Gaze allocation during gestural enhancement of degraded speech in native and non-native listeners

Julija Vaitonytė
Gaze allocation during gestural enhancement of degraded speech in native and non-native listeners

Master thesis
Julija Vaitonytė

Supervisors: Prof. dr. Asli Özyürek and Linda Drijvers
Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Second reader: Dr. Susanne Brouwer
Centre for Language Studies, Radboud University Nijmegen
Nijmegen, The Netherlands

Student number: 4608089
Program: Research Master in Language and Communication
Faculty of Arts, Radboud University Nijmegen
Submission date: 30th August, 2018
Acknowledgements

This thesis grew out of a research project embedded in the Multimodal Language and Cognition Lab at the Max Planck Institute for Psycholinguistics. Thus, I wanted to thank the leader of the MLC group, prof. dr. Asli Özyürek, for the opportunity to work in her lab and develop hands-on skills and relevant expertise.

My special thanks also go to Linda Drijvers, who supervised me on a daily basis. Thank you for challenging me in all possible ways and encouraging great independence. The lessons learned with you will stick with me.
Abstract

In natural communication, humans mostly focus on the speaker’s face and devote very little visual attention to gestures, which play a significant role in language comprehension. To date, explicit attention to gestures has only been studied in optimal listening conditions, and only with native speakers of a language. We investigated how native and non-native listeners allocate attention to speech relevant visual information (visible speech and iconic gestures) when speech is both clear and degraded. We recorded participants’ eye movements while they watched videos of an actress who uttered a high frequency Dutch verb (e.g., rijden, to drive in English) in either clear or 6-band noise-vocoded speech (moderate speech signal degradation), and who also either performed or not the accompanying iconic gesture. We found that the face was the locus of sustained attention in clear and degraded speech, suggesting gestures, when they were presented, attracted little overt attention in (non-)native listeners. While both listener groups gazed more to gestures in clear than degraded speech, overall, non-native listeners gazed significantly more to gestures compared to native listeners. Taken together, the results suggest explicit attention to the face is functional for the extraction of information from visual speech articulators while gestural information can be extracted peripherally. Yet, the fact that gestures attract more visual attention from non-native listeners than would be normally expected may suggest that non-native listeners conceive of visual semantic information conveyed by gestures as an especially rich source of information that can aid language comprehension under both good and suboptimal listening conditions.
# Table of Contents

Acknowledgements

Abstract

1. Introduction .................................................................................................................................................. 6
2. Present study .................................................................................................................................................. 10
3. Methods ....................................................................................................................................................... 12
   3.1 Participants ........................................................................................................................................... 12
   3.2 Materials ............................................................................................................................................. 13
   3.3 Experimental paradigm ....................................................................................................................... 14
   3.4 Procedure ............................................................................................................................................ 14
   3.5 Data analysis ........................................................................................................................................ 16
      3.5.1 Behavioral data ............................................................................................................................ 16
      3.5.2 Ocular data ................................................................................................................................... 16
      3.5.3 Cluster-based Permutation Tests .............................................................................................. 17
      3.5.4 Comparing mean proportion of looking between native and non-native listeners .................. 18
4. Results ......................................................................................................................................................... 19
   4.1 Behavioral results .................................................................................................................................. 19
      4.1.1 Reaction times ............................................................................................................................ 19
      4.1.2 Accuracy of responses .............................................................................................................. 20
      4.1.3 Error rates ................................................................................................................................... 22
   4.2 Ocular results ........................................................................................................................................ 23
      4.2.1 Visual attention to the face in native and non-native listeners: within group comparisons ......... 23
      4.2.2 Visual attention to the face in native and non-native listeners: between-group comparisons ....... 27
      4.2.3 Visual attention to gestures in native and non-native listeners: within-group comparisons ......... 28
      4.2.4 Visual attention to gestures in native and non-native listeners: between-group comparisons ....... 31
      4.2.5 Relationship between the amount of overt attention to gestures and performance on a cued-recall task in native and non-native listeners ........................................................................... 32
5. Discussion .................................................................................................................................................... 34
   5.1 Prolonged attention to the face in degraded speech in both listener groups ................................. 35
5.2 More explicit attention to gestures in non-native listeners compared to native listeners under good and adverse listening conditions .......................................................... 36

5.3 Language proficiency as the modulator of overt attention to gestures during multimodal language comprehension ................................................................. 37

6. Conclusion ......................................................................................................................... 39

References ............................................................................................................................ 40
1. Introduction

In natural communication, language comprehension is intrinsically multisensory due to the involvement of auditory (speech) and visual signals, such as visible speech, which is a by-product of speech articulation (Yehia et al., 1998), and hand gestures (McNeill, 1992; 2000). The fact that listeners continuously integrate auditory and visual information when comprehending language is most apparent when listening conditions become challenging (but see Rosenblum, 2008 and Özyürek et al., 2014 for reviews on multimodal language comprehension when speech is clear). When speech intelligibility is compromised by noise or signal degradation, it has been shown, the information derived from visible speech (Eber, N., 1975; Drijvers & Özyürek, 2017; MacLeod & Summerfield, 1987; Ross et al., 2007; Sumby & Pollack, 1954), and that conveyed by gestures (Drijvers & Özyürek, 2017; Holle et al., 2010) can facilitate language comprehension. Recently, the role that both visual articulators, visible speech and gestures, play in aiding degraded speech comprehension has been studied by presenting them in a joint context (Drijvers & Özyürek, 2017).

The aim of the study by Drijvers & Özyürek (2017) was to investigate whether iconic gestures, which illustrate object attributes, actions, and space (McNeill, 1992) and can complement a spoken message on a semantic level (Beattie & Shovelton, 1999; 2006; Gerwing & Allison, 2009) can, on top of visible speech, facilitate comprehension of degraded speech. Native listeners of Dutch were presented with videos, in which a speaker uttered a highly frequent action verb (e.g., drinken, to drink in English) that was either clear or noise-vocoded by means of using 6-band (moderate speech degradation) and 2-band noise-vocoding (severe degradation). Besides adjusting the clarity of speech input, the degree of visual information was manipulated in three ways: the speaker’s lips were either blurred, or visible speech was present, or visible speech and a gesture were present. It was revealed participants were most able to accurately identify the verb when information from both visible speech and gestures was available, suggesting visible speech and iconic gestures together can aid comprehension of degraded speech to a greater extent than when only one of the visual inputs is present. This effect (termed the double enhancement of degraded speech) was largest when speech was 6-band noise-vocoded, suggesting moderate rather than severe speech degradation represents conditions under which integration of the auditory with the visual streams has a maximal effect on comprehension. That is, comprehension is facilitated not on the basis of solely relying on the visual information, but through integrating the auditory cues that are still moderately reliable with the two streams of visual information.

The degree to which each of the visual articulators help disambiguate degraded speech when one is a native listener may, however, differ from that experienced by a non-native listener. Non-native listeners’ speech recognition being overall more hindered by the speech signal distortions compared to native listeners (Mayo et al., 1997) allows to assume that non-native listeners may also exploit visual information differently from natives. For instance, non-native listeners may rely more heavily on visual semantic information when speech is degraded. Listeners with a low proficiency in a second language have been shown to draw substantially on iconic gestures in order to boost language comprehension even when speech is clear (Dahl & Ludvigsen, 2014; Sueyoshi & Hardison, 2005). The line of research that involved relatively proficient second language learners suggests that they use phonological information from visible speech to aid non-
native phoneme comprehension (Hazan et al., 2006). Yet, at the sentence level, *highly-proficient* non-native listeners make use of the semantic contextual-information in order to resolve the impoverished auditory cues to a lesser degree than native listeners (Bradlow & Alexander, 2007, Mayo et al., 1997). It has been proposed that the benefits of the contextual information in aiding language comprehension are dependent on the quality of the speech signal (Bradlow & Alexander, 2007). Essentially, contextual cues need to be coupled to the acoustic cues which in turn requires experience with the sound inventory of the language. Under adverse listening conditions, due to the reduced experience with the sound system of the target language, non-natives are also limited in their capacity to bind the contextual cues to the acoustic information relative to native listeners.

On the basis of these observations, a subsequent study by Drijvers & Özyürek (resubmitted) set out to answer whether non-native listeners (i.e., native speakers of German with high proficiency in Dutch) experienced the added benefit from visual semantic information on top of the enhancement conferred by visible speech by means of employing the same design and stimuli as described above. In a non-native listener study, Drijvers & Özyürek observed a pattern of results that was highly reminiscent of those reported with native listeners (Drijvers & Özyürek, 2017), only reduced in scale. What is of note, despite that the interpretation must be cautious when making direct comparisons between native and non-native listeners, was seemingly more reliance on the part of non-native subjects on gestural information when speech was 2-band noise-vocoded. When speech was severely degraded, non-native listeners were mostly able to aid auditory signal comprehension with the semantic information extracted from gestures rather than from both types of visual articulators. It has been proposed that when the auditory input becomes increasingly more ambiguous (going from clear to 2-band noise vocoded speech), it is also more and more difficult to map visible speech information to the auditory cues that can still be recovered from the speech signal that is degraded, which, as argued by Drijvers & Özyürek (resubmitted), may be attributed to the level of language proficiency. That the extent to which a non-native listener derive benefit from visual information is reliant on language proficiency is in line with previous work showing that increased auditory proficiency was associated with a greater audio-visual benefit on a phonemic discrimination task (Hazan et al., 2006). That is, the higher the level of language proficiency of a second-language learner, the greater the degree of audiovisual benefit that they experienced. It is, however, important to bear in mind two aspects. First, the study by Drijvers & Özyürek (resubmitted) employed non-native listeners with high language proficiency. Second, highly-proficient non-native listeners did benefit from both visual articulators (visible speech and gestures) when speech was 6-band noise-vocoded. This further implies that highly-proficient non-native listeners integrate iconic gestures with speech that is moderately degraded.

The idea of iconic gesture integration with clear and degraded speech in native and non-native listeners has been recently examined by using electrophysiology (Drijvers & Özyürek, 2018). Speech-gesture integration by means of EEG is commonly tested by investigating the N400 ERP component that is sensitive to semantic unification operations (Holle & Gunter, 2007; Özyürek et al., 2007). Typically, researchers employ a paradigm, in which visual semantic information conveyed by gestures can be either congruent or incongruent with the spoken message. The increased load of integrating conflicting sources of information on a semantic level (i.e., when gesture mismatches speech) is reflected in enhanced amplitudes of the N400 component. The more
neural resources are needed to reconcile the incongruent visual and auditory inputs, the more negative the amplitudes, and this difference between the N400 amplitudes observed in the congruent vs incongruent conditions is the N400 effect (Kutas & Federmeier, 2014).

For an EEG study, Drijvers & Özyürek (2018) recruited native and highly proficient non-native listeners of Dutch and adapted a subset of the stimuli taken from their previous work (i.e., videos with clear or 6-band noise-vocoded speech, and either a matching or a mismatching gesture). Highly-proficient non-native listeners were selected to participate in the study as participants with a low proficiency in a language would potentially be only able to pick up information from gestures, thus preventing to study gestural enhancement of degraded speech. It was shown that for native listeners of Dutch the N400 effect was present in both clear and degraded speech. Yet the N400 effect was larger when speech was clear than when it was degraded, implying semantic manipulations are easier to observe when the speech signal is of good quality. Regarding non-native listeners, the N400 effect was larger in clear speech than degraded speech, which was driven by the absence of the N400 effect when speech was degraded. Curiously though, the N400 effect in clear speech that was observed for non-native listeners was larger than the one evidenced in the native listener group. Compared to native listener group, the heightened susceptibility of non-native listeners to noticing speech-gesture mismatches when speech is clear may be indicative of non-native listeners allocating more visual attention to gestures. In degraded speech, however, the failure to spot speech-gesture incongruences may be attributed to non-native listeners’ increased difficulty in extracting auditory cues and combining them with visual semantic information from gestures, which, most likely, stems from a non-native language ability. Although being highly proficient in the target language, subjects are still hindered by their non-native listener status, implying less experience as compared to native listeners with the phonological system of the language.

On the basis of the electrophysiological results, there is reason to propose the existence of quantitative differences in terms of how native and non-native listeners allocate visual attention when they disambiguate degraded speech. The present study, largely motivated by these potential differences in visual attention, sought to understand how visual attention unfolds in a joint context under adverse but also good listening conditions, i.e., when (non)-native listeners integrate the auditory speech with the information derived from visual speech and iconic gestures. In terms of visual attention under suboptimal auditory conditions, the studies so far have been conducted with the observers who were presented with the face only instead of the face together with all the body features below the neck. It has been revealed that during audiovisual speech perception with the diminished quality of the auditory channel, attention to the lower part of the face (mouth) increases (Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998). This suggests that perceivers are drawn to a spatial location that they may consider will offer information helpful in resolving uncertainty about the speech signal. Regarding attention allocation to gestures, to-date it has only been studied when speech is clear. It has been indicated that observers (native listeners of a language) direct very little explicit attention to gestures (Gullberg & Holmqvist, 1999, 2006; Gullberg & Kita, 2009), and the face is the default location for sustained visual attention. Listeners/viewers explicitly gaze at the speaker’s gestures only around 0.5% of the time during face-to-face interaction as well as on video while the face is attended over 90% of the time.
The likelihood of allocating overt attention to gestures generally increases under two circumstances, one of which is autofixations (a speaker gazing at their own gesture) while the other is gestural holds (a hold is a temporary cessation of gestural movement in space) (Gullberg & Kita, 2009). In the first instance, listener/viewer is drawn to look at the speaker’s gesture because the latter breaks the mutual eye contact, further suggesting overt attention to gestures in this case is subject to social factors. Regarding gestural holds, the reason for explicit attention to gestures is modulated by physical factors, in that, as has been argued by Gullberg & Kita (2009), gesture with a hold may challenge peripheral vision. It has been maintained (Gullberg & Kita, 2009; Beattie, 2016, p. 131) that when gesture temporarily ceases, the observer needs to fixate gesture as peripheral vision, being very suitable for processing motion information, is no longer sufficient for extracting information from that gesture.

It remains to be seen whether there are other factors than the social and physical aspects that drive explicit attention to gestures more than would be generally expected when listeners understand speech in a multimodal context. We refer here to those other factors that may modulate attention allocation as the external (speech degradation) and internal (non-nativeness) characteristics during multimodal language comprehension. It is worth stressing that this is how language comprehension unfolds in everyday situations, in which listening conditions are commonly less than ideal yet listeners can aid language comprehension by coupling to the speech input visual cues that naturally unfold during face-to-face communication.
2. Present study

The present study aimed to examine how native and non-native listeners allocate attentional resources to both types of visual input, visible speech and iconic gestures, when comprehending language under adverse listening conditions but also when speech is clear. Visual attention was studied by means of employing the eye tracking methodology, which permits recording eye movements with high temporal resolution. Through measuring eye position, inferences about overt attention allocation can be made, in that attention is tightly linked to gaze behavior (Hoffman & Subramaniam, 1995). While it is generally possible to allocate attention without moving one’s eyes, it is more common, and, also argued, more efficient (Eckstein et al., 2017) to explicitly attend to a spatial location. One is usually looking at what they are processing thus measuring where and when observers look permits a moment-by-moment assessment of thought processes. In other words, eye tracking is an unintrusive and inexpensive method that facilitates a window into one’s cognition, through which behavior on a cognitive task can be explained.

In this study, participants’ eye movements were recorded while they watched videos of an actress who uttered a highly frequent Dutch verb (e.g., rjden, to drive in English) in either clear or 6-band noise-vocoded speech, and who also either performed or not the accompanying iconic gesture. The study thus employed four conditions: clear speech (C), clear speech + gesture (CG), degraded speech (D), degraded speech + gesture (DG). Participants were given a cued-recall task, in that after having watched each video they had to choose from the four answer options which verb they think they heard. In addition to ocular measures (the proportion of looking to different areas of interest as defined for face and body), the behavioral data (accuracy and reaction times, RTs) were collected and analyzed.

We strived to answer not only where and how much (non)-native listeners look when speech is clear and when it is degraded but also whether and how gesture presentation can impact gaze behavior and thus allocation of overt attention to gestures in clear and degraded speech. It was also of interest whether overt attention to gestures is associated with behavioral performance reflected in accuracy and RTs. For behavioral performance, it was expected, in line with previous studies that native (Drijvers & Özyürek, 2017; Drijvers, Özyürek & Jensen, 2018), and non-native listeners (Drijvers & Özyürek, resubmitted) would show gestural enhancement of degraded speech (i.e., participants will be faster and more accurate in verb identification when gestures are present than when they are absent). It was also predicted that native listeners would outperform non-native listeners in terms of accuracy of responses and RTs.

Regarding ocular measures, we hypothesized that in the absence of gestures, and irrespective of the clarity of speech, native and non-native listeners would gaze solely at the face. We reasoned that the face would receive explicit attention due to participants’ learned experiences of the face conferring valuable interpersonal information (including but not limited to emotions and mental states, Argyle & Cook, 1976; Baron-Cohen et al., 2001), and linguistically relevant cues helpful in turn-taking (Ho et al., 2015) and audiovisual speech processing when speech is clear (Rosenblum, 2008) and when it is compromised by signal degradation and noise (Drijvers & Özyürek, 2017; Sumby & Pollack, 1954).

When gestures are presented in clear and degraded speech, we expected that they would attract attention to some extent from both native and non-native listeners, in that “movement is
able to bring about reflexive eye movements” as put by Holmqvist et al. (2011, p. 82). Another reason to expect attention allocated to gestures, and therefore more attention to gestures from both groups in clear than degraded speech, is that in clear speech the attentional system may not be taxed by resolving poor auditory cues. Yet, we proposed that in clear speech there would still be more attention allocated to the face than gestures on the basis of previous work showing that the face is the main locus of attention even when gestures are present (Gullberg & Holmqvist, 1999, 2006; Gullberg & Kita, 2009). On the basis of the EEG results from Drijvers & Özyürek (2018), we expected that non-native listeners would allocate more overt attention to gestures than native listeners.

With respect to gestures being presented in degraded speech, it was hypothesized that native and non-native listeners would gaze at the face, based on research showing increased attention to the lower part of the face for word identification under suboptimal listening conditions (Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998), but also gestures, which convey visual semantic information that help in disambiguation of degraded speech (Drijvers & Özyürek, 2017). Regarding differential gaze behavior in degraded speech + gesture condition between native and non-native listeners, no clear predictions were made due to two possible scenarios. One possibility was that in degraded speech + gesture condition gaze behavior of the two groups would be highly similar. That is, both native and non-native listeners would devote an equally small amount of explicit attention to gestures in degraded speech + gesture condition because they would largely be focusing on the face with the attempt to lip-read. Alternatively, non-native listeners might devote more attention to gestures presented in degraded speech than native listeners. The N400 effect being absent in the non-native listener group does not rule out the possibility that non-native listeners might allocate more visual attention to gestures even when speech is degraded, only that mapping visual semantic information to the auditory stream is more difficult when one is a non-native listener.
3. Methods

3.1 Participants

A total of 41 participants took part in the experiment, of whom 20 were native speakers of Dutch (7 males, mean age = 26.00, SD = 7.58) and 21 were native speakers of German (5 males, mean age = 23.05, SD = 2.62) with high proficiency in Dutch. For the final analyses, the data of 40 participants were used due to having discarded the data of one German participant (female) who failed to meet the inclusion criteria as described below. Participants reported no neurological, language, hearing or motor disorders, and had either normal or corrected-to-normal vision. Dutch participants were recruited through Radboud Research Participation System (SONA) while the recruitment of German participants was carried out through advertising the experiment on social media as well as via personal contacts. All participants gave informed written consent and received a monetary compensation for their participation. The two groups did not differ in age (t(38) = 1.64, p = 0.108) and both Dutch and German participants had obtained education at the university level. The majority of participants in both groups were, at the time of testing, students at Radboud University Nijmegen, except for several participants who reported working.

We recruited German participants on the basis of the following criteria: having obtained a score of 60% or higher on the LexTALE test (Lemhöfer & Broersma, 2012), having lived/studied in the Netherlands for at least one year, having used Dutch on a regular basis (at least once a week), and having started to acquire Dutch after the age of 12. The female participant whose data were not used for the analyses reported having started to acquire Dutch at the age of 5 (the start of acquisition of the Dutch language for the rest of non-native participants ranged between 12 and 22 years, mean age = 18.25, SD = 2.80). There were two German participants who reported having lived in the Netherlands for 7 and 8 months. Yet, given their high scores on the LexTALE test (77.5% and 80.0% respectively), the data of these participants were used for further analyses.

The LexTALE test (Lemhöfer & Broersma, 2012), which all German participants completed online prior to coming to the lab, is a lexical decision test that has been used to assess vocabulary knowledge and language proficiency. We specifically asked for a score of ≥60% on this test as we wanted to ensure that our non-native listener group was indeed proficient in the Dutch language. A score of 60% and higher is known to correlate with B2 level and higher. A foreign language user with level B2, as based on the Common European Network for Languages (Council of Europe, 2001), has the fluency to communicate with native speakers of a language effortlessly and spontaneously in a range of contexts.

At the end of the testing session, non-native listeners took part in the adapted version of the LexTALE test (further referred to as the LexTALE II) that consisted of 60 items (verbs used in the videos and pseudoverbs), which we administered in order to determine that German participants were familiar with the verbs presented in the videos. We gave Dutch participants to fill in both the LexTALE test and the LexTALE II in order to ensure that all participants followed the same experimental procedure. As we expected, significant differences between the two groups were found when native listeners’ scores on the LexTALE test (M = 91.91, SD = 6.07) were compared with those of non-native listeners (M = 78.12, SD = 9.21), with the native listener group
scoring significantly higher on the test, \( t(38) = 5.58, p < .001 \). Similarly, native listeners obtained significantly higher scores on the LexTALE II (\( M = 96.33, SD = 2.73 \)) than non-native listeners (\( M = 78.59, SD = 8.84 \), \( t(38) = 8.56, p < .001 \)).

### 3.2 Materials

The stimuli used in the present experiment were drawn from a larger set of videos employed in a previous study (for a detailed description of the stimuli, see Drijvers & Özyürek, 2017). All videos were recorded with a JVC GY-HM100 camcorder and were on average 2 seconds in length. Specifically, the stimulus material consisted of 160 videos that displayed a woman, who uttered a highly frequent Dutch verb (e.g., *rijden*, *to drive* in English) and either performed the accompanying iconic gesture (\( n = 80 \)) or not.

In each video, the woman, who wore neutrally colored clothing, was standing in the middle of the screen in front of a unicolored background with her hands placed by her sides. In the stimuli where gesture was absent, the woman was standing still throughout the entire video and would utter a verb without any other observable bodily motion except for visible speech. In each video, speech would start at around 680 msec and last until approximately 1500 msec.

In those videos where the woman produced gesture (see Fig. 1), gesture preparation would always start at around 120 msec, with the most meaningful part of the gesture (stroke) occurring on average at 550 msec. Gesture retraction would start at around 1380 msec and gesture would end at approximately 1780 msec. Note, stroke always preceded the speech onset by around 130 msec, which was done in order to maximize the overlap between the stroke and the speech segment. As shown by previous research, speech and gestures are highly synchronized, with gesture typically preceding the linguistic equivalent (Ferré, 2010; Habets et al., 2011; Nobe, 2000). Yet, as also indicated by Habets et al. (2011), multimodal integration is precluded if gesture is temporally too apart from the speech segment, suggesting speech and gesture have to co-occur within a certain time interval. The study by Habets et al., which employed EEG to study speech-gesture synchronization, for instance, showed that gestures were not integrated if they preceded speech by 360 msec.

![Diagram](image)

*Figure 1. Illustration of the structure of the video stimuli, adapted from Drijvers & Özyurek (2018)*
It is also important to stress that gestures were not unambiguous without speech. As pointed out in the original study by Drijvers & Özyürek (2017), who pretested the video stimuli, the mean rate of gestures being identified without speech was 59% across all gesture videos. During the pretest, a group of participants were presented with silent videos displaying only gestures without the associated auditory information and were asked to write down the verb gesture depicted. After this, the participants were given the original verb that the authors associated with gesture displayed in the video and were asked to indicate the degree to which they thought the verb matched the gesture using a 7-point scale. This procedure was carried out in order to create ecologically valid stimuli as in natural communication gestures typically need speech to be disambiguated (Kraus et al., 1991; 1996; Hadar & Pinchas-Zamir, 2004). In other words, the semantic contribution that gestures make to the overall message would not be sufficient on its own for effective comprehension of that message.

In order to create videos in which speech was 6-band noise-vocoded (n = 80), the auditory sound files were first intensity-scaled to 70 dB and de-noised in Praat (Boersma & Weenink, 2015). By means of using a custom-made Praat script, a 6-band noise-vocoding was added to each audio file, which was then recombined with its corresponding video file in Adobe Premiere Pro. Noise-vocoding degrades the spectral content of the speech signal while the temporal envelope is pertained (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). The speech signal is generally intelligible up to a certain degree and more bands correspond to a more intelligible speech signal. In the previous experiment (see Drijvers & Ozyürek, 2017), a 6-band noise-vocoding level was identified as the optimal speech degradation level, at which contribution of iconic gestures to the enhancement of degraded speech is largest. Due to this reason, videos with 6-band noise vocoded speech were chosen for the present experiment.

### 3.3 Experimental paradigm

The study employed two within-subject factors and a between-subject factor (listener status: native vs non-native). The two within-subject factors, speech and gesture, each consisted of two levels: speech could be either clear or degraded (i.e., 6-band noise-vocoding was used, which represents moderate speech signal degradation) whereas gesture could be either present or absent. Thus each participant saw videos in four conditions: clear speech condition (C), clear speech + gesture condition (CG), degraded speech condition (D), and degraded speech + gesture condition (DG). All conditions consisted of 40 unique videos with unique verbs and gestures. Each participant saw a total of 160 videos.

### 3.4 Procedure

The experimental procedure began with first the participant filling out the consent form and, if necessary, performing the LexTALE test (native speakers of Dutch were given a paper version of the LexTALE test upon the arrival to the experiment whereas German participants completed the LexTALE test online prior to coming to the lab). The participant then received the instructions both in word and writing, and was set up for the experiment. Participants were instructed to watch videos and indicate by pressing one of the four buttons on the button box which word they think they heard in the video. The task was a cued-recall task, in that participants were able to choose between the four answer options presented on the screen (a, b, c and d), which
consisted of the correct verb (the target), phonological competitor, semantic competitor or an unrelated verb. This was done in order to ensure participants indeed were paying attention to the videos.

Participants were tested individually in a dimly-lit soundproof booth. They were seated around 70 cm from the computer screen and fitted with headphones. The stimuli, which were randomized across participants, were presented on a 1650x1080 monitor using the Experiment Builder (SR Research Ltd.). Eye movements were recorded with an SR Research Eyelink 1000 Tower mount system that has a sampling rate of 2000 Hz. Movement was minimized by using the chin rest and forehead support. In addition, participants sat in a stable chair with back support.

Participants first saw brief instructions on the screen followed by a blank screen after which the calibration and validation procedure was performed. The in-built 9-point calibration routine was used. If necessary, calibration was repeated several times in order to ensure optimal accuracy. Calibration was considered successful if the average discrepancy between the corresponding point and the participant’s gaze was <0.5-0.75º.

All trials started with a fixation cross presented centrally on a white screen for 1000 msec, after which the video played for 2000 msec (see Fig. 2). Having watched the video, participants would see a blank screen for 1000 msec and would then perform the cued-recall task. Participants had 5000 msec to submit their answer, the time limit after which a new trial would begin.

The testing session lasted on average 20-23 minutes, no break was administered. Participants’ eye gaze was tracked throughout the entire experiment. Having completed the eye tracking part, the participants took part in the LexTALE II.
3.5 Data analysis

3.5.1 Behavioral data

In order to test for difference in accuracy of responses and reaction times, two separate three-way mixed repeated measures ANOVAs with speech (clear, degraded) and gesture (present, absent) as within-subject factors and language (native, non-native) as a between-subject factor were run.

We also conducted a Chi-square test in order to determine whether a particular type of errors (phonological, semantic or other) occurred more frequently than would be expected on the basis of a given condition.

3.5.2 Ocular data

All pre-processing and analysis of the ocular data was done in R by means of using an open-source package eyetrackingR (http://www.eyetracking-r.com/). First, the data were segmented into trials starting from 500 msec before and ending 500 msec after the onset of the video. The data were then checked for trackloss. All trials with trackloss >25% were removed. The average number of contributed trials per participant in the native listener group was 121.25 while in the non-native listener group it was 137.6.

For each trial two areas of interest (AOI) were defined: the face and body (see Fig 3, left panel). Gazes that fell outside of these areas were excluded from further analysis. For each participant the proportion of looking at each AOI was calculated and served as the dependent variable. In order to test for differences in degraded speech vs clear speech condition in specifically looking to the lower part of the face as it was expected that participants in both of these conditions would gaze solely at the face, we defined an additional AOI: the mouth area (see Fig 3, right panel).

Figure 3. Illustration of the different areas of interest (AOIs). The panel on the left and the right depict a still image from one of the videos employed in the study with an outline of the AOIs: the face (upper) and body (lower) while the right panel displays the additional AOI: the mouth.
3.5.3 Cluster-based Permutation Tests

In order to test for differences between conditions within and between participants as well as for the effect of gestural enhancement, non-parametric cluster-based permutation tests (dependent and independent t-tests) were used (Maris & Oostenveld, 2007; Nichols & Holmes, 2002). While in this study, clustering was performed along one dimension (temporal), cluster-based permutation procedure allows to identify significant clusters between two conditions in a contrast along a spatial and/or temporal dimension. For our purposes, cluster-based permutation procedure was especially apt as we were interested not only whether gaze behavior differs between the different conditions but also where in time significant differences between the two contrasted conditions emerge, which can then be explained on the basis of the video structure depicted in Fig. 1.

First, the proportion of looking to the relevant AOI was calculated across participants per timestamp (i.e., 0-2000 msec). Then, a t statistics was calculated for every trial of two conditions of the relevant comparison. All trials whose t value was higher than a pre-determined threshold (97.5 quantile) were selected and clustered on the basis of their temporal adjacency. By summing the t values within each cluster, cluster-level statistics was calculated. Then, by randomly reassigning (1,000) times the two conditions and their corresponding values, the permutation distribution was created. Clusters that fell in the highest or lowest 2.5th percentile of the distribution were considered significant (see Maris and Oostenveld (2007)).

We assessed whether looking behavior was different between the two groups by calculating the difference in the proportion of looking to a respective AOI (i.e., the face, body, and the mouth) in a given contrast. Specifically, we took the proportion of looking to the AOI averaged across participants in one condition and subtracted this value from the other relevant condition. This way we obtained the following contrasts: D – C, DG – D, CG – C, and DG – CG, which were compared between the two listener groups. That is, native listeners (D – C) vs non-native listeners (D – C); native listeners (DG – D) vs non-native listeners (DG – D); native listeners (CG – C) vs non-native listeners (CG – C), and native listeners (DG – CG) vs non-native listeners (DG – CG).

In order to test for gestural enhancement, the procedure was similar. We computed the difference of (DG – D) and compared it to the difference of (CG – C) in native and non-native listener group. For the between-group comparisons, the effect of gestural enhancement was tested by contrasting the score of ((DG – D) – (CG – C)) in native listeners with that of non-native listeners.

Given that we aimed to determine whether the effect of gestural enhancement was significantly related to the amount of overt attention to gestures in each listener group, we obtained Spearman correlation coefficient by means of correlating the participant’s average proportion of looking to gestures in CG condition and DG condition in a set time interval with their interaction score of gestural enhancement in terms of the percentage of correct responses and RTs on a cued-recall task in. That is, Step 1 entailed calculating the mean proportion of looking to the AOI defined for body during the time interval from 550 msec to 1380 msec (starting from the meaningful part of gesture until its offset) separately for clear speech + gesture and degraded speech + gesture condition. Then, Step 2 included computing the interaction score for accuracy as well as RTs by subtracting the participant’s score of ((DG – D) from their score in (CG – C)). Thus ((DG – D) – (CG – C)) was obtained per participant for the accuracy and RTs separately. The value of Step 1 (in CG and DG) and Step 2 were correlated for each participant group.
3.5.4 Comparing mean proportion of looking between native and non-native listeners

In order to determine how much attention on average gestures received from each participant group when gestures unfolded under good and adverse listening conditions, we conducted independent $t$-tests. The mean proportion of looking to the AOI (body) in CG condition and DG condition (starting from the meaningful part of gesture until its offset, from 550 msec to 1380 msec) in native listeners was compared with that in the non-native listener group. That is, we first computed the proportion of looking to body per participant in a set time window (550 - 1380 msec) in native listeners and in non-native listeners and then conducted two separate $t$-tests (for CG and for DG condition).
4. Results

4.1 Behavioral results

4.1.1 Reaction times

A three-way mixed repeated measures ANOVA with two within-subject factors (speech, gesture) and one between-subject factor (language) revealed that there was a main effect of speech, $F(1, 38) = 213.83, p < .001, \eta^2 = 0.47$, suggesting participants were faster to respond when speech was clear than when it was degraded, a main effect of gesture, $F(1, 38) = 62.72, p < .001, \eta^2 = 0.05$, suggesting response latencies were shorter when gestures were present than when they were absent, and a main effect of language, $F(1, 38) = 4.80, p = 0.03, \eta^2 = 0.09$, indicating native listeners were faster to respond than non-native listeners of Dutch. There was also a two-way interaction between speech and gesture, $F(1, 38)= 24.73, p < .001, \eta^2 = 0.01$ as well as speech x gesture x language interaction, $F(1, 38) = 6.97, p = .01, \eta^2 = 0.004$, suggesting that a two-way interaction was significantly different in Dutch and German participants. We further examined this three-way interaction by submitting native listeners’ RTs in ((CG – C) – (DG – D)) and non-native listeners’ RTs in ((CG – C) – (DG – D)) to a statistical analysis in order to ascertain whether speech x gesture interaction (i.e., the effect of gestural enhancement) was significantly larger in one of the two groups. We found that the mean in ((CG – C) – (DG – D)) for native listeners ($M = 267.45, SD = 215.35$) was significantly larger than that for non-native listeners ($M = 82.00, SD = 228.85$), $t(38) = 2.63, p = 0.011$, suggesting the presence of gestures when speech was degraded sped up performance more for native listeners than non-native listeners (see Fig. 4 and Fig. 5).

![Figure 4](image-url)
Figure 5. The graph displays reaction times (in milliseconds) from a cued-recall task for the non-native listener group across the four conditions. The jittered points show the spread of the raw participant data while the split violin with the boxplot elements (representing the median, the interquartile range and the 95% CI) shows the distribution of the data.

4.1.2 Accuracy of responses

A second three-way mixed repeated measures ANOVA with the same factors revealed that there was a main effect of speech, $F(1, 38) = 193.90, p < .001, \eta^2 = 0.64$, suggesting more correct responses were given when speech was clear than when it was degraded, a main effect of gesture, $F(1, 38) = 63.91, p < .001, \eta^2 = 0.15$, suggesting the accuracy of responses was higher when gestures were present than when they were absent, and a main effect of language, $F(1, 38) = 6.42, p = 0.01, \eta^2 = 0.06$, indicating that native listeners of Dutch gave more correct answers than non-native listeners on a cued-recall task. There was also a two-way interaction between speech and gesture, $F(1, 38) = 70.81, p < .001, \eta^2 = 0.15$, suggesting that the effect of speech was dependent on the presence of gestures, with more correct answers given in degraded speech + gesture condition than in degraded speech condition. Speech x gesture x language interaction was non-significant, $F(1, 38) = 0.0002, p = 0.987$ (see Fig. 6 and Fig. 7).
Figure 6. The graph displays the accuracy of responses (in percentages) from a cued-recall task for the native listener group across the four conditions. The jittered points show the spread of the raw participant data while the split violin with the boxplot elements (representing the median, the interquartile range and the 95% CI) shows the distribution of the data.

Figure 7. The graph displays the accuracy of responses (in percentages) from a cued-recall task for the non-native listener group across the four conditions. The jittered points show the spread of the raw participant data while the split violin with the boxplot elements (representing the median, the interquartile range and the 95% CI) shows the distribution of the data.
4.1.3 Error rates

Please note that only significant results are reported. It was found that the frequency of phonological errors (but not semantic or other errors) and the experimental condition were not independent. In both listener groups, phonological errors occurred more frequently in degraded speech condition than would be expected, which was reflected in a significant result for native listeners, $\chi^2 (3, N = 80) = 16.32, p = 0.0009$ and non-native listeners, $\chi^2 (3, N = 80) = 9.55, p = 0.02$. 
4.2 Ocular results

4.2.1 Visual attention to the face in native and non-native listeners: within group comparisons

The overview of the proportion of looks devoted to the face across the four conditions for native and non-native listeners is given in Fig. 8. As is visually evident, both native and non-native listeners looked almost exclusively to the face in clear speech and degraded speech condition. In the two conditions in which gestures were present, non-native listeners devoted to the face approximately similar amount of attention regardless of the quality of the speech signal while native listeners gazed more to the face when speech was degraded compared to when it was clear.

![Figure 8. Overview of the proportion of looks directed to the face in native listeners (left) and non-native listeners (right). The shaded area represents SE estimating the interval into which the parameter (i.e., proportion of looking to the face) of the original population will fall.](image)

In order to statistically compare whether the amount of attention devoted to the face differed across conditions, we performed cluster based permutation tests for the relevant contrasts in each listener group. We found that native and non-native listeners looked significantly more to the face in degraded speech (D) than clear speech (C) condition in the second half on the video (see the grey-shaded area in Fig. 9). For native listeners, a positive cluster, meaning the proportion of attention to the face was higher in degraded speech than clear speech ($C < D, p < .0001$) was observed from 1019 msec to 2000 msec. For non-native listeners, two positive clusters were present. The first positive cluster, suggesting non-native listeners, too, looked more to the face in degraded than clear speech ($C < D, p < .0001$) was present from 259 msec to 385 msec. The second positive cluster ($C < D, p < .0001$) was present from 1243 msec to 2000 msec, indicating that towards the end of the video non-native listeners looked more to the face in degraded speech than clear speech condition. Both groups thus only looked significantly more to the face when speech was degraded than clear towards the end of the video and in a highly similar time interval. Note that the finding of non-native listeners looking significantly more to the face in a very short time window between 259-385 msec when speech was degraded cannot be accounted for by our experimental manipulation as speech onset was always at around 680 msec.
As an additional step, in order to determine whether more attention to the face in the second half of the video was driven by looking to the mouth area we thus zoomed into the face and contrasted the proportion of looks devoted to the mouth in degraded speech condition vs clear speech condition (see the blue-shaded area in Fig. 9). For native listeners, we observed a positive cluster (C < D, p < .001) from 1199 msec to 1990 msec, suggesting in this time interval, native listeners looked more to the mouth in degraded compared to clear speech. For non-native listeners, a positive cluster (C < D, p < .0001) was evidenced from 1351 msec to 2000 msec thus non-native listeners also looked more to the mouth when speech was degraded than when it was clear.

![Figure 9. The graphs display the proportion of looks in degraded speech (D) vs clear speech (C) condition for native listeners (left) and for non-native listeners (right); the grey-shaded area represents the time window when the proportion of looking to the face is significantly different between D and C while the blue-shaded area shows the significant time window of the same contrast for the proportion of looks devoted to the mouth area. Note that the grey-shaded and the blue-shaded areas overlap.](image)

Significant differences in the amount of attention devoted to the face for each listener group were also found for the contrasts of degraded speech + gesture vs clear speech + gesture (see the blue-shaded area in Fig. 10), clear speech + gesture vs clear speech (see the beige-shaded area in Fig. 10), as well as degraded speech + gesture and degraded speech (see the purple-shaded area in Fig. 11).

Having compared degraded speech + gesture (DG) vs clear speech + gesture (CG) condition, for native listeners, we observed a positive cluster (DG > CG, p < .002) from 1216 msec to 1926 msec, suggesting participants looked more to the face when gestures were present in degraded speech than when they were present in clear speech. For non-native listeners, a positive cluster (DG > CG, p < .008) was observed from 1500 msec to 1883 msec, indicating that non-native listeners as well looked more to the face when speech was degraded and gestures were present compared to gestures present in clear speech. For the contrast of clear speech + gesture (CG) and clear speech (C) in native listener group, we observed a negative cluster (CG < C, p < .004) in the time window 538-1301 msec. A negative cluster indicates that native listeners looked significantly more to the face in clear speech than clear speech + gesture condition. Similarly, for non-native listeners, there was a negative cluster (CG < C, p < .0001) present from 221 msec to
1378 msec, suggesting that non-native listeners allocated more attention to the face when speech was clear than when speech was clear and gestures were present.

![Figure 10. The graphs display the proportion of looks to the face within the significant time windows for two contrasts: clear speech vs clear speech + gesture (the beige-shaded area) and degraded speech + gesture vs clear speech + gesture (the blue-shaded area) in native (left) and non-native (right) listeners.]

When the proportion of looks devoted to the face was compared in degraded speech + gesture (DG) and degraded speech (D) conditions, it emerged that they were different for both native and non-native listeners in a highly similar time window (see the purple-shaded area in Fig. 11). Native listeners looked more to the face in degraded speech than degraded speech + gesture condition evidenced in a negative cluster (DG < D, p < .001) present from 249 msec to 1969 msec, which indicates that the proportion of looks directed to the face was higher in degraded speech than degraded speech + gesture condition. Similarly, for non-native listeners, a negative cluster (DG < D, p < .0001) was present from 297 msec to 2000 msec, suggesting that non-native listeners looked more to the face in degraded speech condition compared to degraded speech + gesture condition.
Finally, in order to test for the effect of gestural enhancement of degraded speech in terms of attention allocation to the face in each listener group, we subtracted the proportion of looks devoted to the face in degraded speech + gesture condition from that in clear speech + gesture condition and compared this difference (DG – CG) to the difference in the proportion of looking to the face in degraded speech condition and clear speech condition (D – C) (see the grey-shaded area in Fig. 12). For native listeners we found a positive cluster (DG – CG < D – C, \( p = 0.002 \)) from 1602 msec to 2000 msec. For non-native listeners there was a positive cluster (DG – CG < D – C, \( p = 0.0001 \)) present from 1464 msec to 2000 msec. The positive cluster observed in both groups suggests that the proportion of looks devoted to the face was larger when there were no gestures compared to when gestures were present in both clear and degraded speech.
4.2.2 Visual attention to the face in native and non-native listeners: between-group comparisons

Having examined differences in visual attention allocated to the face across the four conditions for each listener group separately, we similarly used cluster-based permutation tests in order to determine whether the two groups differed in terms of gaze behavior with respect to the face. We first computed the proportion of attention devoted to the face in a given contrast in each listener group in order to be able to conduct between-group comparisons. This was done by subtracting the proportion of looks devoted to the face in one condition from that in the other condition (i.e., D – C, DG – D, CG – C, and DG – CG). The relevant contrasts were then statistically compared between the two listener groups: native listeners (D – C) vs non-native listeners (D – C); native listeners (DG – D) vs non-native listeners (DG – D); native listeners (CG – C) vs non-native listeners (CG – C), and native listeners (DG – CG) vs non-native listeners (DG – CG). For the effect of gestural enhancement this included ((DG – D) – (CG – C)) in native listeners vs ((DG – D) – (CG – C)) in non-native listeners.

Note, we present the results for the between-group comparisons that reached the statistical significance, non-significant results are not reported.

We found a negative cluster present from 264 msec to 425 msec (native listeners (D – C) > non-native listeners (D – C), p = 0.03). This indicates that native listeners looked significantly more to the face in a very early time window in the first half of the video than non-native listeners in the conditions in which gestures were absent. This significant result in gaze behavior between the two groups cannot be explained by our experimental manipulation as speech would always start at around 680 msec. The lack of significant findings for other contrasts in terms of looking to the face suggests that gaze behavior in the two groups was similar, in that both groups looked at the face when gestures were absent and they also largely did so when gestures were presented under good and adverse listening conditions.

Finally, it was of interest to examine the amount of attention allocated to the face when the meaningful part of gesture unfolds (530-1380 msec). We computed the mean proportion of looks devoted to the face in clear speech + gesture condition during this time interval in native listeners (M = 0.82, SD = 0.19) as well as non-native listeners (M = 0.74, SD = 0.22), and also compared the two group means. The result was non-significant, t(38) = 1.12, p = 0.2, indicating that even though non-native listeners appear to have gazed less to the face as can also be ascertained visually in Fig. 8 (the proportion fluctuates from 0.7 to 0.8 for non-native listeners), native and non-native listeners are similar in the amount of attention that the face receives throughout this time interval.

We also ascertained whether the mean proportion of looks devoted to the face in degraded speech + gesture condition during the time interval when the meaningful part of gesture unfolds (530-1380 msec) differed between native listeners (M = 0.91, SD = 0.07) and non-native listeners (M = 0.77, SD = 0.19), a significant result was observed, t(38) = 2.95, p = .005, indicating that native listeners devoted larger amount of attention to the face than native listeners during this time interval.
4.2.3 Visual attention to gestures in native and non-native listeners: within-group comparisons

Similar to the statistical analyses performed for the proportion of looking to the face, we compared the proportion of looks devoted to the AOI defined for body across the four conditions in native and non-native listeners and for the relevant contrasts between the two groups. The overview of the proportion of looking to body across the four conditions for native and non-native listeners is displayed in Fig. 13. It can be seen that non-native listeners allocated more visual attention to gestures than native listeners under good and adverse listening conditions.

![Figure 13](image)

*Figure 13. Overview of the proportion of looks directed to body in native listeners (left) and non-native listeners (right). The shaded area represents the SE estimating the interval into which the parameter (i.e., proportion of looking to the face) of the original population will fall.*

Having compared the proportion of looks devoted to body in clear speech vs degraded speech condition for native listeners, we observed a negative cluster (D < C, p < .001) that was present from 769 msec to 2000 msec (see the purple-shaded area in Fig. 14). This indicates that native listeners looked more to body when speech was clear than when it was degraded. In the non-native listener group, for the same contrast there was a negative cluster (D < C, p < .001) present from 259 msec to 2000 msec, indicating that non-native listeners attended significantly more to body in clear speech than degraded speech.
Figure 14. The graphs display the proportion of looks to body within the significant time windows for two contrasts: degraded speech + gesture vs degraded speech (the red-shaded area) and clear speech vs degraded speech (the purple-shaded area) in native listeners (left) and non-native (right). Note that the areas overlap for non-native listeners.

For the comparison of degraded speech + gesture vs degraded speech in native listeners we observed a positive cluster (DG > D, p < .001) present in the time window 245-2000 msec (see the red-shaded area in Fig. 14), suggesting the amount of attention allocated to body was significantly higher when speech was degraded and gestures were present than in degraded speech condition. Similarly, in non-native listeners, there was a positive cluster (DG > D, p < .001) present from 264 msec to 2000 msec, indicating that non-native listeners looked more to body when speech was degraded and when gestures were present rather than absent.

We also found that both native and non-native listeners looked more to body in clear speech + gesture than clear speech condition, which for native listeners was evidenced in a positive cluster (CG > C, p < .0001) that was present from 233 msec to 1425 msec while for the non-native listener group this was evidenced in a positive cluster (CG > C, p < .0001) that was observed from 217 msec to 1448 msec (see the beige-shaded area in Fig. 15).
When we contrasted degraded speech + gesture and clear speech + gesture conditions in native listener group, a negative cluster (CG > DG, $p < .013$) was observed from 1353 msec to 1884 msec (see the blue-shaded area in Fig. 15), meaning that native listeners looked more to body when speech was clear and gestures were present than when gestures were present in degraded speech. For the same contrast in non-native listeners, a negative cluster (CG > DG, $p < .002$) was evidenced from 1521-1759 msec. This suggests that both participant groups looked more to body in the second half of the video when gestures were presented in clear speech than when they were present in degraded speech.

In order to test for the effect of gestural enhancement of degraded speech in terms of the proportion of looking to the AOI defined for body in native and non-native listeners, we contrasted the proportion of looks to body in (DG – CG) with that in (D – C). For native listeners, we found a positive cluster (DG – CG > D – C, $p < .001$) that was present from 1648 msec to 2000 msec. For non-native listeners there was a positive cluster (DG – CG > D – C, $p < .0001$) present from 1641 msec to 2000 msec (see the grey-shaded area in Fig. 16). This result in terms of both groups suggests that native and non-native listeners looked more to body when gestures were presented in clear as well as degraded speech than when gestures were absent under both good and adverse listening conditions.
4.2.4 Visual attention to gestures in native and non-native listeners: between-group comparisons

In order to test for differences in the amount of attention with respect to the AOI defined for body between the two groups, we repeated the same procedure described in section on page 26 (Visual attention to the face in native and non-native listeners: between-group comparisons).

Note, we present the results for the between-group comparisons that reached the statistical significance, non-significant results are not reported. In terms of differential gaze behavior in native vs non-native listeners for the relevant contrasts, we observed two significant results.

It was found that non-native listeners allocated more overt attention to gestures than native listeners under good listening conditions, evidenced in a positive cluster (non-native listeners (CG–C) > native listeners (CG–C), \(p = 0.008\)) that was present from 731 msec to 952 msec. A positive cluster indicates that the proportion of looks devoted to body was higher in non-native listeners than the one in the native listener group in the respective time window. It was also revealed that non-native listeners allocated more overt attention to gestures than native listeners under adverse listening conditions, evidenced in a positive cluster (non-native listeners (DG–D) > native listeners (DG–D), \(p = 0.029\)) that was present from 743 msec to 1087 msec. This suggests that the proportion of looks devoted to body was larger in non-native listeners compared to native listeners.

Finally, we examined whether the mean proportion of looking to body in clear speech + gesture condition during the time interval when the meaningful part of gesture unfolds (530-1380 msec) differed between native listeners (\(M = 0.17, SD = 0.19\)) and non-native listeners (\(M = 0.25, SD = 0.22\)), a non-significant result was observed, \(t(38) = -1.12, p = 0.26\), indicating that on average throughout this whole time interval mean proportion of looking to body in native listeners and non-native listeners was similar.

Having compared the mean proportion of looks devoted to body in degraded speech + gesture condition during the time interval when the meaningful part of gesture unfolds (530-1380 msec) in native listeners (\(M = 0.11, SD = 0.10\)) vs non-native listeners (\(M = 0.22, SD = 0.19\)), a
significant result was observed, $t(38) = -2.03, p = 0.04$, indicating that under adverse listening conditions, non-native listeners devoted larger amount of attention to gestures in the time interval when speech and gestures were unfolding compared to native listeners.

4.2.5 Relationship between the amount of overt attention to gestures and performance on a cued-recall task in native and non-native listeners

In native listener group, having correlated the effect of gestural enhancement in terms of the percentage of correct responses on a cued-recall task and the amount of overt attention to gestures in degraded speech + gesture condition in the time interval from 550 msec to 1330 msec (when the meaningful part of gesture unfolds), we observed a significant relationship between the two, $r_s = .50, p = .02$, suggesting that the more a native listener looked to gestures, the more correct answers s/he was able to obtain (see Fig. 17). The correlation for reaction times was non-significant, $r_s = .22, p = .33$, indicating that looking more to gestures did not speed up the performance on a cued-recall task.

Figure 17. Interaction score of gestural enhancement (in terms of accuracy) plotted against the amount of overt attention allocated to gestures in degraded speech + gesture condition in native listeners

In non-native listener group, having correlated the effect of gestural enhancement in terms of percentage of correct responses on a cued-recall task with the amount of overt attention to gestures in degraded speech + gesture condition during the time interval from 550 msec to 1330 msec (when the meaningful part of gesture unfolds), the relationship was non-significant, $r_s = .08, p = .71$, nor was it for RTs, $r_s = -0.41, p = .07$. 
We correlated the gestural enhancement effect (in terms of both the percentage of correct responses and RTs) on a cued-recall task with the amount of attention devoted to gestures in clear speech + gesture condition for the native as well as the non-native listener group, none of the correlations were significant.

Native listeners:

Attention to gestures in CG condition and the gestural enhancement effect in terms of percentage of correct responses, $r_s = 0.20, p = .37$, and in terms of RTs, $r_s = 0.04, p = .83$.

Non-native listeners:

Attention to gestures in CG condition and the gestural enhancement effect in terms of percentage of correct responses, $r_s = 0.21, p = .37$, and in terms of RTs, $r_s = -0.37, p = .10$. 
5. Discussion

The present study used eye tracking to investigate how native and non-native listeners allocate visual attentional resources when integrating auditory speech with visual information under clear and adverse listening conditions. We manipulated the external (speech intelligibility) and internal (listener status) factors by presenting participants, native and non-native listeners of Dutch, with videos of an actress who uttered an action verb in Dutch (clear or 6-band noise vocoded speech), accompanied by an iconic gesture or not. Along with the ocular measures (proportion of looking to different areas of interest), we collected and analyzed the behavioral data from a cued-recall task (how accurately and quickly the participants identified the verb presented in the videos).

In line with previous studies (Drijvers & Özyürek, 2017; Drijvers, Özyürek & Jensen, 2018; Holle et al., 2010), the behavioral data revealed the effect of speech degradation and gesture presence as well as their interaction on participants’ performance. In both listener groups, response latencies were longer and accuracy dropped in degraded speech condition and the condition in which gestures were absent, with non-native listeners being more hindered by this overall. The fact that speech degradation was detrimental to both listener groups was also manifested in having observed more phonological errors than would be expected in degraded speech condition. The participants’ performance was improved by gesture presence when speech was degraded reflected in shorter RTs and better accuracy, suggestive of the effect of gestural enhancement of degraded speech. The gestural enhancement effect was, however, larger for native listeners compared to non-native listeners, the former group benefitted more from the presence of gestures when speech was degraded (this was evidenced in terms of RTs yet not accuracy as we did not observe a significant three-way interaction for the percentage of correct responses on a cued-recall task). Consequently, we observed a significant relationship between more overt attention to gestures when speech was degraded and higher accuracy for native listeners only. Those native listeners who devoted more explicit visual attention to gestures in the time interval when gestures were unfolding and speech was degraded (from the stroke and onwards, to gesture offset), were able to obtain a higher percentage of correct responses on a cued-recall task across all conditions. Taken together, the results seem to suggest that native listeners are able to derive more benefit from visual semantic information conveyed by gestures when comprehending degraded speech than non-native listeners.

With respect to the results from ocular measures, several conclusions can be drawn. First, in the absence of gestures, native and non-native listeners focused almost exclusively on the face when speech was clear and when it was degraded. When gestures were presented, relative to the face, they attracted a small amount of attention from both participant groups, with overall more overt attention being allocated to gestures when speech was clear than when it was degraded in both groups. Importantly, non-native listeners allocated more visual attention to gestures compared to native listeners irrespective of the quality of the speech input and in the time interval when speech and gestures unfolded.
5.1 Prolonged attention to the face in degraded speech in both listener groups

In the absence of gestures, the face is a strong locus of attention for both listener groups in clear speech and degraded speech. The strong preference for faces can be explained by faces being a frequently encountered class of stimuli in daily life that convey various information, including social, emotional and identity (Baron-Cohen et al., 2001; Letourneau & Mitchell, 2011) as well as linguistically relevant cues for audiovisual speech processing (Rosenblum, 2008) and turn-taking (Ho et al., 2015).

In our data, the fact that listeners focus on the face for the visual correlates of speech when speech is degraded has been evidenced by significant differences in the proportion of looking to the face in clear speech vs degraded speech condition that emerge in the second half of the video. We observed a more rapid decline in the amount of attention allocated to the face in clear than degraded speech. It is reasonable to assume that this pattern of sustained attention to the face when speech is degraded in both participant groups reflects listeners seeking to extract visible speech cues that are functional in speech perception under suboptimal listening conditions (Benoît, Mohamadi, & Kandel, 1994; Sumby & Pollack, 1954; Ross et al., 2007). Our additional analysis specifically showed that participants gazed more to the mouth area in the second half of the video in degraded speech than clear speech, which is in line with previous eye tracking studies that show that with the decreasing quality of the speech input, attention to the mouth area increases (Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998, Yi et al., 2013). Mouth cues are a valuable source of information for word identification while the visual correlates of prosody and phonetics of speech are distributed all over the face and even beyond, in that they can be perceived from head motion (Munhall et al., 2004; Yehia et al., 2002).

It is of note though that the effect of looking more to the mouth in degraded speech than clear speech for both groups emerges relatively late (from 1199 msec and 1351 msec in native and non-native listeners, respectively). Based on the fact that visual speech information can be extracted from other areas than the jaw and lips and that even participants who gazed at the eyes for around 45% of the time at the highest level of masking noise were able to correctly identify a large proportion of the sentences (Vatikiotis-Bateson et al., 1998), we may propose that our participants, too, need not to immediately orient to the mouth when speech is degraded. Research shows that the eyes and the mouth are the two regions in the face that attract most attention when speech is presented in noise (Buchan et al., 2007), however, this was not found in the study by Vatikiotis-Bateson et al. (1998).

We should, however, consider the alternative explanation why attention to the face was sustained towards the end of the video when speech was degraded than clear. It may well be that as soon as the listeners became confident having understood the verb conveyed in the video, they started to direct their gaze to the middle of the screen in order to prepare for giving a response to a cued-recall task. This happened faster in clear than degraded speech as in clear speech participants were not hindered by the quality of the speech signal in understanding the verb.
5.2 More explicit attention to gestures in non-native listeners compared to native listeners under good and adverse listening conditions

Before the relevant findings with respect to gaze behavior when gestures are presented in clear and degraded speech are discussed in both listener groups, it is apt to bear in mind that attention to gestures is essentially attention away from the face. This notion, namely, gestures drawing attention away from the face is further conceptualized as the “gesture attraction” effect.

The “gesture attraction” effect is relatively small in clear and degraded speech in both native and non-native listeners. Participants largely gazed at the face when the meaningful part of gesture unfolded in clear speech and they did so even more under adverse listening conditions. Despite only a small amount of attention devoted to gestures relative to the amount of attention that the face received in both groups, non-native listeners was the group that gazed more to gestures in the time interval when both speech and gestures unfolded under good and adverse listening conditions.

Based on work on attention allocation during multimodal language comprehension under good listening conditions (Gullberg & Holmqvist, 1999, 2006; Gullberg & Kita, 2009), we thus expected a substantial amount of attention being allocated to the face relative to gestures in native listeners but also the non-native listener group. As indicated, non-native listeners gazed more to gestures compared to native listeners but even our native listener group paid more attention to gestures than has been reported in previous studies. The reason why this study reports a much larger amount of attention devoted to gestures by native listeners compared to previous studies, in which only native listeners/observers were tested (Gullberg & Holmqvist, 1999, 2006; Gullberg & Kita, 2009), may be attributed to differences in the task. While in previous work addressees were instructed to attend to the story conveyed by the speaker in order to later be able to answer questions (e.g., in Gullberg & Kita, 2009, subjects answered questions by drawing), our participants were asked to watch videos and identify the verb said in each video. To some extent, the task in the present study can be said to be more goal-driven compared to the ones in previous studies. The strategy of our participants may have been to explicitly draw on all available sources of information as much as possible in order to complete the task accurately and quickly. Given that gaze is very sensitive to task demands (Yarbus, 1967), this may explain why the differences in the amount of attention devoted to gestures have arisen between our and previous work. Regardless of this overall somewhat greater amount of visual attentional resources devoted to gestures in our study, our findings are in support of the previous observations that the face dominates attention during multimodal language comprehension.

The likely candidate why the face is the default location during multimodal language comprehension is that information conveyed by gestures can be extracted by peripheral vision as has been maintained by Gullberg & Kita (2009). An array of studies that used eye tracking to investigate visual attention during communication in sign language seem to also align with this argument. It has been, for instance, revealed that signers fixate the face more than other visual cues, such as the hands (Agrafiotis et al. 2003; Muir & Richardson, 2005; Emmorey et al., 2009). Maintaining the face in the centre of the visual field enables signers to extract detailed movements of the facial articulators (eyebrows, lips) while the information from the manual signs is extracted by means of the peripheral vision (Swisher et al., 1989). Due to movements of hands and arms evolving in a relatively large space (compared to the face) in front of the speaker’s body, they can be attended in the periphery, which is in particular suited for processing motion information.
(Lappin et al., 2009). Along similar lines, Beattie (2016) has argued that as long as motion is ongoing, peripheral vision is suitable for extracting information from gestures. Evidence from work by Gullberg & Kita (2009) indeed shows that information from gestures that are not overtly attended can be remembered and used for the task completion (specifically, participants were able to correctly draw the direction of movement, the information contained only in gestures but not speech). Under certain circumstances, as already mentioned, gestures do receive explicit attention from listeners/viewers (Gullberg & Kita, 2009). Our study shows that in addition to the two factors, autofixations and gestural holds, there is a third one that drives overt attention to gestures more than one would expect, thus, being a non-native listener.

5.3 Language proficiency as the modulator of overt attention to gestures during multimodal language comprehension

Based on the electrophysiological results from Drijvers & Özyürek (2018), it was in line with our expectations the finding that non-native listeners gaze significantly more at gestures than native listeners when speech is clear. In the study by Drijvers & Özyürek (2018), the heightened amount of overt attention to gestures has been reflected in the enhanced N400 effect in clear speech for non-native listeners, for whom also the topographical distribution of this effect was more right-lateralized compared to that of native listeners. The recruitment of right-lateralized areas playing a prominent role during speech-gesture integration (Holle et al., 2010; Willems et al., 2007) may indicate that non-native listeners even though they are highly proficient in the language seek to enhance auditory language comprehension with visual semantic cues even when speech is clear. Our study is in support of the notion that proficiency level may modulate the extent to which a listener/viewer explicitly draws on gestural information in good listening conditions by showing more overt attention devoted to gestures in non-native compared to native listeners. Previous behavioral work (Sueyoshi & Hardison, 2005) also indicates that the degree of reliance on gestures may hinge on language proficiency. Specifically, the study by Sueyoshi & Hardison (2005) showed that the English learners of low proficiency benefitted more from gestures for content comprehension than learners who had high proficiency. It is reasonable then to propose that as the proficiency level of the language increases, the degree of explicitly relying on gestures for comprehension when speech is clear may decrease, but even at the very high level of language proficiency non-native listeners seem to be more alert to visual semantic cues conveyed by gestures relative to native listeners.

Under adverse listening conditions, our data point to a reduced amount of visual attention allocated to gestures relative to when gestures are presented in clear speech in both listener groups. Attending less to gestures in degraded speech + gesture condition than clear speech + gesture condition happens at the expense of focusing more on the face. The diminished quality of the speech signal may cause the observers to zoom into the face in order to extract information from visible speech while gestural information can be attended in the periphery as is the case when speech is clear. Yet, even under suboptimal listening conditions, non-native listeners allocate more overt attention to gestures than native listeners, and specifically, when speech and gesture unfold. The fact that in degraded speech non-native listeners pay more attention to gestures than the native listener group suggests that the former group may well prioritize visual semantic information to a greater extent than native listeners. It is also reasonable to assume that non-native listeners attempt to integrate visual semantic information more strongly than native listeners and thus they devote a greater amount of visual resources to gestures. Despite that perception can occur without direct
fixations, processing is more efficient when one allocates overt attention to a spatial location (Eckstein et al., 2017).

Our data revealed that in degraded speech + gesture condition, non-native listeners devote more attention to visual semantic information than native listeners in the time window from ~ 700 msec to 1000 msec. It is relevant to note that this time window is reminiscent of the time interval from a previous MEG study that used the same video materials to examine the oscillatory dynamics of gestural enhancement (however, in native listeners of Dutch) (Drijvers, Özyürek & Jensen, 2018). In the MEG study, during the time interval from 700 msec to 1100 msec one of the brain areas that was significantly engaged was the right superior temporal sulcus (STS). The STS is the brain structure associated with hosting audiovisual speech integration (Callan et al., 2003) and, more broadly, the integration of multimodal stimuli (Beauchamp, 2005). Thus, the ocular and the brain data seem to point to the idea that non-native listeners are actively integrating visual semantic information with the auditory stream when speech is degraded. Yet, as the auditory cues are degraded, processing may be hampered more for non-native listeners than native listeners, in that visual semantic information needs to be coupled to the auditory input. This may be the reason why the N400 effect for non-native listeners was absent when speech was degraded (Drijvers & Özyürek, 2018) even though our study shows that non-native listeners allocate more visual attention to gestures than native listeners during the time window when gestures unfold. In consequence, the increased difficulty in binding the visual and the auditory streams may also be the reason why we observed a significant relationship between more overt attention to gestures in degraded speech and higher percentage of correct responses on a cued-recall task for native listeners only.

Non-native listeners compared to native listeners may be less able to benefit from gestural information under adverse listening conditions as visual semantic cues need to be coupled to the acoustic cues in a similar manner as visible speech is mapped onto the auditory speech to derive audiovisual benefit. The benefit of visual speech information coupled to the auditory cues are highly dependent on the ability to perform the acoustic-phonetic mapping. In order to bind auditory information to its visual correlate good familiarity with the sound system of the language is required, and this is where non-native listeners are disadvantaged compared to native listeners, and especially so under adverse listening conditions (Bradlow & Bent, 2007; Mayo et al., 1997; Takata & Nabelek, 1990). For a native listener who has a life-long experience of how each phonetic contrast may map onto a visual correlate, visible speech confers larger benefit in comparison to a non-native listener. Our study suggests that the mechanisms that seem to modulate the degree of audiovisual benefit in terms of non-native listening may also be at play when a non-native listener attempts to integrate visual semantic information conveyed by gestures with the auditory speech that is reduced in quality. The present study thus underlines the role that language proficiency plays in modulating the degree of audiovisual benefit, both in terms of visible speech as shown by previous work (Bradlow & Bent, 2007; Mayo et al., 1997; Takata & Nabelek, 1990) and visual semantic cues as revealed by our study.

This study by following the line of research demonstrating that language comprehension is multimodal under both good and adverse listening conditions, has revealed the differential role that visual semantic information conveyed by iconic gestures have in terms of native and non-native language processing. While non-native listeners may attach more significance to gestural information as shown by allocating more visual attention to gestures than native listeners,
essentially the amount of benefit that one can derive from visual semantic information by integrating it with the auditory speech cues under adverse listening conditions is highly dependent on the language proficiency of the listener.

6. Conclusion

The impetus for this study has been the potential differences in visual attention allocation between native and highly-proficient non-native listeners as suggested by previous work on speech-gesture integration under good and adverse listening conditions. We have been able to show that while both listener groups largely gaze at the face in clear and degraded speech, non-native listeners indeed allocate more visual attention to iconic gestures compared to native listeners. This finding is a novel contribution to the literature on visual attention during multimodal language comprehension as previously it has only been reported that explicit attention to gestures is mainly driven by two factors (autofixations and gestural holds). We here show that the non-native language proficiency is yet another aspect that influences more overt attention being allocated to gestures than would be normally expected and therefore irrespective of the speech quality. More specifically, language proficiency can be conceptualized as the modulating factor in terms of the degree to which listeners explicitly resort to gestures in order to enhance language comprehension. Non-native listeners in this study even though they are highly-proficient in the language may conceive of gestural information as a rich source of information and thus purposively seek to enhance language comprehension with visual semantic information conveyed by iconic gestures.
References


Drijvers, L., & Ozyürek, A. (resubmitted). Highly proficient non-native listeners benefit less than natives from visual context during degraded speech comprehension.


