Brain and Language 177–178 (2018) 7–17

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l

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

Linda Drijvers[a,b,*], Asli Özyürek[a,b,c]

[a] Radboud University, Centre for Language Studies, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands
[b] Radboud University, Donders Institute for Brain, Cognition, and Behaviour, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands
[c] Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

## 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, but in degraded speech only for natives. 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.

## 1. 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; Drijvers, Ozyurek, & Jensen, in press). 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, Dolk, & Gunter, 2012b; Drijvers et al., in press). 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, Florentine, & Buus, 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 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, and 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.

* Corresponding author at: Radboud University, Centre for Language Studies, Donders Institute for Brain, Cognition and Behaviour, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands.
*E-mail address:* linda.drijvers@mpi.nl (L. Drijvers).

https://doi.org/10.1016/j.bandl.2018.01.003
Received 28 January 2017; Received in revised form 5 January 2018; Accepted 15 January 2018

### 1.1. 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, 1999a, 1999b; Beattie & Shovelton, 2002; Holle & Gunter, 2007; Holler, Kelly, Hagoort, & Ozyurek, 2010; Holler, Shovelton, & Beattie, 2009; Holler et al., 2014; Kelly, Barr, Church, & Lynch, 1999; Kelly, Healey, Özyürek, & Holler, 2015; Obermeier, Holle, & Gunter, 2011; 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; Holle et al., 2012; 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, Ruschemeyer, Hennenlotter, & Iacoboni, 2008; Holle, Obleser, Rueschemeyer, & Gunter, 2010b; Willems, Özyürek, & Hagoort, 2007; Willems, Özyürek, & Hagoort, 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, Kravitz, & Hopkins, 2004; Kelly et al., 1999; Obermeier et al., 2011; Wu & Coulson, 2005, 2007a, 2007b) have focused on the N400 component to assess differences in semantic processing. The N400 is a negative-going ERP component between 300 and 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 & Lee, 2012; Kelly, Ward, Creigh, & Bartolotti, 2007; Kelly et al., 2004; Ozyürek, Willems, Kita, & Hagoort, 2007; Sheehan, Namy, & Mills, 2007; Wu & Coulson, 2007a, 2007b), 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 & Ozyürek, 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 & Ozyürek, 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, Dick, & Mills, 2006; Boulenger, Hoen, Jacquier, & Meunier, 2011; Obleser & Kotz, 2011; Strauß, Kotz, & Obleser, 2013). For example, Obleser and 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, Kotz, and Obleser (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, Nusbaum, & Small, 2006; Skipper, Wassenhove, Nusbaum, & Steven, 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.

### 1.2. 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, Rosen, & Scott, 2009; Mayo et al., 1997; Oliver, Gullberg, Hellwig, Mitterer, & Indefrey, 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 and 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.

### 1.3. 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 et al., 2011; Holle & Gunter, 2007; Ozyürek et al., 2007; Wu & Coulson, 2007a, 2007b), 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 & Ozyürek, 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 cued-recall after each item that asked which verb they had heard in the videos.

Behaviorally, and in line with previous work (Drijvers & Ozyürek, 2017; Holle et al., 2010), we expected that native listeners would benefit from gestures during degraded speech comprehension, resulting in more correct answers on the cued-recall 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 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, and might result 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).

## 2. Methods

### 2.1. 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.

### 2.2. 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.

### 2.3. Stimulus materials

The materials in this experiment are partially based on a subset of pretested stimuli which are described in more detail in Drijvers and Ozyürek (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 s (see Fig. 1B). 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 Fig. 1A).

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 680 ms, 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 Fig. 1C).

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 and Ozyürek (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, Morrel-Samuels, & Colasante, 1991), and that they were to an extent dependent on speech to be disambiguated (Habets et al., 2011, see Drijvers & Ozyürek, 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, Zeng, Kamath, Wygonski, & Ekelid, 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 & Ozyürek, 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 & Ozyürek, 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 Fig. 1 for an overview). All conditions consisted of 40 unique videos with unique verbs and gestures.

### 2.4. 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 1650 × 1080 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 cued-verb recall 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 min. After the experiment, participants filled in the adapted version of the LexTALE to test their knowledge of the verbs used in these videos.

### 2.5. EEG data acquisition

The participants EEG was continuously recorded throughout the

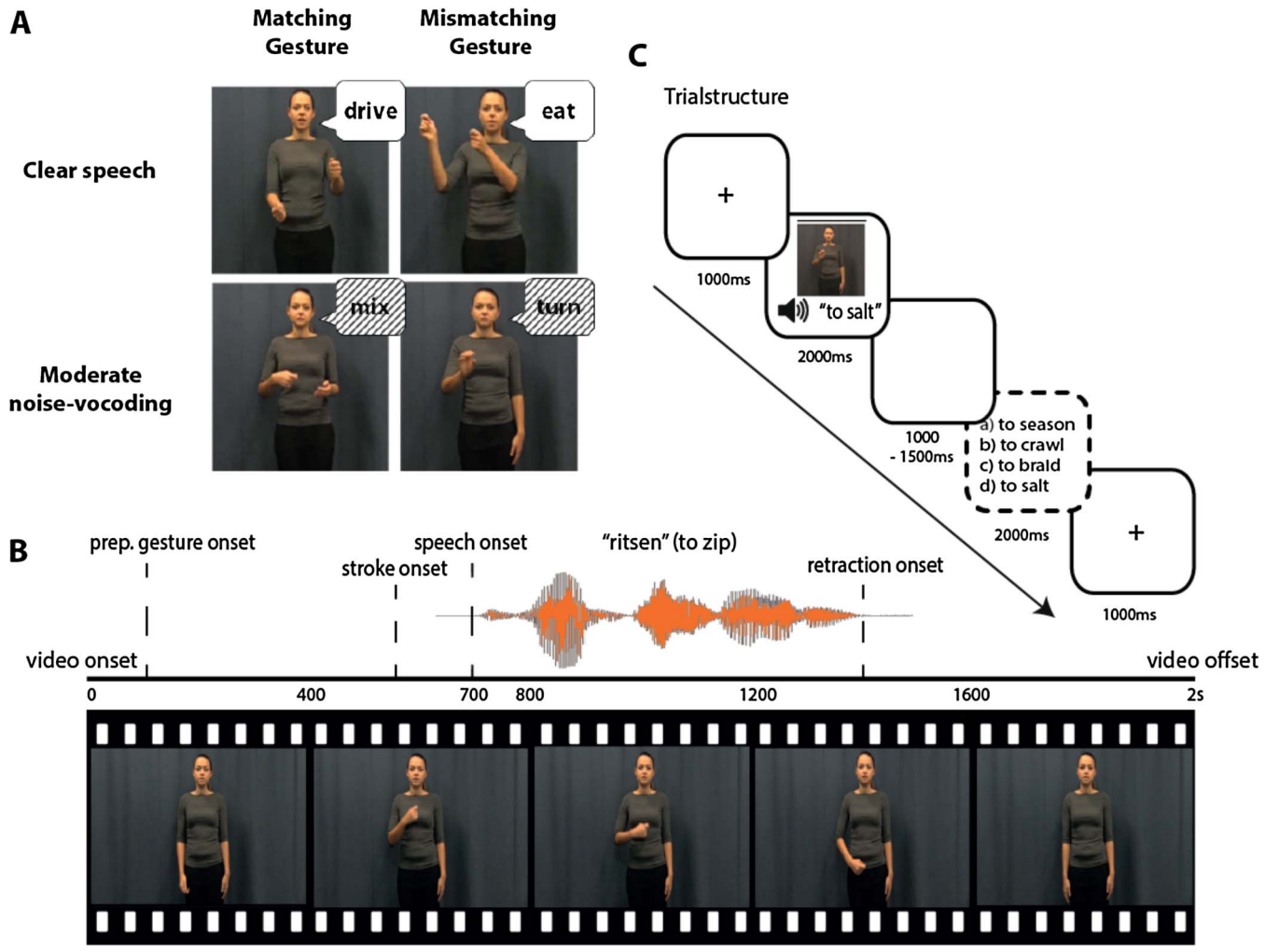


**Fig. 1.** Experimental overview. (A) Overview of conditions. (B) Video structure. (C) Trialstructure.

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 KOhm. The EEG was filtered through a 0.02–100 Hz band-pass filter and digitized on-line with a sampling frequency of 500 Hz.

### 2.6. EEG data analysis

We analyzed the EEG data by using Fieldtrip (Oostenveld, Fries, Maris, & Schoffelen, 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 -0.4 s to -0.2 s. 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 & Oostenveld, 2007).

## 3. Results

### 3.1. 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, $t(44) = 10.892$, $p = < .001$) 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, $t(44) = 7.34$, $p < .001$). 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.

### 3.2. Behavioral results - cued verb-recall task

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).

### 3.3. 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 ($F(1, 22) = 140.95$, $p < .001$, Wilks' Lambda = 0.135, $\eta 2 = 0.87$). We also found an effect of Gesture ($F(1, 22) = 128.87$, $p = < .001$, 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 ($F(1, 22) = 112.20$, $p = < .001$, 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 $t(22) = 11.12$, $p_{bon} < .001$, between clear-match and degraded-match $t(22) = 7.89$, $p_{bon} < .001$ and between clear-mismatch and degraded-mismatch $t(22) = 12.42$, $p_{bon} < .001$ (see Fig. 2).

We found a similar pattern in terms of reaction times and found a main effect of Noise-vocoding ($F(1, 22) = 74.11$, $p = < .001$, 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 ($F(1, 22) = 69.20$, $p = < .001$, 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 ($F(1, 22) = 43.87$, $p = < .001$, 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, $t(22) = -7.80$, $p_{bon} < .001$, between clear-match and degraded-match $t(22) = -6.97$, $p_{bon} < .001$, and between clear-mismatch and degraded-mismatch, $t(22) = -8.73$, $p_{bon} < .001$.

### 3.4. 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 ($F(1, 22) = 165.47$, $p < .001$, 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 ($F(1, 22) = 69.65$, $p < .001$, 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 ($F(1, 22) = 82.91$, $p < .001$, 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, $t(22) = -0.92$, $p = .367$, but did show significant differences between degraded-match and degraded-mismatch, $t(22) = -9.55$, $p < .001$, between clear-match and degraded-match $t(22) = -6.74$, $p < .001$, and between clear-mismatch and degraded-mismatch, $t(22) = -15.29$, $p < .001$.

We observed a similar pattern in reaction times as in response accuracy: We observed a significant main effect of Noise-vocoding ($F(1, 22) = 104.554$, $p < .001$, 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

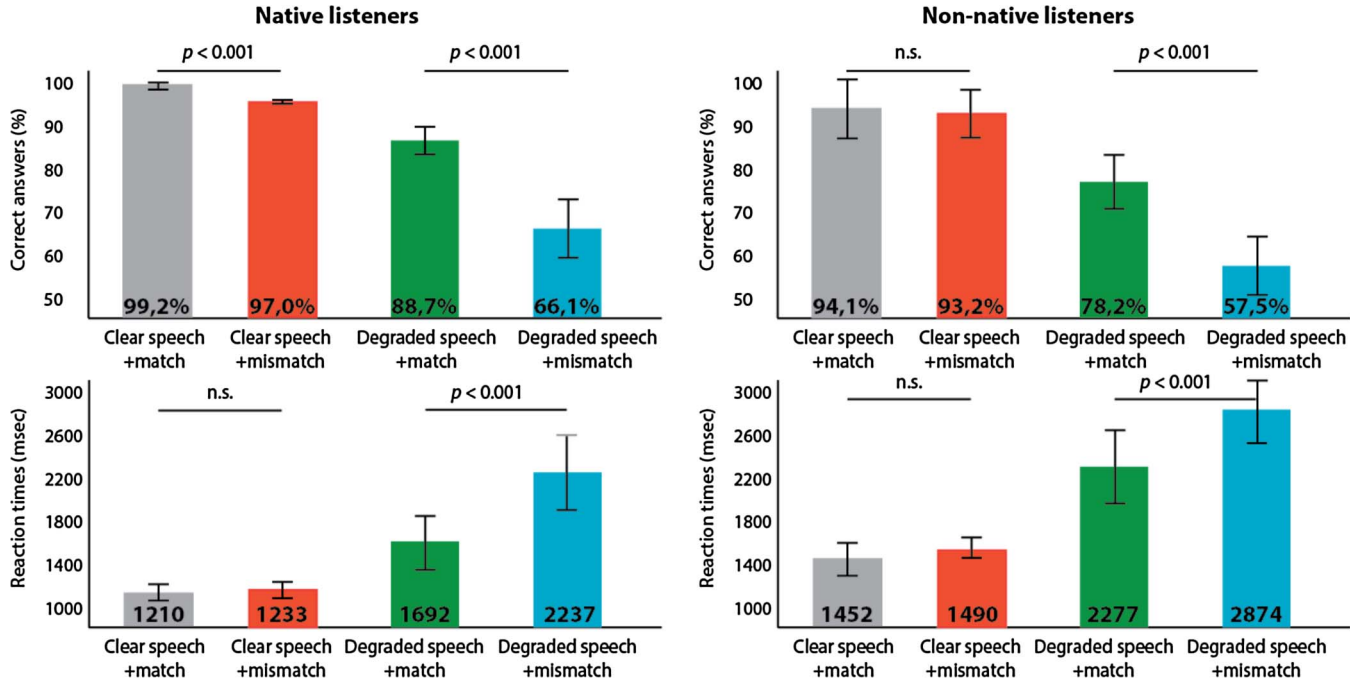


**Fig. 2.** Behavioral results of the cued-verb recall task. Top panels represent correct answers in percentages per listener group. Error bars present standard deviations. Lower panels represent reaction times (msec).

speech signal ($F(1, 22) = 53.42$, $p < .001$, 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 ($F(1, 22) = 59.53$, $p < .001$, 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, $t(22) = -8.576$, $p_{bon} < .001$, between clear-match and degraded-match $t(22) = -8.48$, $p_{bon} < .001$, and between clear-mismatch and degraded-mismatch, $t(22) = -10.76$, $p_{bon} < .001$.

Note that non-native listeners show a similar behavioral pattern as native listeners, but that they showed an overall lower accuracy and slower reaction times in the degraded speech conditions (see Fig. 2).

### 3.5. 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 300 ms 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, $p < .001$), the degraded-match and the degraded-mismatch condition (degraded-mismatch > degraded-match, $p < .05$), between the clear-match condition and degraded-match condition (clear-match < degraded-match, $p < .001$) and the clear-mismatch and degraded-mismatch condition (clear-mismatch < degraded-mismatch, $p < .001$). Fig. 3 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 ($p = .041$).

### 3.6. EEG data - non-native listeners

In non-native listeners, we observed a significant difference between clear-match and clear-mismatch (clear-mismatch > clear-match, $p < .001$), but not between degraded-match and degraded-mismatch ($p = .16$). We observed a significant difference between clear-match and degraded-match (clear-match < degraded-match, $p < .001$) and between clear-mismatch and degraded-mismatch (clear-mismatch < degraded-mismatch, $p < .05$) 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.

### 3.7. EEG data - native versus non-native listeners

Although the difference between native and non-native listeners in

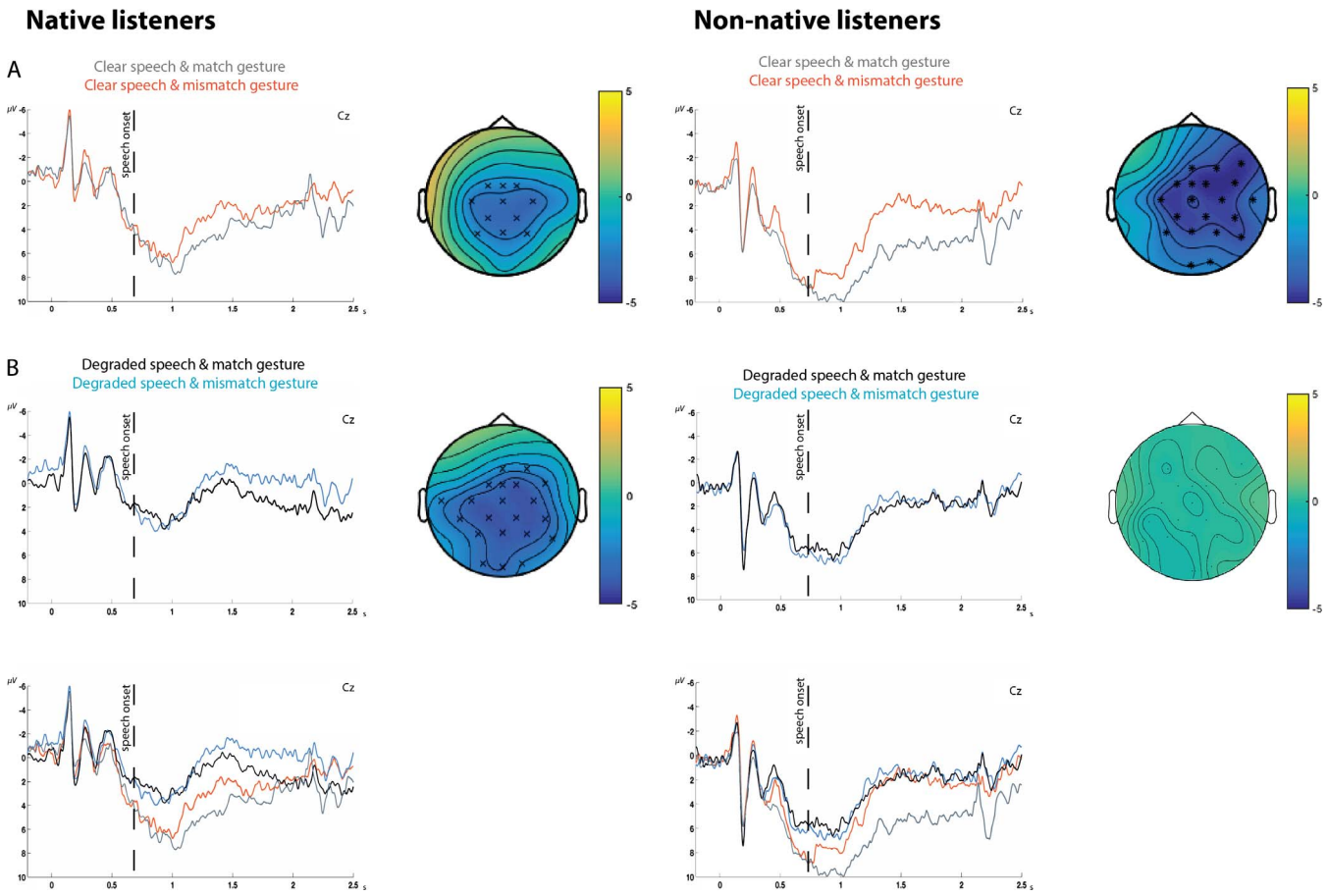


**Fig. 3.** Grand-average waveforms for ERPs elicited in the different conditions at electrode Cz. Negativity is plotted upward. Waveforms are timelocked to the onset of the video.

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.

## 4. 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.

### 4.1. 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 demonstrated lower overall accuracy scores and slower reaction times in the degraded conditions than in 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 et al. (2017)). Our EEG results provided more evidence for this claim.

### 4.2. EEG results - native listeners

We observed a more negative N400 amplitude when gestures mismatched compared to matched clear speech (in line with e.g., Habets et al., 2011; Holle & Gunter, 2007; Kelly et al., 2004; Ozyürek et al., 2007), 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 and 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 ambigious without the speech context.

In Fig. 3A 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 (1000 ms for clear speech vs. 1280 ms for degraded speech) Previous studies (e.g., Obleser & Kotz, 2011; Strauβ 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).

### 4.3. 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 (Bradlow & Alexander, 2007; Bradlow & Bent, 2002; Gat & Keith, 1978; Golestani et al., 2009; Mayo et al., 1997; 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, or could be explained by a ceiling effect. In the cued-recall 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 et al., 2014; Skipper, Goldin-Meadow, Nusbaum, & Small, 2009; Straube, Green, Weis, & Kircher, 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 et al., 2006, 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.

## 5. 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.

## Conflict of interest

The authors declare no conflict of interest.

## 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.

## References

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. http://dx.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. http://dx.doi.org/10.1080/01690965.2010.542671.

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

Beattie, G., & Shovelton, H. (1999b). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology, 18*(4), 438–462. http://dx.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. http://dx.doi.org/10.1348/000712602162526.

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language, 124*(2), 143–152. http://dx.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. http://dx.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. http://dx.doi.org/10.1016/j.cortex.2014.11.018.

Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer.

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. http://dx.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. http://dx.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. http://dx.doi.org/10.1121/1.1487837.

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., López, V., & Núñez, R. E. (2009). Gesture and metaphor comprehension: Electrophysiological evidence of cross-modal coordination by audiovisual stimulation. *Brain and Cognition, 70*(1), 42–52. http://dx.doi.org/10.1016/j.bandc.2008.12.005.

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. http://dx.doi.org/10.1111/j.1540-4781.2014.12124.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. http://dx.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. http://dx.doi.org/10.1162/jocn_a_00963.

Drijvers, L., & Ozyürek, A. (2017). Visual context enhanced: the joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language & Hearing Research, 60*, 212–222. http://dx.doi.org/10.1044/2016_JSLHR-H-16-0101.

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* (in press).

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.

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.

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.

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. http://dx.doi.org/10.1002/hbm.20753.

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. http://dx.doi.org/10.1162/jocn.2010.21462.

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. http://dx.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. http://dx.doi.org/10.1016/j.neuropsychologia.2015.04.018.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience, 19*(7), 1175–1192. http://dx.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. http://dx.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. http://dx.doi.org/10.3389/fpsyg.2012.00074.

Holle, H., Obleser, J., Rueschemeyer, S.-A., & Gunter, T. C. (2010a). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage, 49*(1), 875–884. http://dx.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 (pp. 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*. http://dx.doi.org/10.1093/scan/nsu047.

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. http://dx.doi.org/10.1007/s10919-008-0063-9.

Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language, 40*, 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. http://dx.doi.org/10.1162/jocn.2009.21254.

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. http://dx.doi.org/10.3758/s13423-014-0681-7.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89*(1), 253–260. http://dx.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. http://dx.doi.org/10.1080/01690965.2011.581125.

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. http://dx.doi.org/10.1016/j.bandl.2006.07.008.

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. http://dx.doi.org/10.1006/brln.1998.1979.

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology, 61*(5), 743–754. http://dx.doi.org/10.1037/0022-3514.61.5.743.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*(12), 463–470. http://dx.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). *Annual Review of Psychology, 62*, 621–647. http://dx.doi.org/10.1146/annurev.psych.093008.131123.Thirty.

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. http://dx.doi.org/10.3758/s13428-011-0146-0.

Macedonia, M., & Kriegstein, K. Von. (2012). Gestures enhance foreign. *Language Learning*, 393–416.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190. http://dx.doi.org/10.1016/j.jneumeth.2007.03.024.

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. http://dx.doi.org/10.1044/jslhr.4003.686.

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

Obermeier, C., Dolk, T., & Gunter, T. C. (2012b). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 48*(7), 857–870. http://dx.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. http://dx.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. http://dx.doi.org/10.1016/j.neuroimage.2010.12.020.

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. doi: 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. http://dx.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 of London. Series B, Biological Sciences, 369*(1651), 20130296. http://dx.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. http://dx.doi.org/10.1162/jocn.2007.19.4.605.

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. http://dx.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. http://dx.doi.org/10.1016/S0385-8146(01)00133-X.

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. http://dx.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*. http://dx.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. *Cerebral Cortex, 17*(10), 2387–2399. http://dx.doi.org/10.1093/cercor/bhl147.Hearing.

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. http://dx.doi.org/10.1371/journal.pone.0051207.

Straußß, A., Kotz, S. A., & Obleser, J. (2013). Narrowed expectancies under degraded speech: Revisiting the N400. *Journal of Cognitive Neuroscience, 25*(8), 1383–1395. http://dx.doi.org/10.1162/jocn.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*(4), 661–699. http://dx.doi.org/10.1111/j.0023-8333.2005.00320.x.

Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: An ERP study. *Neuropsychologia, 51*(13), 2847–2855. http://dx.doi.org/10.1016/j.neuropsychologia.2013.09.027.

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. http://dx.doi.org/10.1121/1.1512289.

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. http://dx.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. http://dx.doi.org/10.1016/j.neuroimage.2009.05.066.

Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology, 42*(6), 654–667. http://dx.doi.org/10.1111/j.1469-8986.2005.00356.x.

Wu, Y. C., & Coulson, S. (2007a). How iconic gestures enhance communication: An ERP study. *Brain and Language, 101*(3), 234–245. http://dx.doi.org/10.1016/j.bandl.2006.12.003.

Wu, Y. C., & Coulson, S. (2007b). 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.

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. Doi: 10.3389/fpsyg.2016.00908.