396-9

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America*, *131*(2), 1449–1464. https://doi.org/10.1121/1.3675943

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, *2*(1), 1–13. https://doi.org/10.1080/17470910601043644

Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., & Hallett, M. (2002). Neural correlates of cross-modal binding. *Nature Neuroscience*, *6*, 190. Retrieved from http://dx.doi.org/10.1038/nn993

Butterworth, B., & Beattie, G. (1978). Gesture and Silence as Indicators of Planning in Speech. In R. Campbell & P. T. Smith (Eds.), *Recent Advances in the Psychology of Language: Formal and Experimental approaches*. New York: Plenum. https://doi.org/10.1007/978-1-4684-2532-1

Buzsáki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents — EEG , ECoG , LFP and spikes. *Nature Reviews Neuroscience*, *13*(June), 407–420. https://doi.org/10.1038/nrn3241

Caetano, G., Jousmaki, V., & Hari, R. (2007). Actor's and observer's primary motor cortices stabilize similarly after seen or heard motor actions. *PNAS*, *104*(21), 9058–9062.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, *14*(17), 2213–2218. https://doi.org/10.1097/01.wnr.0000095492.38740.8f

Callan, D. E., Jones, J. a, Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet

filtered visual speech gesture information. *Journal of Cognitive Neuroscience*, *16*(5), 805–816. https://doi.org/10.1162/089892904970771

Calvert, G. A. (2001). Crossmodal Processing in the Human Brain : Insights from Functional Neuroimaging Studies. *Cerebral Cortex*, *11*, 1110–1123.

Carter, C., Macdonald, A., Botvinick, M., Ross, L. L., Stenger, V., Noll, D., & Cohen, J. D. (2000). Parsing executive processes: strategic vs. evaluative functions of the anterior cingulate cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(4), 1944–1948. https://doi.org/10.1073/pnas.97.4.1944

Clark, H. H. (1996). *Using Language*. Cambridge University Press. Retrieved from https://books.google.com/books?id=DiWBGOPYnoC&pgis=1

Connolly, J. F., Philips, N. A., Stewart, S. H., & Brake, W. G. (1992). Event-Retated Potential Sensitivity to Acoustic and Semantic Properties of Terminal Words in Sentences. *Brain and Language*, *18*(43), 1–18.

Cornejo, C., Simonetti, F., Ibanez, A., Aldunate, N., Ceric, F., Lopez, V., & Nunez, R. E. (2009). Gesture and metaphor comprehension: Electrophysiological evidence of cross-modal coordination by audiovisual stimulation. *Brain and Cognition*, *70*(1), 42–52. https://doi.org/10.1016/j.bandc.2008.12.005

Cutler, A., Garcia Lecumberri, M. L., & Cooke, M. (2008). Consonant identification in noise by native and non-native listeners: Effects of local context. *The Journal of the Acoustical Society of America*, *124*(2), 1264–1268. https://doi.org/10.1121/1.2946707

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, *116*(6), 3668–3678. https://doi.org/10.1121/1.1810292

Dahl, T. I., & Ludvigsen, S. (2014). How I See What You're Saying: The Role of Gestures in Native and Foreign Language Listening Comprehension. *The Modern Language Journal*, *98*(3), 813–833. https://doi.org/10.1111/j.1540-4781.2014.12124.x

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical Processing in Spoken Language Comprehension. *The Journal of Neuroscience*, *23*(8), 3423–3431.

Dick, A. S., Goldin-Meadow, S., Solodkin, A., & Small, S. L. (2012). Gestures in the developing brain. *Dev. Sci.*, *15*(2), 165–180. https://doi.org/10.1111/j.0013-9580.2004.458002.x

Dick, A. S., Mok, E. H., Raja Beharelle, A., 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. https://doi.org/10.1002/hbm.22222

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. https://doi.org/10.1162/jocn_a_00963

Drijvers, L., Mulder, K., & Ernestus, M. (2016). Alpha and gamma band oscillations index differential processing of acoustically reduced and full forms. *Brain and Language*, *153–154*, 27–37. https://doi.org/10.1016/j.bandl.2016.01.003

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). https://doi.org/10.1044/2016_JSLHR-H-16-0101

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*(January 2017), 7–17. https://doi.org/10.1016/j.bandl.2018.01.003

Drijvers, L., Özyürek, A., & Jensen, O. (2018). Alpha and Beta Oscillations Index Semantic Congruency between Speech and Gestures in Clear and Degraded Speech. *Journal of Cognitive Neuroscience*, *30*(8), 1086–1097. https://doi.org/10.1162/jocn_a_01301

Drijvers, L., Özyürek, A., & Jensen, O. (2018). Hearing and seeing meaning in noise: Alpha, beta, and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human Brain Mapping*. https://doi.org/10.1002/hbm.23987

Drijvers, L., & Trujillo, J. P. (2018). Commentary: Transcranial Magnetic Stimulation over Left Inferior Frontal and Posterior Temporal Cortex Disrupts Gesture-Speech Integration. *Frontiers in Human Neuroscience*, *12*(256). https://doi.org/10.3389/fnhum.2018.00256

Eckert, M. a, Menon, V., Walczak, A., Ahlstrom, J., Denslow, S., & Dubno, J. R. (2009). At the Heart of the Ventral Attention System: the Right Anterior Insula. *Hum Brain Mapp.*, *30*(8), 2530–2541. https://doi.org/10.1002/hbm.20688.At

Eisner, F., McGettigan, C., Faulkner, A., Rosen, S., & Scott, S. K. (2010). Inferior frontal gyrus activation predicts individual differences in perceptual learning of cochlear-implant simulations. *Journal of Neuroscience*, *30*(21), 7179–7186. https://doi.org/10.1523/JNEUROSCI.4040-09.2010.Inferior

Emmorey, K., Thompson, R., & Colvin, R. (2009). Eye gaze during comprehension of American sign language by native and beginning signers. *Journal of Deaf Studies and Deaf Education*, *14*(2), 237–243. https://doi.org/10.1093/deafed/enn037

Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2013). The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *Journal of Neuroscience*, *33*(26), 10688–10697. https://doi.org/10.1523/JNEUROSCI.4596-12.2013

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423–425. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5808871

Erber, N. P. (1971). Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by

children with impaired hearing. *Journal of Speech and Hearing Research*, *14*(3), 496–512. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5163883

Erber, N. P. (1975). Auditory-visual perception of speech. *J Speech Hear Disord*, *40*(4), 481–492.

Ernestus, M., Dikmans, M., & Giezenaar, G. (2017). Advanced second language learners experience difficulties processing reduced word pronunciation variants. *Dutch Journal of Applied Linguistics*, *6*(1), 1–31.

Federmeier, K. D., & Kutas, M. (2001). Meaning and Modality : Influences of Context, Semantic Memory Organization, and Perceptual Predictability on Picture Processing. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *27*(1), 202–224. https://doi.org/10.1037//027

Federmeier, K. D., & Kutas, M. (2002). Picture the difference: Electrophysiological investigations of picture processing in the two cerebral hemispheres. *Neuropsychologia*, *40*(7), 730–747. https://doi.org/10.1074/jbc.M403429200

Flege, J. E. (1992). The intelligibility of English vowels spoken by British and Dutch talkers. *Intelligibility in Speech Disorders: Theory, Measurement, and Management*, 157–232. Retrieved from http://jimflege.com/files/Flege_in_Kent_1992.pdf

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, *9*(10), 474–480. https://doi.org/10.1016/j.tics.2005.08.011

Fries, P., Nikolić, D., & Singer, W. (2007). The gamma cycle. *Trends in Neurosciences*, *30*(7), 309–316. https://doi.org/10.1016/j.tins.2007.05.005

Galambos, R., Makei, S., & Talmachoff, P. J. (1981). A 40-Hz auditory potential recorded from the human scalp. *Proceedings of the National Academy of Sciences*, *78*(4), 2643–2647.

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. https://doi.org/10.1016/j.neuroimage.2004.06.004

Gat, I. B., & Keith, R. W. (1978). An Effect of Linguistic Experience: Auditory Word Discrimination by Native and Non-Native Speakers of English. *Audiology*, *17*, 339–345.

Giani, A. S., Ortiz, E., Belardinelli, P., Kleiner, M., Preissl, H., & Noppeney, U. (2012). Steady-state responses in MEG demonstrate information integration within but not across the auditory and visual senses. *NeuroImage*, *60*(2), 1478–1489. https://doi.org/10.1016/j.neuroimage.2012.01.114

Giard, M. H., & Peronnet, F. (1997). Auditory-Visual Integration during Multimodal Object Recognition in Humans : A Behavioral and Electrophysiological Study. *Journal of Cogntive Neuroscience*, *11*(5), 473–490.

Goldin-Meadow, S. (2005). *Hearing Gesture: How Our Hands Help Us Think*. Retrieved from https://books.google.com/books?hl=nl&lr=&id=LCJ5eQdsolsC&pgis=1

Golestani, N., Rosen, S., & Scott, S. K. (2009).

Native-language benefit for understanding speech-in-noise: The contribution of semantics. *Bilingualism (Cambridge, England)*, *12*(3), 385–392. https://doi.org/10.1017/S1366728909990150

Graham, J. A., & Argyle, M. (1975). International Journal of A Cross-Cultural Study of the Communication of Extra- Verbal Meaning by Gesture A cross-cultural study of the communication. *International Journal of Psychology*, *10*(1), 57–67.

Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for auditory-visual consonant recognition. *The Journal of the Acoustical Society of America*, *100*(4 Pt 1), 2415–2424. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8865647

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. https://doi.org/10.1002/hbm.20753

Gross, J., Kujala, J., Hamalainen, M., Timmermann, L., Schnitzler, A., & Salmelin, R. (2001). Dynamic imaging of coherent sources: Studying neural interactions in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 694–699. https://doi.org/10.1073/pnas.98.2.694

Gulbinaite, R., van Viegen, T., Wieling, M., Cohen, M. X., & VanRullen, R. (2017). Individual alpha peak frequency predicts 10 Hz flicker effects on selective attention. *The Journal of Neuroscience*, 1163–17. https://doi.org/10.1523/JNEUROSCI.1163-17.2017

Gullberg, M., & Holmqvist, K. (1999). Keeping an eye on gestures: Visual perception of gestures in face-to-face communication. *Pragmatics & Cognition*, *7*(1), 35–63. https://doi.org/10.1075/pc.7.1.04gul

Gullberg, M., & Holmqvist, K. (2002). Visual Attention towards Gestures in Face-to-Face Interaction vs . on Screen *. *International Gesture Workshop*, 206–214.

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. https://doi.org/10.1075/pc.14.1.05gul

Gullberg, M., & Kita, S. (2009). Attention to Speech-Accompanying Gestures: Eye Movements and Information Uptake. *Journal of Nonverbal Behavior*, *33*(4), 251–277. https://doi.org/10.1007/s10919-009-0073-2

Habets, B., Kita, S., Shao, Z., Ozyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*(8), 1845–1854. https://doi.org/10.1162/jocn.2010.21462

Hagoort, P. (2013). *MUC (Memory, Unification, Control) and beyond. Frontiers in Psychology* (Vol. 4). Elsevier Inc. https://doi.org/10.3389/fpsyg.2013.00416

Hagoort, P., & Van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 801–811. https://doi.org/10.1098/

rstb.2007.2089

Hannah, B., Wang, Y., Jongman, A., Sereno, J. A., Cao, J., & Nie, Y. (2017). Cross-modal association between auditory and visuospatial information in Mandarin tone perception in noise by native and non-native perceivers. *Frontiers in Psychology*, *8*(DEC), 1–15. https://doi.org/10.3389/fpsyg.2017.02051

Hannemann, R., Obleser, J., & Eulitz, C. (2007). Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain Research*, *1153*, 134–143. https://doi.org/10.1016/j.brainres.2007.03.069

Hardison, D. M. (2010). Visual and auditory input in second-language speech processing. *Language Teaching*, *43*(December 2009), 84. https://doi.org/10.1017/S0261444809990176

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*(June 2017), 135–157. https://doi.org/10.1016/j.cognition.2018.06.018

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non- native consonant contrasts. *The Journal of the Acoustical Society of America*, *119*(3), 1740–1751. https://doi.org/10.1121/1.2166611

He, Y., Gebhardt, H., Steines, M., Sammer, G., Kircher, T., Nagels, A., & Straube, B. (2015). The EEG and fMRI signatures of neural integration: An investigation of meaningful gestures and corresponding speech. *Neuropsychologia*, *72*, 27–42. https://doi.org/10.1016/j.neuropsychologia.2015.04.018

He, Y., Gebhardt, H., Str, L., Rondinone, I., & Straube, B. (2011). The Missing Power: Language Mediates Sensorimotor-related Beta Oscillations during On-line Comprehension of Different Types of Co-speech Gesture. *Unpublished*.

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*, *0*(0), 1–17. https://doi.org/10.1007/s00429-018-1674-5

Herring, J. D. (2017). *Driving visual cortex to study neuronal oscillations*. Doctoral thesis.

Herrmann, C. S. (2001). Human EEG responses to 1-100 Hz flicker: Resonance phenomena in visual cortex and their potential correlation to cognitive phenomena. *Experimental Brain Research*, *137*(3–4), 346–353. https://doi.org/10.1007/s002210100682

Herrmann, C. S., Munk, M. H. J., & Engel, A. K. (2004). Cognitive functions of gamma-band activity: Memory match and utilization. *Trends in Cognitive Sciences*, *8*(8), 347–355. https://doi.org/10.1016/j.tics.2004.06.006

Hervais-Adelman, A. G., Carlyon, R. P., Johnsrude, I. S., & Davis, M. H. (2012). Brain regions recruited for the effortful comprehension of noise-vocoded words. *Language and Cognitive Processes*, *27*(7–8), 1145–1166. https://doi.org/10.1080/01690965.2012.662280

Higby, E., Kim, J., & Obler, L. K. (2013). Multilingualism and the brain. *Annual Review of Applied Linguistics*, *33*(August 2014), 68–101. https://doi.org/10.1017/S0267190513000081

Hipp, J. F., Engel, A. K., & Siegel, M. (2011). Oscillatory synchronization in large-scale cortical networks predicts perception. *Neuron*, *69*(2), 387–396. https://doi.org/10.1016/j.neuron.2010.12.027

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*(April), 298–310.

Hirata, Y., Kelly, S., Huang, J., & Manansala, M. (2014). Effects of Hand Gestures on Auditory Learning of Second-Language Vowel Length Contrasts. *Journal for Speech, Language & Hearing Research*, *57*(December), 2090–2101. Retrieved from http://www.colgate.edu/docs/default-source/default-document-library/effects-of-hand-gestures-on-auditory-learning-of-second-language-vowel-length-contrasts.pdf?sfvrsn=0

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

Holle, H., Gunter, T. C., Ruschemeyer, S. A., Hennenlotter, A., & Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *NeuroImage*, *39*(4), 2010–2024. https://doi.org/10.1016/j.neuroimage.2007.10.055

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*(March), 74. https://doi.org/10.3389/fpsyg.2012.00074

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. https://doi.org/10.1016/j.neuroimage.2009.08.058

Holler, J., Kelly, S., Hagoort, P., & Ozyurek, A. (2010). When gestures catch the eye : The influence of gaze direction on co-speech gesture comprehension in triadic communication, 467–472.

Holler, J., Kokal, I., Toni, I., Hagoort, P., Kelly, S. D., & Ozyürek, A. (2014). Eye'm talking to you: speakers' gaze direction modulates co-speech gesture processing in the right MTG. *Social Cognitive and Affective Neuroscience*, 1–7. https://doi.org/10.1093/scan/nsu047

Holler, J., Schubotz, L., Kelly, S., Hagoort, P., Schuetze, M., & Özyürek, A. (2014). Social eye gaze modulates processing of speech and co-speech gesture. *Cognition*, *133*(3), 692–697. https://doi.org/10.1016/j.cognition.2014.08.008

Holler, J., Shovelton, H., & Beattie, G. (2009). Do Iconic Hand Gestures Really Contribute to the Communication of Semantic Information in a Face-to-Face Context? *Journal of Nonverbal Behavior*, *33*(2), 73–88. https://doi.org/10.1007/s10919-008-0063-9

Hoskin, J., & Herman, R. (2001). The

communication, speech and gesture of a group of hearing-impaired children. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, *36 Suppl*, 206–209. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11340783

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

Ibanez, 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. https://doi.org/10.1016/j.neulet.2010.01.009

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57. https://doi.org/10.1016/S0010-0277(02)00198-1

Jensen, O., Kaiser, J., & Lachaux, J.-P. (2007). Human gamma-frequency oscillations associated with attention and memory. *Trends in Neurosciences*, *30*(7), 317–324. https://doi.org/10.1016/j.tins.2007.05.001

Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Frontiers in Human Neuroscience*, *4*(November), 186. https://doi.org/10.3389/fnhum.2010.00186

Jokisch, D., & Jensen, O. (2007). Modulation of gamma and alpha activity during a working memory task engaging the dorsal or ventral stream. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(12), 3244–3251. https://doi.org/10.1523/JNEUROSCI.5399-06.2007

Jongman, A., Wang, Y., & Kim, B. H. (2003). Contributions of Semantic and Facial Information to Perception of Nonsibilant Fricatives. *Journal of Speech Language and Hearing Research*, *46*(6), 1367. https://doi.org/10.1044/1092-4388(2003/106)

Josse, G., Joseph, S., Bertasi, E., & Giraud, A. L. (2012). The brain's dorsal route for speech represents word meaning: evidence from gesture. PLoS One, 7(9), e46108.

Jung, T.-P. P., Makeig, S., Humphries, C., Lee, T.-W. W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*(2), 163–178. https://doi.org/10.1111/1469-8986.3720163

Kawase, S., Hannah, B., & Wang, Y. (2014). The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers. *The Journal of the Acoustical Society of America*, *136*(3), 1352–1362. https://doi.org/10.1121/1.4892770

Kayser, C., & Logothetis, N. K. (2009). Directed interaction between auditory and superior temporal cortices and their role in sensory integration. *Frontiers in Integrative Neuroscience*, *3*(7), 1–11. https://doi.org/10.3389/neuro.07

Keitel, C., Quigley, C., & Ruhnau, P. (2014). Stimulus-Driven Brain Oscillations in the Alpha

Range: Entrainment of Intrinsic Rhythms or Frequency-Following Response? *Journal of Neuroscience*, *34*(31), 10137–10140. https://doi.org/10.1523/JNEUROSCI.1904-14.2014

Keitel, C., Schröger, E., Saupe, K., & Müller, M. M. (2011). Sustained selective intermodal attention modulates processing of language-like stimuli. *Experimental Brain Research*, *213*(2–3), 321–327. https://doi.org/10.1007/s00221-011-2667-2

Kelly, S. ., 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*(October 2014), 313–334. https://doi.org/10.1080/01690960802365567

Kelly, S., Bailey, A., & Hirata, Y. (2017). Metaphoric Gestures Facilitate Perception of Intonation More than Length in Auditory Judgments of Non-Native Phonemic Contrasts. *Collabra: Psychology*, *3*(1), 7. https://doi.org/10.1525/collabra.76

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

Kelly, S. D., Creigh, P., & Bartolotti, J. (2010). Integrating speech and iconic gestures in a Stroop-like task: evidence for automatic processing. *Journal of Cognitive Neuroscience*, *22*(4), 683–694. https://doi.org/10.1162/jocn.2009.21254

Kelly, S. D., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology*, *5*(July), 673. https://doi.org/10.3389/fpsyg.2014.00673

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, *89*(1), 253–260. https://doi.org/10.1016/S0093-934X(03)00335-3

Kelly, S. D., & Lee, A. L. (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Language and Cognitive Processes*, *27*(6), 793–807. https://doi.org/10.1080/01690965.2011.581125

Kelly, S. D., Manning, S. M., & Rodak, S. (2008). Gesture Gives a Hand to Language and Learning: Perspectives from Cognitive Neuroscience, Developmental Psychology and Education. *Language and Linguistics Compass*, *2*(4), 569–588. https://doi.org/10.1111/j.1749-818X.2008.00067.x

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

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267. https://doi.org/10.1177/0956797609357327

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. https://doi.org/10.1016/j.bandl.2006.07.008

Kelly, S., Healey, M., Özyürek, A., & Holler, J. (2015). The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review*, *22*(2), 517–523. https://doi.org/10.3758/s13423-014-0681-7

Kelly, S., Hirata, Y., Simester, J., Burch, J., Cullings, E., & Demakakos, J. (2008). Effects of hand gesture and lip movements on auditory learning of second language speech sounds. *The Journal of the Acoustical Society of America*, *6*(October), 2357–2362. https://doi.org/10.1121/1.2933816

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Retrieved from https://books.google.nl/books/about/Gesture.html?id=UHILQAAACAAJ&pgis=1

Kiefer, M., Weisbrod, M., Kern, I., Maier, S., & Spitzer, M. (1998). Right hemisphere activation during indirect semantic priming: Evidence from event-related potentials. *Brain and Language*, *64*(64), 377–408. https://doi.org/10.1006/brln.1998.1979

Kilman, L., Zekveld, A., Hällgren, M., & Rönnberg, J. (2014). The influence of non-native language proficiency on speech perception performance. *Frontiers in Psychology*, *5*(JUL), 1–9. https://doi.org/10.3389/fpsyg.2014.00651

Kilner, J. M., Marchant, J. L., & Frith, C. D. (2009). Relationship between activity in human primary motor cortex during action observation and the mirror neuron system. *PloS One*, *4*(3), e4925. https://doi.org/10.1371/journal.pone.0004925

Kim, J., & Davis, C. (2014). Comparing the consistency and distinctiveness of speech produced in quiet and in noise. *Computer Speech and Language*, *28*(2), 598–606. https://doi.org/10.1016/j.csl.2013.02.002

Kim, J., Sonic, A., & Davis, C. (2011). Hearing speech in noise: Seeing a loud talker is better. *Perception*, *40*(7), 853–862. https://doi.org/10.1068/p6941

Kircher, T., Straube, B., Leube, D., Weis, S., Sachs, O., Willmes, K., … Green, A. (2009). Neural interaction of speech and gesture: Differential activations of metaphoric co-verbal gestures. *Neuropsychologia*, *47*(1), 169–179. https://doi.org/10.1016/j.neuropsychologia.2008.08.009

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception 36 ECVP Abstract Supplement.*

Klepp, A., Niccolai, V., Buccino, G., Schnitzler, A., & Biermann-Ruben, K. (2015). Language-motor interference reflected in MEG beta oscillations. *NeuroImage*, *109*, 438–448. https://doi.org/10.1016/j.neuroimage.2014.12.077

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: the inhibition-timing hypothesis. *Brain Research Reviews*, *53*(1), 63–88. https://doi.org/10.1016/j.brainresrev.2006.06.003

Koelewijn, T., van Schie, H. T., Bekkering, H., Oostenveld, R., & Jensen, O. (2008). Motor-cortical beta oscillations are modulated

by correctness of observed action. *NeuroImage*, *40*(2), 767–775. https://doi.org/10.1016/j.neuroimage.2007.12.018

Kopell, N., Kramer, M. A., Malerba, P., & Whittington, M. A. (2010). Are Different Rhythms Good for Different Functions? *Frontiers in Human Neuroscience*, *4*(November), 1–9. https://doi.org/10.3389/fnhum.2010.00187

Kotz, S. A., Meyer, M., Alter, K., Besson, M., Von Cramon, D. Y., & Friederici, A. D. (2003). On the lateralization of emotional prosody: An event-related functional MR investigation. *Brain and Language*, *86*(3), 366–376. https://doi.org/10.1016/S0093-934X(02)00532-1

Kouh, M., & Poggio, T. (2008). A Canonical Neural Circuit for Cortical Nonlinear Operations. *Neural Computation*, *1451*(20), 1427–1451.

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, *61*(5), 743–754. https://doi.org/10.1037/0022-3514.61.5.743

Krizman, J., Bradlow, A. R., Lam, S. S.-Y., & Kraus, N. (2016). How bilinguals listen in noise: linguistic and non-linguistic factors. *Bilingualism: Language and Cognition*, *FirstView*, 1–10. https://doi.org/10.1017/S1366728916000444

Król, M. E. (2018). Auditory noise increases the allocation of attention to the mouth, and the eyes pay the price: An eye-tracking study. *PLoS ONE*, *13*(3). https://doi.org/10.1371/journal.pone.0194491

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. https://doi.org/10.1016/S1364-6613(00)01560-6

Kutas, M., & Federmeier, K. D. (2014). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annu Rev Psychol.*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123.Thirty

Kutas, M., & Hillyard, S. A. (1984). Brain Potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(3947), 161–163.

Lamminmäki, S., Parkkonen, L., & Hari, R. (2014). Human Neuromagnetic Steady-State Responses to Amplitude-Modulated Tones, Speech, and Music. *Ear and Hearing*, *35*(4), 461–467.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/Doi 10.1038/Nrn2532

Lecumberri, M. L. G., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, *52*(11–12), 864–886. https://doi.org/10.1016/j.specom.2010.08.014

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: a quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Leonard, M. K., Torres, C., Travis, K. E., Brown, T. T., Hagler, D. J., Dale, A. M., … Halgren, E. (2011). Language proficiency modulates the recruitment of non-classical language areas in bilinguals. *PLoS ONE*, *6*(3). https://doi.org/10.1371/journal.pone.0018240

Lombard, E. (1911). Le signe de l'elevation de la voix. *Annals Maladiers Oreille, Larynx, Nez, Pharynx, 37*, 101–119.

Lozano-Soldevilla, D., Ter Huurne, N., Cools, R., & Jensen, O. (2014). GABAergic modulation of visual gamma and alpha oscillations and its consequences for working memory performance. *Current Biology*, *24*(24), 2878–2887. https://doi.org/10.1016/j.cub.2014.10.017

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

Macedonia, M., & Kriegstein, K. Von. (2012a). Gestures Enhance Foreign Language Learning. *Biolinguistics*, *6*(3–4), 393–416.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

Mayo, L. H., Florentine, M., & Buus, S. (1997). Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language, and Hearing Research : JSLHR*, *40*(3), 686–693. https://doi.org/10.1044/jslhr.4003.686

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. https://doi.org/10.1016/j.neuropsychologia.2012.01.010

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

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America*, *131*(1), 509–517. https://doi.org/10.1121/1.3664087

Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science (New York, N.Y.)*, *221*(4608), 389–391. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6867718

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *European Journal of Neuroscience*, *48*(7), 2609–2621. https://doi.org/10.1111/ejn.13748

Meyer, L., Obleser, J., & Friederici, A. D. (2013). Left parietal alpha enhancement during working memory-intensive sentence processing. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *49*(3), 711–721.

https://doi.org/10.1016/j.cortex.2012.03.006

Mitra, P. P., & Pesaran, B. (1999). Analysis of dynamic brain imaging data. *Biophysical Journal*, *76*(2), 691–708. https://doi.org/10.1016/S0006-3495(99)77236-X

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory – visual interactions during early sensory processing in humans : a high-density electrical mapping study. *Cognitive Brain Research*, *14*, 115–128.

Morett, L. M., & Chang, L. (2015). Emphasising sound and meaning : pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, *30*(3), 347–353. https://doi.org/10.1080/23273798.2014.923105

Morgan, S. T., Hansen, J. C., Hillyard, S. A., & Posner, M. (1996). Selective attention to stimulus location modulates the steady-state visual evoked potential. *Neurobiology*, *93*(May), 4770–4774. https://doi.org/10.1073/pnas.93.10.4770

Muir, L. J., & Richardson, I. E. G. (2005). Perception of sign language and its application to visual communications for deaf people. *Journal of Deaf Studies and Deaf Education*, *10*(4), 390–401. https://doi.org/10.1093/deafed/eni037

Müller, M. M., Andersen, S., Trujillo, N. J., Valdés-Sosa, P., Malinowski, P., & Hillyard, S. a. (2006). Feature-selective attention enhances color signals in early visual areas of the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(38), 14250–14254. https://doi.org/10.1073/pnas.0606668103

Muller, M. M., Picton, T. W., Valdes-sosa, P., Riera, J., Teder-salejarvi, W. A., & Hillyard, S. A. (1998). Effects of spatial selective attention on the steady-state visual evoked potential in the 20 – 28 Hz range. *Cogn Brain Res.*, *6*(4), 249–261.

Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, *60*(6), 926–940. https://doi.org/10.3758/BF03211929

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12. https://doi.org/10.1007/s00426-005-0031-5

Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., & Rossion, B. (2015). The steady-state visual evoked potential in vision research: A review. *Journal of Vision*, *15*(6), 1–46. https://doi.org/10.1167/15.6.4.doi

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. Psychological review, 115(2), 357.

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. https://doi.org/10.1016/j.cortex.2011.02.007

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. https://doi.org/10.1162/jocn.2010.21498

Obleser, J., & Kotz, S. a. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage*, *55*(2), 713–723. https://doi.org/10.1016/j.neuroimage.2010.12.020

Obleser, J., & Weisz, N. (2012). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cerebral Cortex (New York, N.Y. : 1991)*, *22*(11), 2466–2477. https://doi.org/10.1093/cercor/bhr325

Obleser, J., Wise, R. J. S., Alex Dresner, M., & Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(9), 2283–2289. https://doi.org/10.1523/JNEUROSCI.4663-06.2007

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse listening conditions and memory load drive a common α oscillatory network. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(36), 12376–12383. https://doi.org/10.1523/JNEUROSCI.4908-11.2012

Oliver, G., Gullberg, M., Hellwig, F., Mitterer, H., & Indefrey, P. (2012). Acquiring L2 sentence comprehension: A longitudinal study of word monitoring in noise. *Bilingualism: Language and Cognition*, *15*(May), 841–857. https://doi.org/10.1017/S1366728912000089

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 156869. https://doi.org/10.1155/2011/156869

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: insights from brain and behaviour. *Philosophical Transactions of the Royal Society B*, *369*(August), 1–10. https://doi.org/10.1098/rstb.2013.0296

Ozyürek, 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. https://doi.org/10.1162/jocn.2007.19.4.605

Parks, N. A., Hilimire, M. R., & Corballis, P. M. (2011). Steady-state signatures of visual perceptual load, multimodal distractor filtering, and neural competition. *Journal of Cognitive Neuroscience*, *23*(5), 1113–1124. https://doi.org/10.1162/jocn.2010.21460

Payne, L., & Sekuler, R. (2014). The importance of ignoring: Alpha oscillations protect selectivity. *Curr Dir Psychol Sci*, *23*(3), 171–177. https://doi.org/10.1177/0963721414529145.The

Peelle, J. E. (2018). Listening Effort. *Ear and Hearing*, *39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *68*, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, (10), 437–442.

Perniss, P. (2018). Why We Should Study Multimodal Language. *Frontiers in Psychology*, *9*(June), 1–5. https://doi.org/10.3389/fpsyg.2018.01109

Pfurtscheller, G., & Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, *110*(11), 1842–1857. https://doi.org/10.1016/S1388-2457(99)00141-8

Piai, V., Roelofs, A., Rommers, J., & Maris, E. (2015). Beta oscillations reflect memory and motor aspects of spoken word production. *Human Brain Mapping*, *36*(7), 2767–2780. https://doi.org/10.1002/hbm.22806

Picton, T. W., John, M. S., Dimitrijevic, A., Purcell, D., Picton, T. W., John, M. S., … Purcell, D. (2003). Human auditory steady-state responses : Respuestas auditivas de estado estable en humanos. *International Journal of Audiology*, *42*(4), 177–219. https://doi.org/10.3109/14992020309101316

Posner, M. I. (2016). Orienting of attention: Then and now. *Quarterly Journal of Experimental Psychology*, *69*(10), 1864–1875. https://doi.org/10.1080/17470218.2014.937446

Pulvermüller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *European Journal of Neuroscience*, *21*(3), 793–797. https://doi.org/10.1111/j.1460-9568.2005.03900.x

Rees, G., Frith, C., & Lavie, N. (2001). Processing of irrelevant visual motion during performance of an auditory attention task. *Neuropsychologia*, *39*, 937–949.

Regan, M. P., He, P., & Regan, D. (1995). An audio-visual convergence area in the human brain. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, *106*(3), 485–487. https://doi.org/10.1007/BF00231071

Regan, M. P., & Regan, D. (1988). A Frequency Domain Technique for Characterizing Nonlinearities in Biological Systems. *J. Theor. Biol.*, *133*, 293–317.

Regan, M. P., & Regan, D. (1989). Objective Investigation of Visual Function Using a Nondestructive Zoom-FFT Technique for Evoked Potential Analysis. *Canadian Journal of Neurological Sciences / Journal Canadien Des Sciences Neurologiques*, *16*(2), 168–179. https://doi.org/10.1017/S0317167100028845

Rennig, J., Wegner-Clemens, K., & Beauchamp, M. S. (2018). Face Viewing Behavior Predicts Multisensory Gain During Speech Perception. *BioRxiv*, 1–18. https://doi.org/10.1101/331306

Riseborough, M. G. (1981). Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication. *Journal of Nonverbal Behavior*, *5*(3), 172–183. https://doi.org/10.1007/BF00986134

Rogers, S. L., Speelman, C. P., Guidetti, O., & Longmuir, M. (2018). Using dual eye tracking to uncover personal gaze patterns during social

interaction. *Scientific Reports*, *8*(1), 1–9. https://doi.org/10.1038/s41598-018-22726-7

Rogers, W. T. (1978). the Contribution of Kinesic Illustrators Toward the Comprehension of Verbal Behavior Within Utterances. *Human Communication Research*, *5*(1), 54–62. https://doi.org/10.1111/j.1468-2958.1978.tb00622.x

Rohe, T., & Noppeney, U. (2016). Distinct computational principles govern multisensory integration in primary sensory and association cortices. *Current Biology*, *26*(4), 509–514. https://doi.org/10.1016/j.cub.2015.12.056

Rohe, T., & Noppeney, U. (2018). Reliability-Weighted Integration of Audiovisual Signals Can Be Modulated by Top-down Control. *Eneuro*, *5*(February), ENEURO.0315-17.2018. https://doi.org/10.1523/ENEURO.0315-17.2018

Ross, B., Herdman, A. T., & Pantev, C. (2005). Right hemispheric laterality of human 40 Hz auditory steady-state responses. *Cerebral Cortex*, *15*(12), 2029–2039. https://doi.org/10.1093/cercor/bhi078

Ross, B., Picton, T. W., Herdman, A. T., & Pantev, C. (2004). The effect of attention on the auditory steady-state response. *Neurology & Clinical Neurophysiology*, *2004*(3), 22. https://doi.org/10.1016/j.eclnm.2010.02.002

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*, *17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Saldern, S. Von, & Noppeney, U. (2013). Sensory and Striatal Areas Integrate Auditory and Visual Signals into Behavioral Benefits during Motion Discrimination, *33*(20), 8841–8849. https://doi.org/10.1523/JNEUROSCI.3020-12.2013

Saupe, K., Widmann, A., Bendixen, A., Müller, M. M., & Schröger, E. (2009). Effects of intermodal attention on the auditory steady-state response and the event-related potential. *Psychophysiology*, *46*(2), 321–327. https://doi.org/10.1111/j.1469-8986.2008.00765.x

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia*, *42*(11), 1435–1446. https://doi.org/10.1016/j.neuropsychologia.2004.04.015

Schaller, F., Weiss, S., & Müller, H. M. (2017). EEG beta-power changes reflect motor involvement in abstract action language processing. *Brain and Language*, *168*, 95–105. https://doi.org/10.1016/j.bandl.2017.01.010

Scharenborg, O., Coumans, J. M. J., & van Hout, R. (2018). The effect of background noise on the word activation process in nonnative spoken-word recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(2), 233–249. https://doi.org/10.1037/xlm0000441

Schepers, I. M., Schneider, T. R., Hipp, J. F., Engel, A. K., & Senkowski, D. (2013). Noise alters beta-band activity in superior temporal cortex during audiovisual speech processing. *NeuroImage*, *70*, 101–112. https://doi.org/10.1016/j.neuroimage.2012.11.066

Schneider, T. R., Debener, S., Oostenveld, R., & Engel, A. K. (2008). Enhanced EEG gamma-band activity reflects multisensory semantic matching in visual-to-auditory object priming. *NeuroImage*, *42*(3), 1244–1254. https://doi.org/10.1016/j.neuroimage.2008.05.033

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., Puce, A., Program, S., & Hall, A. S. (2008). Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci.*, *12*(3), 106–113. https://doi.org/10.1016/j.tics.2008.01.002.Neuronal

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69-78. https://doi.org/10.1016/j.cognition.2004.01.006

Scott, S. K. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*(12), 2400–2406. https://doi.org/10.1093/brain/123.12.2400

Senkowski, D., Saint-Amour, D., Höfle, M., & Foxe, J. J. (2011). Multisensory interactions in early evoked brain activity follow the principle of inverse effectiveness. *NeuroImage*, *56*(4), 2200–2208. https://doi.org/10.1016/j.neuroimage.2011.03.075

Senkowski, D., Schneider, T. R., Foxe, J. J., & Engel, A. K. (2008). Crossmodal binding through neural coherence: implications for multisensory processing. *Trends in Neurosciences*, *31*(8), 401–409. https://doi.org/10.1016/j.tins.2008.05.002

Shannon, R., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, *270*(5234), 303–304.

Sheehan, E. A., Namy, L. L., & Mills, D. L. (2007). Developmental changes in neural activity to familiar words and gestures. *Brain and Language*, *101*(3), 246–259. https://doi.org/10.1016/j.bandl.2006.11.008

Shimizu, T., Makishima, K., Yoshida, M., & Yamagishi, H. (2002). Effect of background noise on perception of English speech for Japanese listeners. *Auris Nasus Larynx*, *29*(2), 121–125. https://doi.org/10.1016/S0385-8146(01)00133-X

Siegel, M., Donner, T. H., & Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews. Neuroscience*, *13*(2), 121–134. https://doi.org/10.1038/nrn3137

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2009). Gestures orchestrate brain networks for language understanding. *Current Biology : CB*, *19*(8), 661–667. https://doi.org/10.1016/j.cub.2009.02.051

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2006). Lending a helping hand to hearing: another motor theory of speech perception. *Action to Language via the Mirror Neuron System*, 250–286. https://doi.org/10.1017/CBO9780511541599.009

Skipper, J. I., Wassenhove, V. Van, Nusbaum, H. C., & Steven, L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex*, *17*(10), 2387–2399. https://doi.org/10.1093/cercor/bhl147.Hearing

Spaak, E., de Lange, F. P., & Jensen, O. (2014). Local Entrainment of Alpha Oscillations by Visual Stimuli Causes Cyclic Modulation of Perception. *Journal of Neuroscience*, *34*(10), 3536–3544. https://doi.org/10.1523/JNEUROSCI.4385-13.2014

Stolk, A., Todorovic, A., Schoffelen, J. M., & Oostenveld, R. (2013). Online and offline tools for head movement compensation in MEG. *NeuroImage*, *68*, 39–48. https://doi.org/10.1016/j.neuroimage.2012.11.047

Stothart, G., & Kazanina, N. (2013). Oscillatory characteristics of the visual mismatch negativity: what evoked potentials aren't telling us. *Frontiers in Human Neuroscience*, *7*(August), 1–9. https://doi.org/10.3389/fnhum.2013.00426

Straube, B., Green, A., Bromberger, B., & Kircher, T. (2011). The differentiation of iconic and metaphoric gestures: Common and unique integration processes. *Human Brain Mapping*, *32*(4), 520–533. https://doi.org/10.1002/hbm.21041

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. https://doi.org/10.1016/j.neuropsychologia.2009.09.025

Straube, B., Green, A., Sass, K., & Kircher, T. (2014). Superior temporal sulcus disconnectivity during processing of metaphoric gestures in Schizophrenia. *Schizophrenia Bulletin*, *40*(4), 936–944. https://doi.org/10.1093/schbul/sbt110

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. https://doi.org/10.1371/journal.pone.0051207

Strauss, A., Kotz, S. A., Scharinger, M., & Obleser, J. (2014). Alpha and theta brain oscillations index dissociable processes in spoken word recognition. NeuroImage, 97, 387–395. https://doi.org/10.1016/j.neuroimage.2014.04.005

Strauss, A., Kotz, S. A., & Obleser, J. (2013). Narrowed Expectancies under Degraded Speech : Revisiting the N400. *Journal of Cognitive Neuroscience*, *25*(8), 1383–1395. https://doi.org/10.1162/jocn

Strauss, A., Wostmann, M., & Obleser, J. (2014). Cortical alpha oscillations as a tool for auditory selective inhibition. *Frontiers in Human Neuroscience*, *8*(May), 1–7. https://doi.org/10.3389/fnhum.2014.00350

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661–699. https://doi.org/10.1111/j.0023-8333.2005.00320.x

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

Summerfield, A. Q. (1983). *Audio-visual speech perception, lipreading and artificial stimulation*. *Hearing Science and Hearing Disorders*. Academic Press Inc. https://doi.org/10.1016/B978-0-12-460440-7.50010-7

Tallon-Baudry, C., & Bertrand, O. (1999).

Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences*, *3*(4), 151–162. https://doi.org/10.1016/S1364-6613(99)01299-1

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, *14*(9), 400–410. https://doi.org/10.1016/j.tics.2010.06.008

Tettamanti, M., Moro, A., Messa, C., Moresco, R. M., Rizzo, G., Carpinelli, A., … Perani, D. (2005). Basal ganglia and language: Phonology modulates dopaminergic release. *NeuroReport*, *16*(4), 397–401. https://doi.org/10.1097/00001756-200503150-00018

Tiitinen, H., Sinkkonen, J., Reinikainen, K., Alho, K., Lavaikainen, J., & Naatanen, R. (1993). Selective attention enhanced the auditory 40-Hz transient response in humans. *Nature*, *364*, 59–60.

Toni, I., de Lange, F. P., Noordzij, M. L., & Hagoort, P. (2008). Language beyond action. *Journal of Physiology Paris*, *102*(1–3), 71–79. https://doi.org/10.1016/j.jphysparis.2008.03.005

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing*, *28*(5), 656–668. https://doi.org/10.1097/AUD.0b013e31812f7185

Vaden, K. I., Kuchinsky, S. E., Cute, S. L., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. a. (2013). The Cingulo-Opercular Network Provides Word-Recognition Benefit. *The Journal of Neuroscience*, *33*(48), 18979–18986. https://doi.org/10.1523/JNEUROSCI.1417-13.2013

Van Berkum, J. J. A., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research*, *17*(3), 701–718. https://doi.org/10.1016/S0926-6410(03)00196-4

van Elk, M., van Schie, H. T., Zwaan, R. A., & Bekkering, H. (2010). The functional role of motor activation in language processing: Motor cortical oscillations support lexical-semantic retrieval. *NeuroImage*, *50*(2), 665–677. https://doi.org/10.1016/j.neuroimage.2009.12.123

Van Engen, K. J., & McLaughlin, D. J. (2018). Eyes and ears: Using eye tracking and pupillometry to understand challenges to speech recognition. *Hearing Research*, 1–11. https://doi.org/10.1016/j.heares.2018.04.013

Varela, F., Lachaux, J., Rodriguez, E., & Martinerie, J. (2001). The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, *2*, 229.

Vialatte, F. B., Maurice, M., Dauwels, J., & Cichocki, A. (2010). Steady-state visually evoked potentials: Focus on essential paradigms and future perspectives. *Progress in Neurobiology*, *90*(4), 418–438. https://doi.org/10.1016/j.pneurobio.2009.11.005

Vrba, J., & Robinson, S. E. (2001). Signal processing in magnetoencephalography. *Methods*, *25*(2), 249–271. https://doi.org/10.1006/meth.2001.1238

Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: An ERP study. *Neuropsychologia*, *51*(13), 2847–2855. https://doi.org/10.1016/j.neuropsychologia.2013.09.027

Wang, L., Hagoort, P., & Jensen, O. (2018). Language Prediction Is Reflected by Coupling between Frontal Gamma and Posterior Alpha Oscillations. *Journal of Cognitive Neuroscience*, 1–16. https://doi.org/10.1162/jocn

Wang, L., Jensen, O., van den Brink, D., Weder, N., Schoffelen, J.-M., Magyari, L., … Bastiaansen, M. (2012). Beta oscillations relate to the N400m during language comprehension. *Human Brain Mapping*, *33*(12), 2898–2912. https://doi.org/10.1002/hbm.21410

Wang, L., Zhu, Z., & Bastiaansen, M. (2012). Integration or predictability? A further specification of the functional role of gamma oscillations in language comprehension. *Frontiers in Psychology*, *3*(June), 187. https://doi.org/10.3389/fpsyg.2012.00187

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, *124*(3), 1716–1726. https://doi.org/10.1121/1.2956483

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

Weiss, S., & Mueller, H. M. (2012). "Too Many betas do not Spoil the Broth": The Role of Beta Brain Oscillations in Language Processing. *Frontiers in Psychology*, *3*(June), 201. https://doi.org/10.3389/fpsyg.2012.00201

Weisz, N., Hartmann, T., Müller, N., Lorenz, I., & Obleser, J. (2011). Alpha rhythms in audition: cognitive and clinical perspectives. *Frontiers in Psychology*, *2*(April), 73. https://doi.org/10.3389/fpsyg.2011.00073

Weisz, N., & Obleser, J. (2013). Synchronisation signatures in the listening brain: A perspective from non-invasive neuroelectrophysiology. *Hearing Research*, 1–13. https://doi.org/10.1016/j.heares.2013.07.009

Werner, S., & Noppeney, U. (2010). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral Cortex*, *20*(8), 1829–1842. https://doi.org/10.1093/cercor/bhp248

Werner, S., & Noppeney, U. (2011). The contributions of transient and sustained response codes to audiovisual integration. *Cerebral Cortex*, *21*(4), 920–931. https://doi.org/10.1093/cercor/bhq161

Wijngaarden, S. J. Van, Steeneken, H. J. M., Houtgast, T., van Wijngaarden, S. J., Steeneken, H. J. M., & Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native talkers. *The Journal of the Acoustical Society of America*, *112*(August), 3004–3013. https://doi.org/10.1121/1.1512289

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful Listening: The Processing of Degraded Speech Depends Critically on Attention. *Journal of Neuroscience*, *32*(40), 14010–14021. https://doi.org/10.1523/JNEUROSCI.1528-12.2012

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, *17*(10), 2322–2333. https://doi.org/10.1093/cercor/bhl141

Willems, R. M., Özyürek, A., & Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *NeuroImage*, *47*(4), 1992–2004. https://doi.org/10.1016/j.neuroimage.2009.05.066

Wilsch, A., Henry, M. J., Herrmann, B., Maess, B., & Obleser, J. (2014). Alpha Oscillatory Dynamics Index Temporal Expectation Benefits in Working Memory. *Cerebral Cortex (New York, N.Y. : 1991)*, 1–9. https://doi.org/10.1093/cercor/bhu004

Wong, C., & Gallate, J. (2012). The function of the anterior temporal lobe: A review of the empirical evidence. *Brain Research*, *1449*, 94–116. https://doi.org/10.1016/j.brainres.2012.02.017

Wostmann, M., Herrmann, B., Wilsch, A., & Obleser, J. (2015). Neural Alpha Dynamics in Younger and Older Listeners Reflect Acoustic Challenges and Predictive Benefits. *Journal of Neuroscience*, *35*(4), 1458–1467. https://doi.org/10.1523/JNEUROSCI.3250-14.2015

Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology*, *42*(6), 654–667. https://doi.org/10.1111/j.1469-8986.2005.00356.x

Wu, Y. C., & Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. *Brain and Language*, *101*(3), 234–245. https://doi.org/10.1016/j.bandl.2006.12.003

Wu, Y. C., & Coulson, S. (2007). Iconic gestures prime related concepts: an ERP study. *Psychonomic Bulletin & Review*, *14*(1), 57–63. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17546731

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46.

Zemon, V., & Ratliff, F. (1984). Biological Cybernetics Intermodulation Components of the Visual Evoked Potential : *Biological Cybernetics*, *408*, 401–408. https://doi.org/10.1007/BF00335197

Zhang, L., Li, Y., Wu, H., Li, X., Shu, H., Zhang, Y., & Li, P. (2016). Effects of semantic context and fundamental frequency contours on mandarin speech recognition by second language learners. *Frontiers in Psychology*, *7*(JUN), 1–8. https://doi.org/10.3389/fpsyg.2016.00908

Zhao, W., Riggs, X. K., Schindler, X. I., & Holle, X. H. (2018). Transcranial Magnetic Stimulation over Left Inferior Frontal and Posterior Temporal Cortex Disrupts Gesture-Speech Integration, *38*(8), 1891–1900. https://doi.org/10.1523/JNEUROSCI.1748-17.2017

APPENDICES

APPENDIX I

VERBS USED IN THE EXPERIMENT

Correct item		Item 1	Item 2	Item 3	Item 4
beklimmen		bestijgen	beklimmen	begroeten	bekladden
verzoeken		ritsen	verzoeken	vervloeken	vragen
faxen		fixen	bakken	faxen	sturen
blussen		bewateren	blussen	knopen	bluffen
bidden		baden	verzoeken	bidden	sluiten
ontstoppen		ontsnappen	verschonen	ontstoppen	opscheppen
ventileren		luchten	ventileren	verleren	aaien
zingen		zinken	zingen	fluiten	vissen
drijven		drijven	duwen	schillen	dreigen
aansteken		bespuiten	aansteken	aanmaken	aanstellen
schreeuwen		schreeuwen	schreden	afdrogen	gillen
fluisteren		mompelen	fluisteren	verkorten	kluisteren
luisteren		fluisteren	horen	jongleren	luisteren
yogaen		joggen	ontspannen	blussen	yoga-en
borstcrawlen		borstcrawlen	zwemmen	kammen	borstelen
dobbelen		hobbelen	dobbelen	gokken	duwen
verwijden		vermijden	haken	uitbreiden	verwijden
jongleren		jongleren	opgoooien	pottenbakken	jokeren
knuffelen		knuppelen	knuffelen	kietelen	troetelen
verminderen		begrenzen	balanceren	verminderen	verhinderen
schoppen		knuffelen	trappen	shoppen	schoppen
sporten		slijpen	storten	rennen	sporten
sluipen		stikken	sluipen	slurpen	glippen
vallen		opzoeken	flikkeren	vallen	vullen
ingooien		rollen	ingeven	inslaan	ingooien
hoepelen		ronddraaien	hoepelen	huppelen	flikkeren
speerwerpen		spreken	insmeren	speerwerpen	neerwerpen
begraven		begrijpen	bijzetten	begraven	bestellen
frutselen		kriebelen	peuteren	friemelen	knuffelen
springen		knappen	springen	stikken	swingen
toeschuiven		lopen	toeschuiven	toegeven	aanschuiven
storten		dumpen	storten	sporten	openen
schillen		pellen	schelden	wringen	schillen
slepen		slepen	hakken	trekken	slapen
wissen		uitgummen	steken	vissen	wissen
slijpen		knijpen	plakken	scherpen	slijpen

bespuiten		bespuiten	sproeien	bespelen	nieten
uitkleden		faxen	bekleden	uitdoen	uitkleden
boogschieten		boogschieten	richten	loslaten	verschieten
afkrabben		afkappen	wrijven	afkrabben	dresseren
dansen		sjansen	bewegen	monteren	dansen
stikken		stikken	huilen	smoren	stekken
hooghouden		voetballen	openen	hoogspringen	hooghouden
schoenpoetsen		schoonpoetsen	afvegen	schoenpoetsen	dippen
wiegen		schommelen	wiegen	liegen	persen
duwen		beitelen	schuiven	duwen	huwen
marcheren		markeren	stampen	melken	marcheren
kruiden		kruiden	kruipen	vouwen	zouten
begroeten		zwaaien	beboeten	begroeten	horen
haken		breien	vijlen	raken	haken
wegen		darten	wegen	wagen	wiegen
gamen		fluiten	nemen	spelen	gamen
stofzuigen		uitzuigen	vegen	uitrollen	stofzuigen
gooien		gooien	werpen	dooien	schroeven
drinken		eten	drukken	klinken	drinken
schilderen		lakken	schillen	schilderen	trekken
knopen		knopen	strikken	knippen	liegen
huilen		wenen	ruilen	huilen	steken
wurgen		golven	knijpen	wurgen	wuiven
verkorten		aanbellen	verzorgen	verkorten	verkleinen
vegen		typen	wegen	vegen	dweilen
duiken		ruiken	duiken	huilen	zwemmen
telefoneren		bellen	tolereren	telefoneren	kopieren
tuinieren		fluiten	graven	klieren	tuineren
opscheppen		klimmen	oppeppen	vullen	opscheppen
persen		raspen	passen	duwen	persen
openen		opereren	openen	draaien	kruiden
nieten		hechten	genieten	verschuiven	nieten
kopieren		zagen	kopen	namaken	kopieren
graven		schaven	scheppen	graven	verkleinen
scheuren		schuren	splijten	uitrekken	scheuren
afsluiten		afsluiten	hoepelen	afzetten	besluiten
parfumeren		parafraseren	verplaatsen	parfumeren	geuren

horen		zagen	horen	boren	luisteren
aankleden		aantrekken	aankleden	graven	bekleden
balanceren		telefoneren	wegen	blancheren	balanceren
voetballen		voetballen	schoppen	honkballen	aanbellen
melken		hoepelen	melken	mennen	trekken
duimen		friemelen	duimen	duiken	faxen
vijlen		vijlen	schoppen	veinzen	raspen
regenen		smeren	regenen	regelen	typen
fluiten		verplaatsen	verschuiven	flaneren	fluiten
filmen		fluiten	schroeven	filmen	plukken
stempelen		liften	stemmen	plukken	stempelen
klimmen		klimmen	wiegen	groeien	klappen
raspen		raspen	dresseren	krassen	ontstoppen
schommelen		schoppen	duiken	schommelen	wiegen
skieen		spieken	touwtjespringen	graven	skieen
boksen		sjokken	boksen	hooghouden	stoten
bladeren		bladeren	blazen	dobbelen	waaieren
slingeren		golven	slinken	slingeren	afwassen
verbinden		beitelen	verbinden	koppelen	verbannen
trekken		trekken	strekken	hijsen	tekenen
wenken		jojo-en	rollen	wenken	denken
waaieren		waaien	opensnijden	waaieren	wapperen
krassen		krassen	crossen	strijken	signeren
fotograferen		klikken	frankeren	kopieren	fotograferen
weggooien		werpen	weggooien	haken	weggeven
poederen		kloppen	opscheppen	ploeteren	poederen
kietelen		kietelen	knijpen	gooien	knielen
schrijven		schrijden	stoten	tekenen	schrijven
steken		afstoffen	steken	afvegen	staken
schrobben		schoonmaken	ophangen	schrobben	schrapen
kloppen		kloppen	schoonmaken	klappen	dippen
timmeren		timmeren	tillen	kloppen	geven
rollen		stampen	rollen	draaien	scrollen
stampen		stompen	eten	stampen	trappen
dobbelen		dommelen	schudden	dobbelen	smeren
strijken		strijken	pakken	smeren	bereiken
afdrogen		draaien	afdrogen	kleven	afdragen

markeren		markeren	knopen	parkeren	aanstrepen
slaan		meppen	slaan	staan	mengen
drummen		timmeren	dromen	wringen	drummen
vermengen		vermengen	mixen	vervangen	indrukken
uitrollen		uitrollen	indrukken	schuiven	uitrekken
smssen		smssen	berichten	gamen	wankelen
verplaatsen		kaatsen	vasthouden	verplaatsen	verleggen
knijpen		indrukken	knippen	soppen	knijpen
opzoeken		opzoeken	bezoeken	bladeren	gooien
tekenen		takelen	signeren	voetballen	tekenen
afweren		beweren	afweren	omslaan	stoppen
wassen		wassen	spelen	mixen	wissen
toetsen		poetsen	wegen	toetsen	indrukken
smeren		smeren	opwinden	strijken	scheren
monteren		schroeven	verteren	monteren	wijzen
afstoffen		afstoten	verwijderen	vegen	afstoffen
kogelstoten		kogelstoten	opvegen	werpen	vogelspotten
opensnijden		opensplijten	ritsen	opensnijden	rijden
zwemmen		zwemmen	friemelen	zwichten	wegduwen
verlichten		vermengen	verlichten	indrukken	verliggen
plakken		plakken	steken	kleven	lakken
ondersteunen		drummen	ondersteunen	onderstrepen	dragen
darten		darten	starten	gooien	markeren
ontkurken		ontkomen	trekken	ontkurken	openmaken
verdelen		opdienen	verdelen	vervelen	geven
aansteken		aanstellen	knijpen	indrukken	aansteken
schroeven		monteren	smssen	schroeven	schrapen
golven		golven	fietsen	wiegen	gokken
roeren		klappen	mixen	voeren	roeren
touwtjespringen		skieen	touwtrekken	touwtjespringen	plooien
hameren		timmeren	handelen	draaien	hameren
dresseren		kloppen	adresseren	schrobben	dresseren
afgieten		afgieten	schenken	schieten	omkeren
signeren		tekenen	stampen	signeren	signaleren
vertrappen		stampen	honkballen	verklappen	vertrappen
beitelen		bijtekenen	timmeren	beitelen	krassen
tekenen		rekenen	kleuren	strijken	tekenen

afwassen		afwassen	slaan	schoonmaken	verrassen
opendraaien		ontsluiten	wringen	opendraaien	verfraaien
lakken		plakken	verbinden	lakken	lijmen
vangen		hangen	grijpen	vangen	afdrogen
wijzen		wijzen	eisen	aanbellen	rollen
dippen		dippen	stippen	waaieren	dopen
frankeren		flaneren	frankeren	stempelen	verplaatsen
groeien		opzoeken	bloeien	vergroten	groeien
poolen		spoelen	poolen	stoten	schrijven
breien		friemelen	kleien	breien	slingeren
vlechten		vlechten	boksen	dansen	hechten
verschuiven		verplaatsen	verschuiven	bestuiven	klimmen
jojoen		jojo-en	joggen	jokeren	poederen
indrukken		toetsen	indrukken	uitdrukken	wenken
lopen		lopen	uitkleden	slopen	sluipen
opkloppen		sluipen	verstoppen	vermalen	opkloppen
hijsen		hijsen	wijzen	optillen	begroeten
rijden		strijden	sturen	afgieten	rijden
ophangen		rijden	opleggen	behangen	ophangen
doorgeven		doorgeven	passeren	slingeren	overgeven
uitrekken		scheiden	uitrekken	vertrekken	praten
boren		aanwijzen	boren	waaieren	horen
klappen		kloppen	gamen	samenvoegen	klappen
schudden		mixen	schudden	kogelstoten	schoppen
klikken		indrukken	stampen	prikken	klikken
dirigeren		drummen	dirigeren	regenen	irriteren
praten		vastpakken	eten	verlaten	praten
zagen		dragen	scheiden	snijden	zagen
lezen		bladeren	racen	toeteren	lezen
zouten		zouten	zaaien	kruiden	friemelen
aankruisen		aankruisen	verhuizen	inkleuren	wissen
bewateren		gieten	voeren	bewateren	schateren
hakken		pakken	timmeren	hakken	boksen
plukken		plakken	plukken	grijpen	wringen
stapelen		roken	bouwen	stapelen	stappen
joggen		rennen	pochen	joggen	markeren
doden		afkrabben	doden	snijden	dolen

pellen		bellen	pellen	schrapen	roeien
vissen		weggooien	hengelen	wissen	vissen
toeteren		drummen	stoppen	ploeteren	toeteren
insmeren		aaien	combineren	insmeren	imiteren
ritsen		schrobben	ritsen	splitsen	sluiten
aanbellen		bestellen	aanbellen	indrukken	aanstellen
scrollen		dobbelen	hollen	scrollen	klikken
fietsen		rijden	fietsen	kletsen	gieten
omdraaien		dirigeren	graaien	omdraaien	omslaan
knippen		happen	knippen	slippen	oppompen
naaien		haken	strijken	naaien	zwaaien
schuiven		slaan	stuiven	rijden	schuiven
roeien		roeien	malen	schoenpoetsen	stoeien
stoppen		trekken	sluiten	stoppen	shoppen
mixen		draaien	doden	mixen	niksen
omarmen		ingooien	opwarmen	omarmen	samenvoegen
zwaaien		wijzen	draaien	zwaaien	filmen
uitgummen		wissen	uitgummen	vegen	duimen
basketballen		darten	basketballen	omdraaien	bevallen
eten		eten	happen	buigen	meten
verscheuren		verkleuren	scheiden	verscheuren	uitgummen
strikken		strikken	knopen	tikken	zouten
sjoelen		kammen	voelen	schuiven	sjoelen
dribbelen		kibbelen	dribbelen	stuiteren	bukken
dragen		strijken	vragen	dragen	vasthouden
vliegen		wiegen	bladeren	vliegen	opstijgen
roken		koken	roken	fluiten	drinken
masseren		knijpen	scheren	voetballen	masseren
voeren		aangeven	voeren	roeren	wijzen
stoten		proosten	stoten	plakken	quoten
proosten		filmen	proosten	troosten	stoten
beven		beven	trillen	melken	leven
verlichten		plukken	vijlen	verlichten	verplichten
tellen		tellen	bellen	kietelen	stapelen
jongleren		jongleren	wegen	gooien	omkeren
kammen		kammen	dammen	krabben	lakken
scheiden		verscheuren	snijden	scheiden	kaatsen

wringen		draaien	yoga-en	dringen	wringen
gieten		pellen	inschenken	genieten	gieten
bellen		bellen	oppakken	dirigeren	pellen
kaarten		uitdelen	lakken	tarten	kaarten
zeven		beven	voeren	zeven	schudden
snijden		rijden	hakken	snijden	storten
knikken		knokken	knikken	breien	koken
omslaan		omslaan	vouwen	opstaan	draaien
vouwen		vouwen	bouwen	klappen	verlichten
typen		schrijven	typen	piepen	dribbelen
schoonmaken		schoonmaken	aandrijven	boenen	schaken
schieten		wijzen	gieten	schieten	sjoelen
optillen		gillen	dragen	liften	optillen
oppompen		oppompen	stompen	roeien	scrollen
glijden		stoten	uitrollen	spreiden	glijden
buigen		vliegen	juichen	knikken	buigen
bakken		bakken	schudden	ondersteunen	plakken
aaien		aaien	overslaan	insmeren	maaien
scheren		schoonmaken	scheren	schrapen	leren

APPENDIX II

SUPPLEMENTARY MATERIALS

Supplementary materials - Chapter 3

S1. Alpha power contrast comparisons - analyses of single contrasts

Analyses of the single contrasts of the four conditions concur with these results: In line with previous research on degraded speech comprehension (Weisz et al., 2011; Obleser and Weisz, 2012; Becker et al., 2013; Meyer et al., 2013; Strauß et al., 2014a; Strauß et al., 2014b; Wostmann et al., 2015), alpha power was increased when comparing D to C (one significant positive cluster, $p = .02$, (left)-central and temporoparietal regions, 0.7 - 2.0 s). In a gestural context, however, cluster-based permutation tests revealed that alpha power was more suppressed in response to DG than D (one significant negative cluster, $p < .001$, going from left-temporal regions in an early time window (~0.7 - 0.9s), visual regions in a later time window (~ 0.95 - 1.6s), and back to left-temporal regions and visual regions in the final time window (1.6 - 2.0)). Finally, alpha power was more suppressed in response to CG than C (one significant negative cluster, $p < .001$, visual regions, 0.7 – 2.0), but did not differ when comparing DG and CG ($p = .20$).

S2. Beta power contrast comparisons - analyses of single contrasts

In order to rule out that the observed beta suppression in this interaction effect was not due to simply seeing visible speech (i.e., lips) we compared the single conditions and found that this larger beta suppression only occurred in conditions in which a gesture was present. Beta power was more suppressed over left-temporal, motor and visual areas for DG than CG (one negative cluster, $p < .001$, 0.7 – 2.0), more suppressed over left-temporal, motor and visual areas for DG than D (one negative cluster, $p < .001$, 0.7 - 2.0 s), more suppressed over central-parietal regions in response to CG as compared to C (one negative cluster, $p < .01$, 0.7 - 2.0 s), and more enhanced over left motor areas for D than C (one positive cluster, $p = .03$, 0.95 - 1.55 s). Thus, we observed a power suppression in all contrasts containing gestural information, but not in the one contrast that only contains visible speech (in fact, here the occurence of degraded speech even caused enhanced beta power as compared to clear speech, especially over left-central sensors). This suggests that the observed suppression is driven by the

gesture, but not by other visual information, such as visible speech (which was also present in all conditions).

S3. Gamma power contrast comparisons - analyses of single contrasts

When comparing the single conditions, cluster-based permutation tests revealed differences between DG and CG (one positive cluster, left-temporal areas, $p < .05$, 1.0 - 1.7 s) between DG and D (one positive cluster, bilateral temporal-parietal areas and occipital regions, $p < .05$, 1.0 - 2.0 s) but not between D and C ($p = .15$) or CG and C ($p = .21$). These results suggest that gestural information might require more neuronal computation and active processing when speech is degraded and a gesture is present.

Supplementary materials – Chapter 4

S4.

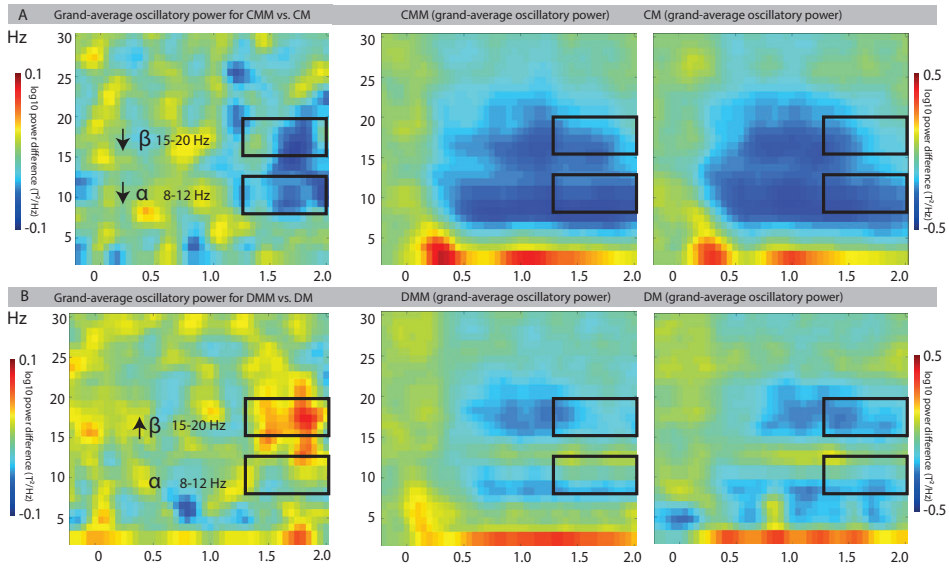


Figure S4: A: Grand-average oscillatory power for CMM vs CM, and corresponding single conditions (CMM, middle panel, CM right panel). B: Grand-average oscillatory power for DMM vs. DM, and corresponding single conditions (DMM, middle panel, DM, right panel).

Supplementary materials - Chapter 8

S5. Eye-tracking results

Native listeners (within-group) - face & mouth

In both clear and degraded speech, native listeners fixated more on the face when no gesture was present (clear speech: CG<CO, p = .004, 538 – 1301 ms, degraded speech: DG<DO, p = 0.001, 249 – 1461 ms). In both clear and degraded speech, listeners gazed more to the mouth when no gesture was present (clear speech: CG<CO, two clusters: p = < 0.001, 262 – 377 ms; p = .023, 822 – 1344 ms, degraded speech DG < DO, two clusters: p = .02, 1059 – 1420 ms, p = .019, 1441 – 1551 ms).

Non-native listeners (within-group - face & mouth)

In both clear and degraded speech, non-native listeners looked more at the face when no gesture was present (clear speech: CG < CO, p < .001, 221 – 1378 ms, degraded speech: DG < DO, p < .001, 297 – 2000 ms).

In both clear and degraded speech, non-native listeners gazed more often to the mouth when a gesture was not present as compared to present (clear speech: CG<CO, three clusters: p = .002, 815 – 1502 ms; p = .0026, 1605 – 1830 ms; p = 0.03, 1845 – 2000 ms, degraded speech: two clusters: p < .001, 191 – 488 ms; p = .01, 537 – 1268 ms).

Native vs. non-native listeners (between-group): face & mouth

No differences between native and non-native listeners were observed when comparing the differences between the proportions of fixations to the face in the CG-CO (p = .35) or DG-CO (p = .27) conditions.

We did not observe differences between native and non-native listeners when comparing the differences in fixations to the mouth in the DG-DO (p = .16) condition.

Native listeners (within-group) - gesture

In both clear and degraded speech, native listeners looked more at the body when a gesture was present than not present (clear speech: CG > CO, $p < .001$, 233 – 1425 ms, degraded speech: DG > DO, $p < .001$, 245 – 2000 ms).

Non-native listeners (within-group) - gesture

In both clear and degraded speech, non-native listeners look more at the body when a gesture is present (clear speech: CG>CO, $p < .001$, 217 – 1448 ms., degraded speech DG>DO, $p = .001$, 264 – 2000 ms).

NEDERLANDSE SAMENVATTING

Stel je voor: het is weekend, en je hebt met een vriendin afgesproken in een een druk café. De muziek staat aan, er zijn malende koffiezetapparaten te horen op de achtergrond, en mensen zijn druk met elkaar in gesprek. Wanneer je vriendin nu over het lawaai heen zou roepen en je zou vragen of je iets te drinken zou willen, dan zou je haar waarschijnlijk niet verstaan. Maar als ze dat zou doen terwijl ze met haar hand een glas uitvbeeldt dat naar de mond wordt gebracht, dan snap je haar waarschijnlijk wel.

Dit proefschrift gaat over dat soort betekenisvolle handbewegingen, die ook wel 'iconische handbewegingen' worden genoemd. In dit proefschrift onderzoek ik hoe iconische handbewegingen iets kunnen bijdragen aan taalbegrip in situaties waarin het lastig is om iemand te verstaan. Dat kan bijvoorbeeld zijn wanneer je je in een lawaaiige omgeving bevindt, zoals in dat café, maar ook wanneer Nederlands je moedertaal niet is. Helpen die handbewegingen je bijvoorbeeld meer of juist minder wanneer Nederlands je moedertaal niet is? En wat gebeurt er nu eigenlijk in de hersenen wanneer je iemand hoort én ziet praten in dat soort lawaaiige omgevingen? In dit proefschrift probeer ik deze vragen te beantwoorden door gebruik te maken van gedragsexperimenten, het bestuderen van oogbewegingen, en door te kijken naar elektrische signalen in het brein.

Het uitgangspunt van mijn proefschrift was dat tot nu toe, het meeste taalonderzoek zich vooral gericht had op gesproken woorden, zonder daarbij de visuele context te betrekken waarin die woorden normaal voorkomen. Veel onderzoeken lieten proefpersonen hierbij taal door een koptelefoon horen. In natuurlijke gesprekken werkt dat natuurlijk anders: daar horen én zien we mensen praten. Tijdens dat soort gesprekken bewegen mensen hun lippen, maar ook hun handen. Uit eerder onderzoek weten we dat luisteraars informatie uit die lip- en handbewegingen halen, zeker wanneer taal lastig te begrijpen is. Daarom vind ik het belangrijk om taal juist in een volledige multimodale context te bestuderen.

Eerder onderzoek heeft bijvoorbeeld aangetoond dat lipbewegingen je kunnen helpen om fonologische informatie te herkennen, zoals spraakklanken. Ander onderzoek heeft ook aangetoond dat iconische handbewegingen, zoals dat drinkgebaar in het café dat ik eerder beschreef, je kunnen helpen de betekenis van een woord te herkennen. De invloed van zowel handbewegingen als lipbewegingen op

taalbegrip was echter tot nu toe nog niet onderzocht in één en dezelfde context.

In hoofdstuk 2 hebben we daarom onderzocht wat die lip- en handbewegingen bijdragen aan taalbegrip wanneer een luisteraar ze in één context waarneemt. Proefpersonen bekeken in dit onderzoek verschillende filmpjes waarin een vrouw te zien was die een werkwoord uitsprak. Soms zag een proefpersoon daarbij haar lippen, en soms niet, en soms maakte ze daarbij een handbeweging, en soms niet. Het geluid in de filmpjes hebben we óf volledig duidelijk gemaakt, zoals bijvoorbeeld in een hele stille kamer, óf een beetje ruizig, óf heel erg ruizig. Doordat we verschillende lawaainiveaus gebruikten, konden we precies zien of lip- en handbewegingen juist extra veel bijdroegen aan taalbegrip wanneer het geluid heel ruizig was, of juist niet. Daarnaast konden we door soms de lippen te laten zien, en soms niet, precies zien wat de informatie van lipbewegingen bijdroeg aan taalbegrip. Hetzelfde gold natuurlijk voor de informatie van handbewegingen, en van de lip- en handbewegingen samen. Aan de proefpersonen hebben we gevraagd of ze konden aangeven welk werkwoord ze hoorden in de video's. Vervolgens hebben wij gekeken hoeveel antwoorden de proefpersonen goed hadden per type filmpje en per type geluid, en hoe erg hun begrip van de werkwoorden verbeterde naar mate het geluid beter werd, en naar mate er meer informatie in de visuele context te vinden was in de vorm van lip- en handbewegingen.

Uit de resultaten bleek dat proefpersonen vooral veel profijt hadden van lip- en handbewegingen wanneer het geluid een beetje ruizig was. Als het geluid namelijk te ruizig was, konden ze de fonologische informatie die door de lipbewegingen werd overgebracht namelijk niet meer koppelen aan de spraak, en daardoor geen profijt meer hebben van zowel lip- en handbewegingen in dezelfde context.

Vervolgens hebben we in hoofdstuk 3 en hoofdstuk 4 bekeken hoe dit nu precies werkt in de hersenen. Wanneer je namelijk iemand hoort en ziet praten zijn er erg veel hersengebieden die iets kunnen bijdragen aan dat proces, zoals hersengebieden die iets te maken hebben met taal, bewegingen, horen, en zicht. Al die hersengebieden moeten vervolgens met elkaar in contact staan om informatie uit te wisselen en om de informatie die je ziet en hoort aan elkaar te koppelen.

Ritmische hersengolven, ook wel 'neurale oscillaties' genoemd, zijn heel

belangrijk voor dit soort processen. Eerder onderzoek heeft bijvoorbeeld al laten zien dat veranderingen in de sterkte van dit soort hersengolven in verband gebracht kon worden met veranderingen in allerlei cognitieve processen, zoals het geheugen, het verwerken en plannen van bewegingen, en het bewustzijn. Daarnaast weten we uit eerder onderzoek dat lokale veranderingen in de sterkte van hersengolven informatief kunnen zijn over hoe en wanneer bepaalde hersengebieden mee doen aan een proces. Op deze manier kunnen we dus neurale oscillaties gebruiken om de neurobiologische mechanismen te onderzoeken die betrokken zijn bij multimodale taalverwerking. Dit helpt ons vervolgens weer om te begrijpen hoe je hersenen er voor zorgen dat wat je ziet gekoppeld wordt aan wat je hoort wanneer je met iemand praat in een natuurlijke context.

In hoofdstuk 3 hebben we gebruik gemaakt van MEG, magnetoencephalografie, waarbij we het magnetische veld dat opgewekt wordt door de elektrische impulsen uit de hersenen hebben onderzocht. In dit onderzoek wilden we erachter komen hoe neurale oscillaties zouden verschillen of overeenkomen wanneer handbewegingen taalbegrip zouden vergroten wanneer spraak ruizig was, of juist niet. Om dit te bestuderen hebben we Nederlandse proefpersonen naar filmpjes laten kijken waarbij een vrouw een werkwoord uitsprak. Soms maakte ze daarbij een handbeweging, en soms niet, en soms was het geluid een klein beetje ruizig, en soms niet. De proefpersonen moesten vervolgens aangeven welk werkwoord ze gehoord hadden door te kiezen uit vier antwoordopties.

In dit onderzoek waren we geinteresseerd in veranderingen in de sterkte van langzame hersengolven, waarbij het signaal langzaam fluctueert (alpha-golven, van 8-12 Hz), iets snellere golven (beta-golven, 13- 30 Hz), en snelle golven (gamma-golven, 30 – 100 Hz). Eerder onderzoek suggereerde dat wanneer de sterkte van de langzame, alpha- en beta-golven omlaag gaat in bepaalde hersengebieden dat deze hersengebieden actief mee doen aan een bepaald proces. Bij gamma-golven werkt dat net andersom: als gamma-golven sterker worden in een bepaald hersengebied, dan duidt dat vermoedelijk aan dat neuronen in een bepaald hersengebied actief bezig zijn met een bepaald proces.

De gedragsresultaten lieten zien dat handbewegingen taalbegrip het meest verbeterden wanneer taal een beetje ruizig was. De neurale resultaten lieten zien

dat wanneer je je in een ruizige situatie bevindt en een handbeweging ervoor zorgt dat je taal beter begrijpt, de hersengebieden voor zicht, bewegingen, en taal extra actief worden om de informatie van een betekenisvolle handbeweging te verwerken. Dit zagen we zowel in de alpha- als de beta-golven. Sterker nog, specifiek het hersengebied dat in verband wordt gebracht met je handen was sterker betrokken bij het proces wanneer spraak ruizig was! Gamma-golven waren het sterkste in hersengebieden die iets te maken hadden met geheugen en in hersengebieden die iets te maken hebben met het verwerken van taal. Dit kan dus betekenen dat wanneer spraak ruizig is, een handbeweging er voor kan zorgen dat je een woord makkelijker kunt ophalen uit je mentale woordenboek. De gedragsresultaten en neurale resultaten konden ook per proefpersoon met elkaar in verband worden gebracht: hoe meer een proefpersoon profijt had van een handbeweging in een ruizige omgeving, hoe minder sterk de alpha- en beta-golven waren, en hoe sterker de gamma-golven waren tijdens dit proces.

In hoofdstuk 4 hebben we onderzocht of het uitmaakt of de betekenis van een handbeweging overeenkomt met de spraak of niet, en of dat verschilt in ruizige en niet-ruizige luisteromgevingen. Dit was interessant omdat dit twee manieren zijn om taalbegrip lastig te maken. De resultaten lieten zien dat wanneer spraak niet ruizig was en de betekenis van een handbeweging niet overeenkwam met de spraak, dat alpha- en beta-golven ervoor zorgden dat hersengebieden die nodig zijn voor de koppeling van handbewegingen en spraak en hersengebieden die iets te maken hebben met bewegingen en zicht extra actief werden. Dit komt waarschijnlijk doordat deze hersengebieden extra hard moeten werken om de niet-overeenkomende informatie te verwerken. Wanneer spraak ruizig was, waren juist hersengebieden die iets te maken hebben met geheugen minder actief. Dat komt waarschijnlijk doordat een handbeweging die niet overeenkomt met de spraak ook niet kan helpen om een woord op te halen uit je mentale woordenboek, zeker niet in een ruizige omgeving.

In hoofdstuk 5 hebben we onderzocht of niet-moedertaalsprekers van het Nederlands net zoveel profijt hadden van betekenisvolle handbewegingen als moedertaalsprekers van het Nederlands. Dit was interessant, omdat dit ook een manier is waarop taal lastiger te begrijpen is, naast het horen van taal in een ruizige

omgeving. Zo zou het bijvoorbeeld kunnen zijn dat niet-moedertaalsprekers juist extra veel profijt hebben van betekenisvolle handbewegingen, omdat de betekenisvolle informatie makkelijk te herkennen is. Aan de andere kant zou het ook zo kunnen zijn dat ze minder profijt hebben van betekenisvolle handbewegingen, omdat het lastiger is voor ze om die betekenis te koppelen aan de ruizige spraak. De niet-moedertaalsprekers die we onderzochten moesten echter wel in staat zijn de spraak te begrijpen. Daarom hebben we niet-moedertaalsprekers gezocht die heel vaardig zijn in het Nederlands, om te kijken of zij net zoveel profijt hebben van de informatie die overgebracht wordt door lip- en handbewegingen als moedertaalsprekers, zeker wanneer de spraak een beetje ruizig is.

Om dit te onderzoeken hebben we daarom hetzelfde experiment als in hoofdstuk 2 uitgevoerd, en de resultaten van de moedertaalsprekers vergeleken met de resultaten van de niet-moedertaalsprekers. Uit de resultaten konden we opmaken dat hoewel de niet-moedertaalsprekers zeker profijt hadden van handbewegingen tijdens het begrijpen van taal in ruis, dat dit toch minder was dan voor de moedertaalsprekers. We zagen daarbij vooral dat de niet-moedertaalsprekers het lastiger vonden om de ruizige spraakklanken te herkennen en de informatie van de lipbewegingen die ze zagen aan de ruizige spraakklanken te koppelen. Daardoor konden ze vervolgens ook niet optimaal gebruik maken van de betekenisvolle informatie die werd overgebracht door de handbewegingen.

In hoofdstuk 6 hebben we gebruik gemaakt van EEG, electroencephalografie, om te kijken of de integratie van handbewegingen en duidelijke of ruizige spraak op dezelfde manier tot stand komt in de hersenen van moedertaalsprekers en niet-moedertaalsprekers van het Nederlands. Dit hebben we gedaan door participanten opnieuw filmpjes te laten zien waarin een vrouw te zien was die een werkwoord uitsprak in ruis of in duidelijke spraak, terwijl ze een handbeweging maakte die overeenkwam met de spraak, of juist niet. Ongeveer 400 milliseconden nadat de proefpersonen de niet-overeenkomende handbeweging zagen en de spraak duidelijk was, was er een verandering te zien in het EEG-signaal van zowel moedertaalsprekers als niet-moedertaalsprekers. De hersenen reageren dus expliciet op de mate van hoe goed de betekenis van een handbeweging bij de spraak past. Dit effect was zelfs sterker voor niet-moedertaalsprekers dan

voor moedertaalsprekers. Wanneer de spraak ruizig was, zag je dit effect alleen nog maar bij de moedertaalsprekers. We concludeerden hieruit dat de niet-moedertaalsprekers misschien meer van het geluid moeten begrijpen om de betekenis van een handbeweging aan een ruizige woord te kunnen koppelen.

In hoofdstuk 7 hebben we onderzocht hoe betekenisvolle handbewegingen het begrip van duidelijke en ruizige taal kunnen vergroten voor niet-moedertaalsprekers. Deze resultaten hebben we vervolgens met de resultaten uit hoofdstuk 3 vergeleken van de moedertaalsprekers om de neurale verschillen en overeenkomsten tussen deze twee groepen te onderzoeken. In dit experiment hebben we daarom dus gebruik gemaakt van dezelfde experimentele opzet als in hoofdstuk 3. De neurale resultaten lieten opnieuw zien dat wanneer een niet-moedertaal spreker zich in een ruizige situatie bevindt en een handbeweging ervoor zorgt dat taal beter te begrijpen is, dat de hersengebieden voor zicht, bewegingen en taal extra actief worden om de informatie van een betekenisvolle handbeweging te verwerken. Dit zagen we zowel in de alpha- als de beta-golven. Opnieuw zagen we ook een sterke relatie tussen hoeveel profijt een proefpersoon had van de betekenisvolle handbewegingen tijdens de gedragstaak en hun hersenactiviteit. We vonden in dit experiment echter geen effect in de gamma band, zoals bij de moedertaalsprekers. Dit zou kunnen komen doordat het lastiger is voor de niet-moedertaalsprekers om de betekenis van een handbeweging te gebruiken om een ruizig woord op te halen uit het mentale woordenboek. Wanneer we de niet-moedertaalsprekers met de moedertaalsprekers vergeleken, zagen we dat de hersengebieden die iets te maken hebben met het koppelen van de ruizige spraak en het gebaar, en het hersengebied dat te maken heeft met mondbewegingen, minder werden betrokken bij het taalverwerkingsproces. Dit laat dus, net als in hoofdstuk 5 en hoofdstuk 6, zien dat het lastiger is voor niet-moedertaalsprekers om de ruizige spraak te koppelen met de betekenisvolle handbewegingen.

In hoofdstuk 8 hebben we onderzocht of moedertaalsprekers en niet-moedertaalsprekers vaker naar lip- en handbewegingen kijken wanneer spraak ruizig is. Op basis van eerdere hoofdstukken zouden we namelijk kunnen verwachten dat luisteraars eerder naar betekenisvolle handbewegingen kijken wanneer spraak ruizig is, zodat ze extra veel betekenisvolle informatie uit de

beweging kunnen halen. Dit bleek iets gecompliceerder: zowel moedertaalsprekers als niet-moedertaalsprekers keken namelijk vaker naar het gezicht, en specifiek naar de lippen, dan naar het lichaam wanneer de spraak een beetje ruizig was. Niet-moedertaalsprekers keken wel vaker naar handbewegingen dan moedertaalsprekers. Dit betekent dus dat hoewel niet-moedertaalsprekers meer naar handbewegingen keken, het toch lastiger was voor hen om deze betekenisvolle informatie te verwerken en te koppelen aan het spraaksignaal. Alleen voor de moedertaalsprekers vonden we een relatie tussen hoeveel ze naar de handbeweging keken en hoeveel profijt ze hadden van die handbeweging wanneer ze ruizige spraak begrepen: hoe meer ze keken, hoe meer profijt ze hadden.

Zoals ik al eerder benadrukte, is het begrijpen van taal een enorm complex proces. In de hersenen zijn er namelijk heel veel processen die tegelijkertijd plaatsvinden wanneer je taal begrijpt. In hoofdstuk 9 hebben we gebruik gemaakt van een nieuwe techniek genaamd 'rapid invisible frequency tagging' (RIFT), zodat we precies in de hersenen konden volgen hoe auditieve en visuele informatie door het brein reist. We lieten hierbij opnieuw filmpjes zien aan onze proefpersonen, opnieuw met ruizige of duidelijke spraak, en een handbeweging of geen handbeweging. Het gedeelte van de video waar de handbeweging te zien was hebben we vervolgens op een specifieke snelheid laten knipperen. Zo snel, dat je het met het blote oog niet kan zien. Dat hebben we ook met het geluid gedaan. Vervolgens hebben we precies gekeken naar deze 'knippersnelheid' in het brein, voor zowel het geluid als de video. Op die manier konden we precies zien hoe de informatie van het geluid en de informatie van de video door het brein reisden, en hoe die informatie aan elkaar gekoppeld werd. Dit bleek precies op de plek in de hersenen te zijn waar we al eerder effecten zagen van de koppeling tussen een betekenisvolle handbeweging en de spraak. RIFT bleek dus een hele veelbelovende techniek te zijn om in de toekomst in te zetten voor vervolgonderzoeken, zeker wanneer je erachter wil komen hoe verschillende informatiestromen in het brein aan elkaar gekoppeld worden, en waar in de hersenen dat precies gebeurt.

In dit proefschrift heb ik laten zien dat het koppelen van betekenisvolle handbewegingen en spraak kan er voor zorgen dat je taal beter begrijpt, ook

wanneer taal lastig te begrijpen is, zoals in een ruizige bar, of wanneer je een niet-moedertaalspreker bent van het Nederlands. Mechanistisch gezien, zorgt het brein hier waarschijnlijk voor door de sterkte van alpha en beta golven te onderdrukken, zodat bepaalde hersengebieden actief mee kunnen doen aan een bepaald proces. Dit lijkt op eenzelfde manier te gaan in beide typen lastige luistersituaties. Toch zitten er ook wat verschillen tussen moedertaalsprekers en niet-moedertaalsprekers. Voor niet-moedertaalsprekers is het lastiger om de ruizige spraakklanken te herkennen dan voor moedertaalsprekers. Hierdoor kan het voor hen lastiger zijn om profijt te hebben van de betekenisvolle informatie die overgebracht wordt door de handen. Dit zagen we zowel in het gedrag, in hun oogbewegingen, en in hun hersenactiviteit. Tot slot heb ik laten zien dat RIFT gebruik kan worden om de koppeling van audiovisuele informatie te onderzoeken. Dit biedt uitstekende kansen voor vervolgonderzoek om te onderzoeken hoe luisteraars hun aandacht verdelen over verschillende bronnen van informatie terwijl ze taal begrijpen, en hoe die verdeling van aandacht verandert over de tijd heen. Uiteindelijk kunnen we dan ook nog beter begrijpen wat voor rol hersengolven spelen in dat proces.

CURRICULUM VITAE

Linda Drijvers was born on December 12, 1990, in 's-Hertogenbosch, The Netherlands. She obtained a bachelor's degree in Dutch Literature and Culture, and a research master's degree in Cognitive Neuroscience (cum laude) from Radboud University. Linda's bachelor's thesis was completed under the supervision of prof. dr. Paula Fikkert, and focused on sound-symbolism in dyslexia. In her master's thesis, Linda used EEG to study the oscillatory dynamics underlying reduced and unreduced word processing under supervision of prof. dr. Mirjam Ernestus. Linda then continued her academic studies and joined the Language in Interaction consortium to work with prof. dr. Asli Ozyurek and prof. dr. Ole Jensen on the studies presented in this doctoral thesis. In addition to her research, Linda worked part-time as a lecturer at Communication & Information Sciences. She was involved in teaching several courses, and supervised many bachelor and master research projects. Linda fulfilled the role of PhD representative of the Language in Interaction consortium for several years, and was involved in organizing multi-day conferences and workshops. In 2015, Linda was elected as 'Face of Science', a division of the The Royal Dutch Academy of Arts and Sciences (KNAW). In this role, she was heavily involved in public outreach, in the form of (radio) interviews, guest lectures, theater performances and blogs about her research. In 2017, Linda received the Christine Mohrmann stipend for outstanding female PhD candidates. She used her stipend to visit the University of Birmingham and University of Oxford. In January 2019, Linda joined the CoSI lab to work with dr. Judith Holler to study the cognitive and neural mechanisms underlying multimodal language comprehension in interactive settings.

AUTHOR PUBLICATIONS

Drijvers, L., Van der Plas, M., Ozyurek, A., Jensen, O. (in press). Native and non-native listeners show similar yet distinct oscillatory dynamics when using gestures to access speech in noise. *NeuroImage* .

Drijvers, L. & Ozyurek, A. (2019). Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language and Speech*.

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

Drijvers, L., & Trujillo, J. P. (2018). Commentary: Transcranial magnetic stimulation over left inferior frontal and posterior temporal cortex disrupts gesture-speech integration. *Frontiers in Human Neuroscience*, 12: 256.

Drijvers, L., Ozyurek, A., & Jensen, O. (2018). Alpha and beta oscillations index semantic congruency between speech and gestures in clear and degraded speech. *Journal of Cognitive Neuroscience*, 30(8), 1086-1097.

Drijvers, L., Ozyurek, A., & Jensen, O. (2018). Hearing and seeing meaning in noise: Alpha, beta and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human Brain Mapping*, 39(5), 2075-2087.

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

Drijvers, L., Mulder, K., & Ernestus, M. (2016). Alpha and gamma band oscillations index differential processing of acoustically reduced and full forms. *Brain and Language*, 153-154, 27-37.

Drijvers, L., Zaadnoordijk, L., & Dingemanse, M. (2015). Sound-symbolism

is disrupted in dyslexia: Implications for the role of cross-modal abstraction processes. In D. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci 2015) (pp. 602-607). Austin, Tx: Cognitive Science Society.

Drijvers, L., Vaitonyte, J., Ozyurek, A. (in revision). Visual attention to gestures reflects processing differences in native and non-native listeners during degraded speech comprehension

Schubotz, L., Holler J., **Drijvers, L.** Ozyurek, A. (in revision). Aging and working memory modulate the ability to benefit from visible speech and iconic gestures during speech-in-noise comprehension.

Mulder, K., **Drijvers, L**., Ernestus, M. (submitted). Processes underlying the comprehension of reduced and full word forms: An ERP Study.

Drijvers, L., Spaak, E., Herring, J., Ozyurek, A., Jensen, O. (in preparation). Speech-gesture integration studied by rapid frequency tagging.

ACKNOWLEDGEMENTS

I could write another book about all the wonderful people that have helped me in one way or another during my time as a PhD student. I feel extremely lucky to have met you all.

First and foremost, it was an honour to learn from two extremely smart, kind and inspiring 'academic parents' – my promotors, **Asli & Ole**.

Asli, thank you for always believing in me and supporting me (even when I was being super eigenwijs), and thank you for being my compass in the academic world. I always had the feeling we could discuss anything, at any moment (even late at night via Twitter...), no matter how busy you were. I greatly valued that (and still do!). Thank you for giving me the freedom to pursue my own interests, and for always challenging me. You are a great role model and a fantastic supervisor, and you certainly taught me how to speak up. It was a privilige to work with you and I will certainly miss it.

Ole, thanks for taking a chance with us psycholinguists. I know our collaboration was a unique match, but this interdisciplinary team has benefitted me enormously. Thank you for always taking the time to listen to me when things were difficult, for always making me smile, and explaining the most complicated concepts in simple doodles and weird analogies. Even though you moved to Birmingham, I always felt like you were close. Many thanks to **Freya, Oskar & Felix** for always being extremely welcoming during my Birmingham visits (even when I came by multiple evenings in one week). You have made my stays in Birmingham feel like a home away from home.

Many thanks to my manuscript committee, **James McQueen, Marianne Gullberg & Jonas Obleser** for reading and evaluating my thesis. I can imagine it was not trivial to plough through 10 chapters, but I'm very grateful that you did. Your comments have greatly improved the content of this thesis, and I look

forward discussing its content with you and the rest of the corona (safe to say before the actual defense, I guess...). Thanks again for your patience and flexibility during the date-planning.

Although my research project was part of the Donders Institute and the Centre for Language Studies, I was mostly based at the MPI. There, I had the privilege to enjoy the amazing support of the **technical group of the MPI**. Many thanks to **Peter, Alex, Ad, Herbert and Tobias** for your patience and supporting me in all my technical 'adventures'. **Johan**, without your support and expert eye we would never have made it to Lowlands festival with our one-way screen. **Nick**, thank you for helping me with my stimuli. You are greatly missed. **Jeroen**, thanks for helping out with all things audio/video related. Without all of you, I couldn't have done most of my work. The same goes for the technical support at the DCCN: thanks to **Jessica & Uriel** for helping out at the MEG. **Paul**, thanks for your help with the MRI.

Apart from the technical groups, I would like to thank the administration and supporting staff at both the MPI and DCCN: Thank you **Ayse**, for dealing with all my administrative and participant-related questions at the DCCN. **Carolin** & **Julia** for helping me with all questions related to Language in Interaction (and beyond). **Jan** & **Robèr**, for always helping out with well, basically anything. We are surely lucky to have you. **Els**, **Dirkje** & **Kevin**, thanks for your guidance as part of the IMPRS Research School. The many activities, but also all the small chats in the corridor greatly contributed to my PhD.

I had the privilege and pleasure of working with excellent master's students and collaborators during my PhD. **René**, thanks for teaching me how the MEG worked, even though I was supposed to be your supervisor. **Mary-Jo**, for testing many of my EEG participants, **Mircea**, for your knowledge on German, and working with me on the MEG project on non-native listeners. **Julija**, for all your efforts on the eye-tracking project. **Mitch** and **Mike**, it was a pleasure to work with you on the typing project. **Eelke**, thanks for bearing with me on the frequency-tagging project. Although I was afraid we were going to blow up the projector/it was never going to work, we made the (in my mind) impossible possible! **Gwilym**, **Mark** & **Tessa**, it was a great pleasure working on the Groot Nationaal Onderzoek. Thanks

again **Gwilym**, for introducing me to Terrace House and some other horribly bad TV shows, and for enjoying and sharing the solitude of the largest hall of the Jaarbeurs with me, when literally no one showed up for our talk on a Friday night. **Kimberley**, I'm aiming for Christmas 2030, but hey, at least we can laugh about it/ be pretty sure no one will scoop us. **James**, for such an effortless collaboration on the TMS commentary. It was a pleasure to be part of the Lowlands adventure with you, **Kimberley** & **James**, and we couldn't have pulled it off without your help, **Sybrine, Hedwig, Eelke, Julia, Marlijn, Annika & Yvonne**.

I was lucky enough to be part of not one, but two amazing labs during my PhD. In no particular order, I'd like to thank all present and past (affiliated) members of the MLC lab: **James, Francie, Louise, Gerardo, David, Judith, Kimberley, Beyza, Zeynep, Dilay, Marlou, Anita, Erce, Lilia, Emma, Els, Patrick, Marlijn, Ezgi, Vicky, Renske, Susanne, Marlijn, Tom, Huku, Christina, & Kazuki**, for the many stimulating discussions and all the fun we had. Thanks to the Neuronal Oscillations lab, **Tjerk, Barbara, Tamas, Thomas, Geoff, Iris, Hyojin, Anne, Cecilia, Rene, Jim, Tom, Lin, Haiteng, Bart, Maarten, Tzvetan, Jorn, Eelke, Tobias,** and **Ana, Sara, Katharina & Elisabet,** for brightening my days in Birmingham!

Nils, thanks for being such a caring friend, and for all your support over the last years. I still remember and cherish your patience with me in the gym/kitchen/ many moves/math lessons/bike repairs/personal crises. I am happy I could spend (almost) every Monday with you to discuss life. You're a great friend, and I hope we'll keep continuing our tradition in the years to come.

Thanks to **Gina**, my oldest Nijmegen-friend, for being the actress in all of my stimuli. Luckily that meant I could 'see' you on an almost daily basis, while you were in Seattle. I'm happy you just got back, and that I now get to see you more often again. To **Caroline**, for being such close friend even though you are now in Leipzig (although it always feels like we spoke to eachother just yesterday). To **Anne**, for all the nice dinners and wine-evenings where we discussed all things ranging from *RuPaul's Drag Race* to our personal lives, both in Lyon, Nijmegen, Seoul, or wherever we would travel. To **Alice**, for our almost weekly meetups in Nijmegen, and all the crazy stories and moments we shared. To **Francie**, for

your everlasting positivity, the best word-joke mug ever ('degraded speech is the greatest speech' – now say that out loud!), and for always being there for me both academically and non-academically, no matter what.

Thanks to **Peter**, for always having time for me, as well as sharing your love for terrible singers, cats, weird videos, tea, and many things more. To **Louise**, for always being righteous, and for your listening ear over the years. To **Nadia**, for always being charmingly honest, and for stimulating my 'burgerlijkheid'. To **David**, for showing me poetry in VR, and joining me for many bike rides. It's called 'justify', by the way. To **Bas** and **Isis**, for our time together. To **Tobias**, for showing me the nasion, to **Johanne**, for our (coffee)breaks and to **Jessica**, for just being you, and joining the Monday dinners. To **Suzanne, Rene, Patrick & Richard**, for including me in the ScienceBattle family. This is something all PhDs should do!

Lettica, thanks for giving me a 'kijkje in de keuken' of research during my bachelor studies. Our time in Stellenbosch, and working with you for two years on the project for my Honours thesis has made me enthusiastic about doing research in the first place. If it wasn't for you, I probably thought I wasn't good enough, and would never have made the switch to Cognitive Neuroscience.

Now on to my paranymphs – super bedankt voor alles (nu al)! You guys are seriously the best. **Lorijn**, we have shared all good and bad things since we started studying together. Thanks for always supporting me and believing in me no matter how crazy you thought my ideas/actions were, or how badly I cooked for you (hope you still fondly remember the quiche-turned-kaasbroodje). I have always greatly valued your honesty, opinion and unconditional friendship, and I feel very lucky to have you in my life. **James**, aka 'the Godfather of Kinect', sometimes I find it scary how well and easy we get along. From the awkward moments we call 'networking', to writing papers together, or working on a crazy (intense) experiment at Lowlands, I'm happy to have shared all of these moments with you. Thanks for bringing wine when we were working late, and for always being super positive. I hope I get to enjoy your company for many years more (especially if that means I get to see your awesome dancemoves again, by the way – yes REALLY). **René**, our coffee breaks and our digital conversations that were

solely made out of GIFs have kept me sane throughout my PhD. Thanks for always making me laugh and for relentlessly making fun of me, you definitely always knew how to get my mind off when I was stressed. I've learned a great deal from you, both academically and non-academically.

Aan **Margit** & **Ronald**, en ook **Rob**, voor alle gezelligheid samen en alle heerlijke etentjes die we samen hebben gehad. Jullie zijn vanaf het begin af aan enorm welkomend geweest, en dat waardeer ik enorm. **Wouter**, zonder jou zag mijn thesis er lang niet zo mooi uit. Enorm bedankt voor alle tijd en moeite!!

Al mijn dank en waardering gaat naar mijn ouders, **Trudie** & **Charles**, en **Astrid** & **Astrid**, voor alle goede zorgen, jullie interesse, en jullie onvoorwaardelijke steun op welke manier dan ook. Mijn allerliefste (en allergrappigste) broertje **Bram**, en natuurlijk **Manon**, bedankt voor alles. **Oma**, dankjewel voor alle hachee en erwtensoep, en het bespreken van de schaatsuitslagen. Aan **Jan** & **Marijke**, nogmaals enorm bedankt voor jullie steun tijdens mijn studie, en aan **Conny, Ton** & **Marga**, met jullie is er altijd iets te beleven. Ik heb maar wat mazzel met zo'n leuke familie.

En, natuurlijk, mijn allerliefste **Eelke**. Ik ben blij dat je ondanks mijn afgrijselijke vertolking van 'No Scrubs' tijdens karaoke in Seoul toch nog voor me bent gevallen. Ik kijk er naar uit al het moois dat nog komen gaat met je te delen. Bedankt voor al je steun, je goede zorgen, je slechte grapjes, je wijze raad, je slimmigheden, en vooral je liefde. Je bent werkelijk de man van mijn dromen.

MPI SERIES IN PSYCHOLINGUISTICS

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda van Turennout*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*
4. The open-/closed-class distinction in spoken-word recognition. *Alette Haveman*
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*
6. Gesture and speech production. *Jan-Peter de Ruiter*
7. Comparative intonational phonology: English and German. *Esther Grabe*
8. Finiteness in adult and child German. *Ingeborg Lasser*
9. Language input for word discovery. *Joost van de Weijer*
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*
11. Producing past and plural inflections. *Dirk Janssen*
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*
13. From speech to words. *Arie van der Lugt*
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*
15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*
16. Language-specific listening: The case of phonetic sequences. *Andrea Weber*
17. Moving eyes and naming objects. *Femke van der Meulen*
18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*
19. Morphology in speech comprehension. *Kerstin Mauth*
20. Morphological families in the mental lexicon. *Nivja H. de Jong*
21. Fixed expressions and the production of idioms. *Simone A. Sprenger*

22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*

23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*

24. Contextual influences on spoken-word processing: An electrophysiological approach. *Daniëlle van den Brink*

25. Perceptual relevance of prevoicing in Dutch. *Petra M. van Alphen*

26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*

27. Producing complex spoken numerals for time and space. *Marjolein Meeuwissen*

28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl J. J. K. Kemps*

29. At the same time…: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*

30. A grammar of Jalonke argument structure. *Friederike Lüpke*

31. Agrammatic comprehension: An electrophysiological approach. *Marlies Wassenaar*

32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*

33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*

34. Phonetic and lexical processing in a second language. *Mirjam Broersma*

35. Retrieving semantic and syntactic word properties. *Oliver Müller*

36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*

37. Sensitivity to detailed acoustic information in word recognition. *Keren B. Shatzman*

38. The relationship between spoken word production and comprehension. *Rebecca Özdemir*

39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*

40. The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation. *Christiane Dietrich*

41. Cognitive cladistics and the relativity of spatial cognition. *Daniel B.M. Haun*

42. The acquisition of auditory categories. *Martijn Goudbeek*

43. Affix reduction in spoken Dutch. *Mark Pluymaekers*

44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Kooijman*

45. Space and iconicity in German Sign Language (DGS). *Pamela Perniss*

46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*

47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*

48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*

49. Phoneme inventories and patterns of speech sound perception. *Anita Wagner*

50. Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*

51. A grammar of Savosavo, a Papuan language of the Solomon Islands. *Claudia Wegener*

52. Prosodic structure in speech production and perception. *Claudia Kuzla*

53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*

54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*

55. Processing the fine temporal structure of spoken words. *Eva Reinisch*

56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*

57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*

58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*

59. Lexical interactions in non-native speech comprehension: Evidence from electro-encephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*

60. Processing casual speech in native and non-native language. *Annelie Tuinman*

61. Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Robinson*

62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*

63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*

64. The meaning and use of ideophones in Siwu. *Mark Dingemanse*

65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*

66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*

67. The relevance of early word recognition: Insights from the infant brain. *Caroline Junge*

68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias J. Sjerps*

69. Structuring language. *Contributions to the neurocognition of syntax. Katrien R. Segaert*

70. Infants' appreciation of others' mental states in prelinguistic communication: A second person approach to mindreading. *Birgit Knudsen*

71. Gaze behavior in face-to-face interaction. *Federico Rossano*

72. Sign-spatiality in Kata Kolok: how a village sign language of Bali inscribes its signing space. *Conny de Vos*

73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*

74. Lexical processing of foreign-accented speech: Rapid and flexible adaptation. *Marijt Witteman*

75. The use of deictic versus representational gestures in infancy. *Daniel Puccini*

76. Territories of knowledge in Japanese conversation. *Kaoru Hayano*

77. Family and neighbourhood relations in the mental lexicon: A cross-language perspective. *Kimberley Mulder*

78. Contributions of executive control to individual differences in word production. *Zeshu Shao*

79. Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing. *Patrick van der Zande*

80. High pitches and thick voices: The role of language in space-pitch associations. *Sarah Dolscheid*

81. Seeing what's next: Processing and anticipating language referring to objects. *Joost Rommers*

82. Mental representation and processing of reduced words in casual speech. *Iris Hanique*

83. The many ways listeners adapt to reductions in casual speech. *Katja Poellmann*

84. Contrasting opposite polarity in Germanic and Romance languages: Verum Focus and affirmative particles in native speakers and advanced L2 learners. *Giuseppina Turco*

85. Morphological processing in younger and older people: Evidence for flexible dual-route access. *Jana Reifegerste*

86. Semantic and syntactic constraints on the production of subject-verb agreement. *Alma Veenstra*

87. The acquisition of morphophonological alternations across languages. *Helen Buckler*

88. The evolutionary dynamics of motion event encoding. *Annemarie Verkerk*

89. Rediscovering a forgotten language. *Jiyoun Choi*

90. The road to native listening: Language-general perception, language-specific input. *Sho Tsuji*

91. Infants'understanding of communication as participants and observers. *Gudmundur Bjarki Thorgrímsson*

92. Information structure in Avatime. *Saskia van Putten*

93. Switch reference in Whitesands. *Jeremy Hammond*

94. Machine learning for gesture recognition from videos. *Binyam Gebrekidan Gebre*

95. Acquisition of spatial language by signing and speaking children: a comparison of Turkish sign language (TID) and Turkish. *Beyza Sümer*

96. An ear for pitch: on the effects of experience and aptitude in processing pitch in language and music. *Salomi Savvatia Asaridou*

97. lncrementality and Flexibility in Sentence Production. *Maartje van de Velde*

98. Social learning dynamics in chimpanzees: Reflections on (nonhuman) animal culture. *Edwin van Leeuwen*

99. The request system in Italian interaction. *Giovanni Rossi*

100. Timing turns in conversation: A temporal preparation account. *Lilla Magyari*

101. Assessing birth language memory in young adoptees. *Wencui Zhou*

102. A social and neurobiological approach to pointing in speech and gesture. *David Peeters*

103. Investigating the genetic basis of reading and language skills. *Alessandro Gialluisi*

104. Conversation Electrified: The Electrophysiology of Spoken Speech Act Recognition. *Rósa Signý Gisladottir*

105. Modelling Multimodal Language Processing. *Alastair Smith*

106. Predicting language in different contexts: The nature and limits of mechanisms in anticipatory language processing. *Florian Hintz*

107. Situational variation in non-native communication. *Huib Kouwenhoven*

108. Sustained attention in language production. *Suzanne Jongman*

109. Acoustic reduction in spoken-word processing: Distributional, syntactic, morphosyntatic, and orthographic effects. *Malte Viebahn*

110. Nativeness, dominance, and the flexibility of listening to spoken language. *Laurence Bruggeman*

111. Semantic specificity of perception verbs in Maniq. *Ewelina Wnuk*

112. On the identification of FOXP2 gene enhancers and their role in brain development. *Martin Becker*

113. Events in language and thought: The case of serial verb constructions in Avatime. *Rebecca Defina*

114. Deciphering common and rare genetic effects on reading ability. *Amaia Carrión Castillo*

115. Music and language comprehension in the brain. *Richard Kunert*

116. Comprehending Comprehension: Insights from neuronal oscillations on the neuronal basis of language. *Nietzsche H.L. Lam*

117. The biology of variation in anatomical brain asymmetries. *Tulio Guadalupe*

118. Language processing in a conversation context. *Lotte Schoot*

119. Achieving mutual understanding in Argentine Sign Language. *Elizabeth Manrique*

120. Talking Sense: the behavioural and neural correlates of sound symbolism. *Gwilym Lockwood*

121. Getting under your skin: The role of perspective and simulation of experience in narrative comprehension. *Franziska Hartung*

122. Sensorimotor experience in speech perception. *Will Schuerman*

123. Explorations of beta-band neural oscillations during language comprehension: Sentence processing and beyond. *Ashley Lewis*

124. Influences on the magnitude of syntactic priming. *Evelien Heyselaar*

125. Lapse organization in interaction. *Elliott Hoey*

126. The processing of reduced word pronunciation variants by natives and foreign language learners: Evidence from French casual speech. *Sophie Brand*

127. The neighbors will tell you what to expect: Effects of aging and predictability on language processing. *Cornelia Moers*

128. The role of voice and word order in incremental sentence processing. *Sebastian Sauppe*

129. Learning from the (un)expected: Age and individual differences in statistical learning and perceptual learning in speech. *Thordis Neger*

130. Mental representations of Dutch regular morphologically complex neologisms. *Laura de Vaan*

131. Speech production, perception, and input of simultaneous bilingual preschoolers: Evidence from voice onset time. *Antje Stoehr*

132. A holistic approach to understanding pre-history. *Vishnupriya Kolipakam*

133. Characterization of transcription factors in monogenic disorders of speech and language. *Sara Busquets Estruch*

134. Indirect request comprehension in different contexts. *Johanne Tromp*

135. Envisioning Language - An Exploration of Perceptual Processes in Language Comprehension. *Markus Ostarek*

136. Listening for the WHAT and the HOW: Older adults' processing of semantic and affective information in speech. *Juliane Kirsch*

137. Let the agents do the talking: on the influence of vocal tract anatomy on speech during ontogeny and glossogeny. *Rick Janssen*

138. Age and hearing loss effects on speech processing. *Xaver Koch*

139. Vocabulary knowledge and learning: Individual differences in adult native speakers. *Nina Mainz*

140. The face in face-to-face communication: Signals of understanding and non-understanding. *Paul Hömke*

141. Person reference and interaction in Umpila/Kuuku Ya'u narrative. *Clair Hill*

142. Beyond the language given: The neurobiological infrastructure for pragmatic inferencing. *Jana Bašnáková*

143. From Kawapanan to Shawi: Topics in language variation and change. *Luis Miguel Rojas-Berscia*

144. On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions. *Linda Drijvers*

MAX PLANCK

On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions

Linda Drijvers

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

VISITING ADDRESS
Wundtlaan 1
6525 XD Nijmegen
The Netherlands

POSTAL ADDRESS
P.O. Box 310
6500 AH Nijmegen
The Netherlands

CONTACT
T +31(0)24 3521 911
F +31(0)24 3521 213
E info@mpi.nl
Twitter @MPI_NL
www.mpi.nl

DONDERS INSTITUTE

CLS | Centre for Language Studies
Radboud University

MPI SERIES 144