Journal of Speech, Language, and Hearing Research

Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension.

Journal:	*Journal of Speech, Language, and Hearing Research*
Manuscript ID	JSLHR-H-16-0101.R1
Manuscript Type:	Research Article
Date Submitted by the Author:	22-Jun-2016
Complete List of Authors:	Drijvers, Linda; Radboud Universiteit Nijmegen, Centre for Language Studies ; Radboud Universiteit Nijmegen Donders Institute for Brain Cognition and Behaviour, Donders Centre for Cognitive Neuroimaging Ozyurek, Asli; Radboud University , Centre for Language Studies ; Max-Planck-Institut fur Psycholinguistik, CLS-M; Radboud Universiteit Nijmegen Donders Institute for Brain Cognition and Behaviour, Donders Centre for Cognitive Neuroimaging
Keywords:	Noise, Language, Cognition, Speech perception, Speech

Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension.

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

*** Correspondence to:**

Linda Drijvers, Radboud University, Centre for Language Studies, Donders Institute for Brain, Cognition and Behaviour, Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands. E-mail: linda.drijvers@mpi.nl, telephone: +31 (0) 24 3521591

Conflict of interest:

The authors declare no conflict of interest.

Abstract

Purpose: This study investigated whether and to what extent iconic co-speech gestures contribute to information from visible speech to enhance degraded speech comprehension at different levels of noise-vocoding. Previously, the contributions of these two visual articulators to speech comprehension have only been studied separately.

Method: Twenty participants watched videos of an actress uttering an action verb and completed a free-recall task. The videos were presented in three speech (2-band; 6-band noise-vocoding; clear), three multimodal (Speech+Lips blurred; Speech+VisibleSpeech; Speech+VisibleSpeech+Gesture) and two visual only conditions (VisibleSpeech; VisibleSpeech+Gesture).

Results: Accuracy levels were higher when both visual articulators were present compared to one or none. The enhancement effects of a) visible speech, b) gestural information on top of visible speech and c) both visible speech and iconic gestures were larger in 6-band than 2-band noise-vocoding or visual only conditions. Gestural enhancement in 2-band noise-vocoding did not differ from gestural enhancement in visual only conditions.

Conclusions: When perceiving degraded speech in a visual context, listeners benefit more from having both visual articulators present compared to one. This benefit was larger at 6-band than 2-band noise-vocoding, where listeners can benefit from both phonological cues from visible speech, and semantic cues from iconic gestures to disambiguate speech.

Introduction

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

Iconic gestures are frequently prevalent in natural, face-to-face communication and have both a temporal and semantic relation with the speech they occur with, causing them to be hard to disambiguate without speech. It has been theorized that iconic gestures are an integral part of language (Kendon, 2004; McNeill, 1992): Speech and iconic gestures are

integrated continuously during comprehension, and target linguistic processing on semantic, syntactic and pragmatic levels (Holle et al., 2012; Kelly, Ozyürek, & Maris, 2010; McNeill, 1992, see for a review and meta-analysis: Hostetter, 2011). Previous research has shown that semantic information from iconic gestures is indeed processed by listeners and that iconic gestures can impact language comprehension, at behavioral and neural levels (e.g. Beattie & Shovelton, 1999; 2002; Holle & Gunter, 2007; Holler et al., 2014; Holler, Kelly, Hagoort, & Ozyurek, 2010; Holler, Shovelton, & Beattie, 2009; Kelly, Barr, Church, & Lynch, 1999; Kelly, Healey, Özyürek, & Holler, 2015; Obermeier, Holle, & Gunter, 2011 see for a review, Özyürek, 2014). For example, in an EEG study, Holle & Gunter (2007) showed participants videos of an actor who uttered a sentence while gesturing. Here, the experimental sentences contained an unbalanced homonym in the first part of the sentence (e.g. 'She controlled the ball'). This homonym was disambiguated in the subsequent clause (e.g. 'which during the game'/'which during the dance'). When the actor uttered the homonym, he would simultaneously produce an iconic gesture that either depicted the dominant (‘game’) or the subordinate meaning (‘dance’) of the homonym. When the gesture was congruent, they found a smaller N400 as compared to an incongruent gesture. This suggests that listeners use the semantic information from gestures to disambiguate speech.

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

the sentence. This effect was also found for hearing-impaired individuals (Obermeier et al., 2012). These results underline that speech-gesture integration can be modulated by specific characteristics of the communicative situation.

Another fMRI study by Holle, Obleser, Rueschemeyer & Gunter (2010) investigated the integration of iconic gestures and speech by manipulating the signal-to-noise ratio (SNR) of the speech to target areas that were sensitive to bimodal enhancement and inverse effectiveness (i.e. greater bimodal enhancement for unimodally least effective stimuli, i.e. the noisiest speech level). Participants watched videos of an actor with a covered face, who uttered short sentences (e.g. 'And now I grate the cheese') with or without an accompanying iconic co-speech gesture. These videos were presented with speech in a good SNR (+2 dB) or in a moderate SNR (-6 dB), using multitalker babble tracks. Their results revealed that the superior temporal sulcus and superior temporal gyrus in both hemispheres were sensitive to bimodal enhancement, and the neural enhancement for bimodal enhancement was even larger when participants were processing the speech and gestures in the degraded speech conditions. On both a neural as a behavioral level (i.e. response accuracy), this study showed that attending to a gesture under adverse listening conditions can significantly enhance speech comprehension and help in the disambiguation of a speech signal that is difficult to interpret. This gestural enhancement had already been described by Rogers (1978), who manipulated noise levels to show that gestures could only benefit speech comprehension when sufficient noise was added to the speech signal.

As Holle and colleagues (2010) note however, their study (and other studies, see e.g. Obermeier et al., 2011; Obermeier et al., 2012) have only focused on one visual articulator in speech-related audiovisual integration, namely iconic gestures. Other visual articulators, such as lip movements, were deliberately excluded from the stimuli that were used by blocking the actor's face with a black mask. Yet, these lip movements are inherently part of natural, face-

to-face communication: Lip movements can provide temporal information about the speech signal (e.g. on the amplitude envelope) and information on the spatial location of a speaker's articulators (e.g. place and manner of articulation), which can be specifically useful when perceiving speech in adverse listening conditions. Additionally, lip movements can convey phonological information, because of the form-form relationship between lip movements and syllables or segments that are present in the speech stream (for a recent review see Peelle & Sommers, 2015)).

The enhancement effect visible speech (consisting of lip movements, tongue movements and information from teeth) has on speech in clear and adverse listening conditions, has been reported by several studies (e.g. Erber, 1969, 1971; Ma, Zhou, Ross, Foxe, & Parra, 2009; Ross et al., 2007; Schwartz et al., 2004; Sumby & Pollock, 1954). Recognizing speech in noise is easier when a visual cue is present than when auditory information is presented alone, and has shown to improve recognition accuracy (Tye-Murray, Sommers, & Spehar, 2007). Previously, studies have argued that this beneficial effect increases as the SNR decreases (Sumby & Pollack 1954; Erber, 1969; 1975, Callan et al., 2003). However, more recent studies have reported that visual enhancement of speech by lip movements seems to be largest at "intermediate" SNR's where the auditory input is at a level between "perfectly audible" and "completely unintelligible" (Ross et al., 2007; 2008, Ma et al., 2009). This has also been reported by Holle et al., (2010), for gestural enhancement of speech in noise. Nevertheless, most studies on lip movements as a visual enhancement of speech have used stimuli that only showed the lips or lower half of the face (e.g. Callan et al., 2003; Ross et al., 2007; Schwartz, Berthommier, & Savariaux, 2004) to eliminate influences from the rest of the face or body. This is similar to studies in the domain of gestural enhancement of speech in noise, where most studies block the face of the speaker, the mouth,

or just show the torso of the speaker, to eliminate influences from visible speech (e.g. Obermeier et al., 2011; Obermeier et al., 2012; Holle et al., 2010).

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

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

visual articulators interact when both can enhance language comprehension, such as when speech is degraded.

The present study

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

We hypothesize that the enhancement from visible speech compared to speech alone (i.e. VisualSpeech enhancement: Speech+VisibleSpeech compared to Speech+Lips blurred) will be larger at an intermediate level of degradation compared to a severe level of degradation, allowing a listener to map the phonological information from visible speech to the speech signal. Additionally, we expect the enhancement from iconic gestures on top of visible speech (i.e. Gestural enhancement: Speech+VisibleSpeech+Gesture - Speech+VisibleSpeech) to be largest at an intermediate level of degradation compared to a severe level of degradation, which would indicate that a listener can benefit more from the semantic information from iconic gestures when there are more clear auditory cues to map this information too. Lastly, we predict that the enhancement of both articulators combined

(i.e. Double enhancement: Speech+VisibleSpeech+Gesture compared to Speech+Lips blurred), to be largest at an intermediate level of degradation compared to severe degradation. Since iconic gestures occur on top of information from visible speech, we expect that that should only be possible when enough auditory cues are available to the listeners. This way, listeners can benefit from both phonological information that is conveyed by visible speech, and from semantic information that is conveyed by iconic gestures.

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

Methods

Participants

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

Stimulus materials

We presented participants with 220 short video clips of a female, native Dutch actress uttering a Dutch action verb. The auditory and visual stimuli consisted of the Dutch high frequent action verbs, to make sure that the verbs could easily be coupled with iconic gestures. All video materials were recorded with a JVC GY-HM100 camcorder. Each recording of an action verb resulted in a video length of 2 seconds with an average speech onset of 680ms after video onset. All videos displayed the female actress from head to knees, appearing in the middle of the screen and wearing neutrally colored clothes (grey and black), in front of a unicolored and neutral background. Upon onset of the recording, the actress' starting position was the same for all videos. She was standing straight, facing the camera, with her arms hanging casually on each side of the body. During recording, she was instructed to utter the action verb while making a hand gesture that she found representative for the verb, without receiving feedback from the experimenter. The gestures she made were not instructed by the experimenter but were created by the actress on the fly. If the actress would have received explicit instructions per gesture, the gestures would have looked unnatural or choreographed, and the conscious effort to make a certain gesture could have drawn the attention to the participants explicitly to the gestures. All gestures that accompanied the action verbs were iconic movements for the actions that the verbs depicted (e.g. a drinking gesture resembling a cup that is raised towards the mouth for the verb 'to drink'). The preparation of all gestures started 120 ms after video onset, and the stroke (the meaning bearing part) of the gestures always coincided with the spoken verb.

The auditory sound files were intensity-scaled to 70 dB and de-noised in *Praat* (Boersma & Weenink, 2015). All sound files were re-combined with their corresponding video files in Adobe Premiere Pro. From each video's clear audio file, we created noise-vocoded degraded versions, using a custom-made script in *Praat*. Noise-vocoding effectively manipulates the spectral or temporal detail while preserving the amplitude envelope of the

speech signal (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). This way, the speech signal remains intelligible to a certain extent, depending on the number of vocoding bands, with more bands resulting in a more intelligible speech signal. We bandpass filtered each sound file between 50 Hz and 8000 Hz, and divided the signal into logarithmically spaced frequency bands between 50 and 8000 Hz. This resulted in cutoff frequencies at 50 Hz, 632.5 Hz and 8000 Hz for 2-band noise-vocoding and 50 Hz, 116.5 Hz, 271.4 Hz, 632.5 Hz, 1473.6 Hz, 3433.5 Hz and 8000 Hz for 6-band noise-vocoding. We used the frequencies to filter white noise in order to obtain six noise bands. We extracted the amplitude envelope of each band by using half-wave rectification. We then multiplied the amplitude envelope with the noise bands and recombined the bands to form the distorted signal.

In addition to clear speech, we included 2-band noise-vocoding and 6-band noise-vocoding in our experiment. In total, eleven conditions were created for the experiment (see Figure 1 for an overview). First, nine conditions were created in a 3 (Speech+Lips blurred, Speech+VisibleSpeech, Speech+VisibleSpeech+Gesture) by 3 (2-band noise-vocoding ('severe' degradation), 6-band noise-vocoding ('moderate' degradation), clear speech) design. Second, we added two extra conditions without sound (VisibleSpeech only, which is similar to lip reading, and VisibleSpeech+Gesture) to test how much information participants can resolve from visual input by itself. These conditions did not contain an audio file, so participants only could utilize the visual input. The final experimental set contained 220 videos with 220 distinct verbs that were divided over these eleven conditions (20 per condition) to test the different contributions of visible speech and gestures to clear speech comprehension and in these two degraded listening conditions.

[Figure 1]

Pre-test

To ensure that the verbs that we chose could be disambiguated by the iconic gestures that we recorded we conducted a pre-test to examine whether the gestures that the actress made in the video indeed depicted the verbs we matched them with in our audio files. In this experiment, twenty native Dutch speakers (10 female, M_{age} = 22;2, SD= 3,3) with no motor, neurological, visual, hearing or language impairments and who did not participate in the main experiment, were presented with 170 video stimuli that contained a gesture (not all 220 videos contained a gesture and these videos thus were used in the other conditions), but without the audio file that contained the verb. All stimuli were presented on a computer screen using Presentation software (Neurobehavioral Systems, Inc.), and presented in a different, randomized order per participant. First, participants were presented with a fixation cross for 1000 ms, after which the video stimulus started playing. After video offset, participants were asked to type down the verbs they associated the movement in the video with. After they filled out the verbs, we showed them the verb we originally matched it with in our auditory stimuli, and asked the participants to indicate on a 7-point-scale (ranging from "does not fit the movement at all" to "fits the movement really well") how iconic they found the movement in the video of the verb that was presented on the screen. This way, we could ensure that in the main experiment, the spoken verbs matched the gesture and participants could use the information from the gestures to disambiguate speech. If the gestures were not a good match with the verb, this gestural information would not enhance speech comprehension. All participants completed the task in approximately 35 minutes and could take self-paced breaks after every 55 items.

The typed answers on the first question of this pre-test ('Which verb do you associate with this video?') were used to determine which verbs had to be renamed to a possibly more occurring synonym, or which verbs were not recognizable and had to be discarded. We coded the answers either as 'correct', when the correct verb or a synonym was given, or as 'incorrect', when the input consisted of an unrelated verb. The results revealed a mean

recognition rate of 59% over all gesture videos. The percentage reported here indicates that the gestures are potentially ambiguous in the absence of speech, which is similar to how they are perceived in everyday communication (Krauss et al., 1991). Although this seems like a low overall consistency between participants, one must note that co-speech gestures, such as the iconic co-speech gestures used in these videos, normally occur in the presence of speech, and a higher overall percentage would have indicated that the gestures in our video were more like pantomimes, which are often understood and produced without speech. Since our study aims to understand the possible effects of iconic co-speech gestures on degraded speech comprehension, we did not use pantomimes.

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

Procedure

In our main experiment, participants were tested in a dimly-lit soundproof booth, and seated in front of a computer with headphones on. Before the experiment started, the experimenter gave a short verbal instruction that prepared the participant for the different videos that were going to be presented. All stimuli were presented full screen on a 1650x1080 monitor using Presentation software (Neurobehavioral Systems, Inc.), at a 70 cm distance in front of the participant. A trial started with a fixation cross of 1000, after which the stimulus was played. Then, in a free-recall task, participants were asked to type which verb they thought the actress tried to convey. After the participants typed in their answers, a new trial began after 500 ms. An answer was coded as 'correct' when a participant wrote down the correct verb, or minor spelling mistakes were made. Synonyms or category-related verbs (e.g. 'to bake' for 'to cook') were counted as incorrect.

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

Results

As a first step, we employed a 3 x 3 repeated measures analysis of variance with the factors Visual Articulator (Speech+Lips blurred; Speech+VisibleSpeech; Speech+VisibleSpeech+Gesture) and Noise-Vocoding Level (2-band noise-vocoding; 6-band noise-vocoding; clear speech) to subject the percentage of correct answers to. Note that we excluded the Visual Only conditions from this analysis (where we only tested VisibleSpeech and VisibleSpeech+Gesture, and not VisibleSpeech+Lips blurred, as this would result in a silent movie with no movement), since this would make our analysis unbalanced. As hypothesized, we found a significant main effect of Noise-Vocoding ($F(2,38) = 1569.78$, $p < .001$, $\eta^2 = .96$) indicating that the more the speech signal was noise-vocoded, the less correct

answers were given by the participants. We also found a main effect of VisualArticulator ($F(2,38) = 504.284$, $p < .001$, $\eta^2 = .98$) indicating that the more visual articulators were added to the signal, the more correct answers were given. In addition, we found a significant interaction between Noise-Vocoding level and VisualArticulator ($F(4,76) = 194.11$, $p < .001$, $\eta^2 = .91$), which seemed to be driven by the relatively higher amount of correct responses in the 6-band noise-vocoding condition compared to the other speech conditions (see Figure 2 for the percentages of correct responses per condition).

[Figure 2]

To further investigate this interaction, we compared the differences between and within the different noise-vocoding levels and visual articulators in a separate analysis. This analysis allowed us to compare the enhancement driven by different visual articulators as well as compare those enhancement effects between noise-vocoding levels. In comparing the enhancement from the different visual articulators, we recognized that calculating the absolute gain in terms of difference scores is limited in appropriately characterizing the maximum gain per condition. This is because there is an inverse relationship that exists between the performance in the Speech+Lips blurred and Speech+VisibleSpeech conditions and the maximum benefit that is derived when calculating the enhancement of the different visual articulators (see Grant & Walden, 1996). For example, we found a 2.75% recognition rate for Speech+Lips blurred in 2-band noise-vocoding as compared to 11.75% in 6-band noise-vocoding. The maximum gain possible on the basis of pure difference scores would therefore be 97.75% for 2-band noise-vocoding, and 88.25% for 6-band noise-vocoding, which would be hard to compare, since the maximal gain that is possible in 2-band noise-vocoding is larger than in 6-band noise-vocoding.

Therefore, to avoid possible floor effects and in keeping with previous studies, such as Sumby & Pollack (1954), we controlled for this by defining three difference scores ((A-B/100-B) (i.e., enhancement types)) for a) VisibleSpeech enhancement: Speech+VisibleSpeech - Speech+Lips blurred; b) Gestural enhancement: Speech+VisibleSpeech+Gesture - Speech+VisibleSpeech; and c) Double enhancement: Speech+VisibleSpeech+Gesture - Speech+Lips blurred, (see Ross et al., 2007 for a discussion of other calculation methods) divided by the maximal possible enhancement (for VisibleSpeech enhancement: 100 -Speech+Lips blurred; for Gestural enhancement: 100 - Speech+VisibleSpeech; for Double enhancement: 100 - Speech+Lips blurred). We subjected these outcomes to a repeated measures ANOVA with the factors Noise-Vocoding (2-band, 6-band, clear) and EnhancementType (VisibleSpeech enhancement, Gestural enhancement, Double enhancement). Our analysis revealed a main effect of Noise-Vocoding ($F(2,38) = 320.23$, $p < .001$, partial $\eta^2 = .94$), indicating that the more degraded the signal was, the less enhancement was present. Moreover, we found a main effect of EnhancementType ($F(1.06, 20.19) = 276.74$, $p < .001$, partial $\eta^2 = .94$, Greenhouse-Geisser corrected), indicating that the more visual information was present, the more participants answered correctly. Importantly, we found a significant interaction between EnhancementType and Noise-Vocoding ($F(1.97, 37.37) = 102.65$, $p < .001$, partial $\eta^2 = .84$, Greenhouse-Geisser corrected). Pairwise comparisons (all Bonferroni corrected) showed a significant difference between Gestural enhancement and VisibleSpeech enhancement in both the 2-band noise-vocoding condition ($t(19) = 9.41$, $p_{bon} < .001$ and the 6-band noise-vocoding condition ($t(19) = 12.94$, $p_{bon} < 0.001$) Furthermore, the difference between Gestural enhancement and VisibleSpeech enhancement was larger for 6-band noise-vocoding than 2-band noise-vocoding ($F(1,19) = 64.48$, $p_{bon} < .001$, partial $\eta^2 = .77$). Finally, Double enhancement was larger at 6-band noise-vocoding than in 2-band noise-vocoding and ($t(19) = -10.035$, $p_{bon} < .001$) (see Figure 3).

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

[figure 3]

Initially, we did not include the two Visual Only conditions (VisibleSpeech only, VisibleSpeech+Gesture) in our main analysis, because they would create an unbalanced design for analyzing all conditions together. However, these conditions were still of interest to determine how much information participants could obtain from visual input alone without speech being present. Therefore, we first tested the difference between the two separate Visual only conditions by means of a paired samples t-test. We found a significant difference between VisibleSpeech only and VisibleSpeech+Gesture ($t(19) = 15.12$, $p < 0.001$), indicating that response accuracy was higher for trials containing both visible speech and gestures, compared to videos that just contained visible speech (see Figure 2). Subsequently, we compared this difference between VisibleSpeech+Gesture and VisibleSpeech Only (i.e. gestural enhancement, computed as the difference between (VisibleSpeech+Gesture – VisibleSpeech only/100 - VisibleSpeech Only)) to the Gestural enhancement in the context of speech (Speech+VisibleSpeech+Gesture -Speech+VisibleSpeech/100 - Speech+VisibleSpeech) both in the 6-band and 2-band noise-vocoding conditions (see Figure 3). Our analysis revealed a significant difference between Gestural enhancement in the Visual

Only conditions and Gestural enhancement in 6-band noise-vocoding, ($t(19) = -3.23$, $p_{bon} < 0.05$), but not compared to 2-band noise-vocoding condition ($t(19) = 1.1$, $p_{bon} > .1$). These results confirmed that Gestural enhancement in 6-band noise-vocoding was significantly greater compared to 2-band noise-vocoding and compared to Gestural enhancement in the Visual only conditions. However, Gestural enhancement in the Visual Only conditions was not larger than Gestural enhancement in 2-band noise-vocoding, indicating that if there are no longer reliable auditory cues available (as in 2-band noise-vocoding), comprehension might be comparable to when there is no auditory input at all (as in Visual Only conditions).

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

Discussion

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

visible speech and iconic gestures present, as compared to having just visible speech present, or having only auditory information present. Here, gestures provide an additional benefit on top of the enhancement of visible speech.

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

In line with previous research, we found a significant benefit of adding information from visible speech to the speech signal (VisibleSpeech enhancement), in response to stimuli from both noise-vocoding levels (e.g. Sumby & Pollack, 1954). This benefit from solely visible speech was significantly larger at a moderate level of noise-vocoding (6-band) than at a severe level of noise-vocoding (2-band). Previously, it has been suggested that the benefit from visible speech continues to increase as the information that is available from auditory

inputs decreases (Sumby & Pollack 1954; Erber, 1969; 1975; Meredith & Stein, 1983), as would be predicted by the principle of inverse effectiveness. However, recent studies have argued that there are minimal levels of auditory information necessary before recognition accuracy can be most enhanced by congruent visible input (Ross et al., 2007). Our data concurs with this latter idea, by finding an optimal range for multimodal integration and enhancement, where auditory cues are moderately reliable, and enhancement from visible speech has its maximal effect.

Importantly, the current results provide novel evidence by showing that iconic gestures can enhance this benefit from visible speech even more: We found a significant difference between Gestural enhancement (Speech+VisibleSpeech+Gesture – Speech+VisibleSpeech) and VisibleSpeech enhancement (Speech+VisibleSpeech – Speech+Lips blurred) at both noise-vocoding levels. In addition, we found significant differences between Double enhancement and Gestural enhancement, as well as significant differences between Double and VisibleSpeech enhancement at both noise-vocoding levels. Our results therefore suggest that although both visual modalities enhance degraded speech comprehension, having both iconic gestures and visible speech present (Double enhancement) in the input enhances speech comprehension most. This is in line with previous literature on the benefits of gestures in language processing and theories of communication that postulate that multimodal information combines with speech to aid language comprehension (McNeill 1992; Clark 1996; Goldin-Meadow 2003, see for a review Kelly, Manning, & Rodak, 2008). Interestingly, the enhancement of both visual articulators (Double enhancement) was significantly larger than VisibleSpeech enhancement at both noise-vocoding levels. This suggests, in line with previous research, that gestures are actively processed and integrated with the speech signal (Kelly et al., 2010; Kendon, 2004), even under conditions where visible speech is visible (also see Holler et al., 2014).

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

In earlier work on the contribution of visible speech and hand gestures to learning non-native speech sounds, Kelly et al. (2008) argued that lip and mouth movements help in auditory encoding of speech, whereas hand gestures only can help to understand the meaning of words in the speech stream when the auditory signal is correctly encoded. Based on their

results, Kelly et al. (2008) argue that the benefits of multimodal input target different stages of linguistic processing. Here, mouth movements seem to aid during phonological stages, whereas hand gestures aid during semantic stages, which, according to the authors, fits with McNeill's (1992) interpretation of speech and gesture forming an integrated system during language comprehension.

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

gestures (i.e. Gestural enhancement) to resolve the form of the speech signal. Based on these results, we suggest that at least in adverse listening conditions where auditory cues are no longer reliable, language processing might be more driven by semantic information that is abstracted from iconic co-speech gestures.

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

gestures and visible speech. Instead, research can tap into the influence of other nonverbal behavior (such as head and brow movements, see e.g., Krahmer & Swerts, 2007) and their co-occurrence with visible speech and gesture to fully understand the optimal conditions for visual enhancement of speech in adverse listening conditions. This, in turn, can further elucidate the results from the current study, but also inform debates on audiovisual training for both clinical populations and educational instruction. Finally, replicating the effects found in this study with hearing-impaired populations will provide a better diagnosis of their speech comprehension in ecologically valid contexts (i.e., in a multimodal context). This in turn could inform debates on audiovisual training for both clinical populations and educational instruction.

Acknowledgements

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

References

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica, 1,* 32-49.

Beattie, G., & Shovelton, H. (1999). 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. doi: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. doi:10.1348/000712602162526

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer. [Computer program]. Version 6.0.19, retrieved 13 June 2016 from http://www.praat.org/

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, *14*(17), 2213–8. doi:10.1097/01.wnr.0000095492.38740.8f

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423–5.

Erber, N. P. (1971). Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by children with impaired hearing. *Journal of Speech and Hearing Research*, *14*(3), 496–512.

Goldin-Meadow, S. (2005). *Hearing Gesture: How Our Hands Help Us Think*. Harvard University Press.

Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for auditory-visual consonant recognition. *The Journal of the Acoustical Society of America*, *100*(4), 2415–24.

Hirata, Y., & Kelly, S. D. (2010). Effects of Lips and Hands on Auditory Learning of Second-Language Speech Sounds. *Journal of Speech, Language and Hearing Research*, *53*(4), 298–310.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, *19*(7), 1175–92.

doi:10.1162/jocn.2007.19.7.1175

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*(3), 74. doi:10.3389/fpsyg.2012.00074

Holle, H., Obleser, J., Rueschemeyer, S.-A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage*, *49*(1), 875–84. doi: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. In *the 34th Annual Meeting of the Cognitive Science Society (CogSci 2012)* (pp. 467-472). Cognitive Society.

Holler, J., Schubotz, L., Kelly, S., Hagoort, P., Schuetze, M., & Özyürek, A. (2014). Social eye gaze modulates processing of speech and co-speech gesture. *Cognition*, *133*(3), 692–7. doi:10.1016/j.cognition.2014.08.008

Holler, J., Shovelton, H., & Beattie, G. (2009). Do Iconic Hand Gestures Really Contribute to the Communication of Semantic Information in a Face-to-Face Context? *Journal of Nonverbal Behavior*, *33*(2), 73–88. doi:10.1007/s10919-008-0063-9

Hoskin, J., & Herman, R. (2001). The communication, speech and gesture of a group of hearing-impaired children. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, *36 Suppl*, 206–9.

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, *137*(2), 297–315. doi:10.1037/a0022128

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. *J.Mem. Lang*, *40*, 577–592.

Kelly, S. D., Manning, S. M., & Rodak, S. (2008). Gesture Gives a Hand to Language and Learning: Perspectives from Cognitive Neuroscience, Developmental Psychology and Education. *Language and Linguistics Compass*, *2*(4), 569–588. doi:10.1111/j.1749-818X.2008.00067.x

Kelly, S. D., Ozyürek, A., & Maris, E. (2010). Two sides of the same coin: speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–7. doi:10.1177/0956797609357327

Kelly, S., Healey, M., Özyürek, A., & Holler, J. (2015). The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review*, *22*(2), 517–23. doi:10.3758/s13423-014-0681-7

Kelly, S., Hirata, Y., Simester, J., Burch, J., Cullings, E., & Demakakos, J. (2008). Effects of hand gesture and lip movements on auditory learning of second language speech sounds. *The Journal of the Acoustical Society of America*, *6*(10), 2357–2362. doi:10.1121/1.2933816

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396-414.

Krauss, R M., Morrel-Samuels, P. & Colsante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology,* 61 (5), 743 - 754.

Lombard, E. (1911). Le signe de l'elevation de la voix. *Annals Maladiers Oreille, Larynx,*

Nez, Pharynx, 37, 101–119.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PloS One*, *4*(3), e4638. doi:10.1371/journal.pone.0004638

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

Obermeier, C., Dolk, T., & Gunter, T. C. (2012). 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), 857–70. doi: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), 1648–63. doi:10.1162/jocn.2010.21498

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. doi:10.1098/rstb.2013.0296

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *68*, 169–81. doi:10.1016/j.cortex.2015.03.006

Rogers, W. T. (1978). the Contribution of Kinesic Illustrators Toward the Comprehension of Verbal Behavior Within Utterances. *Human Communication Research*, *5*(1), 54–62. doi:10.1111/j.1468-2958.1978.tb00622.x

Ross, L. a, Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy

environments. *Cerebral Cortex (New York, N.Y. : 1991)*, *17*(5), 1147–53. doi:10.1093/cercor/bhl024

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69–78. doi:10.1016/j.cognition.2004.01.006

Shannon, R., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, *270*(5234), 303–304.

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), 661–7. doi:10.1016/j.cub.2009.02.051

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

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing*, *28*(5), 656–68. doi:10.1097/AUD.0b013e31812f7185

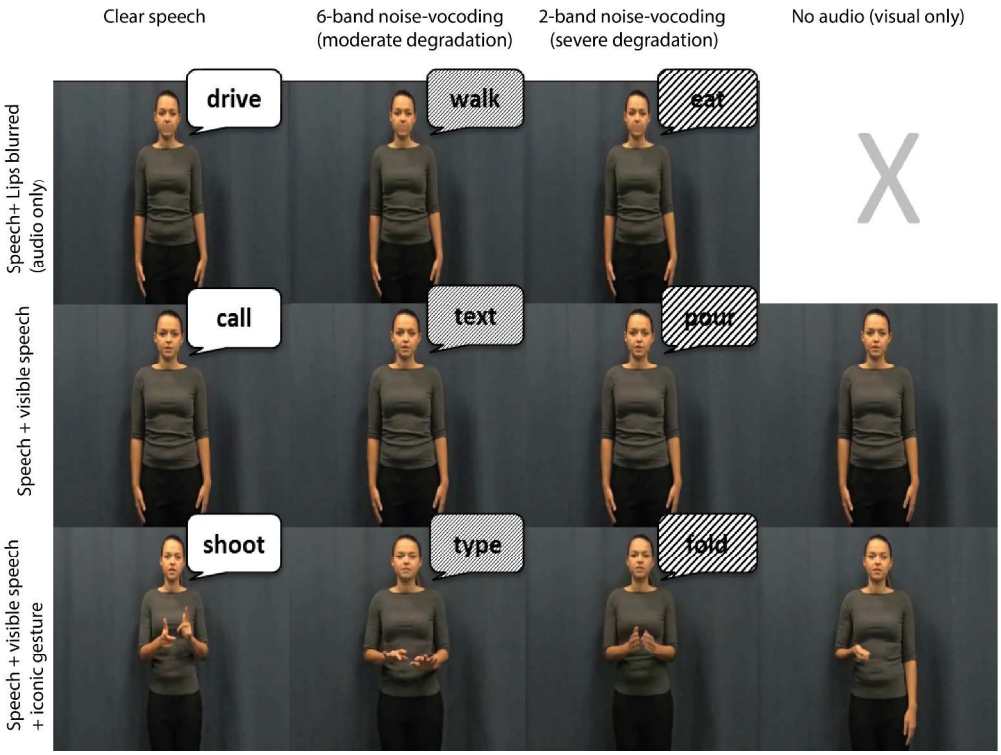
Wu, Y. C., & Coulson, S. (2007). Iconic gestures prime related concepts: an ERP study. *Psychonomic Bulletin & Review*, *14*(1), 57–63. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17546731

Figure captions

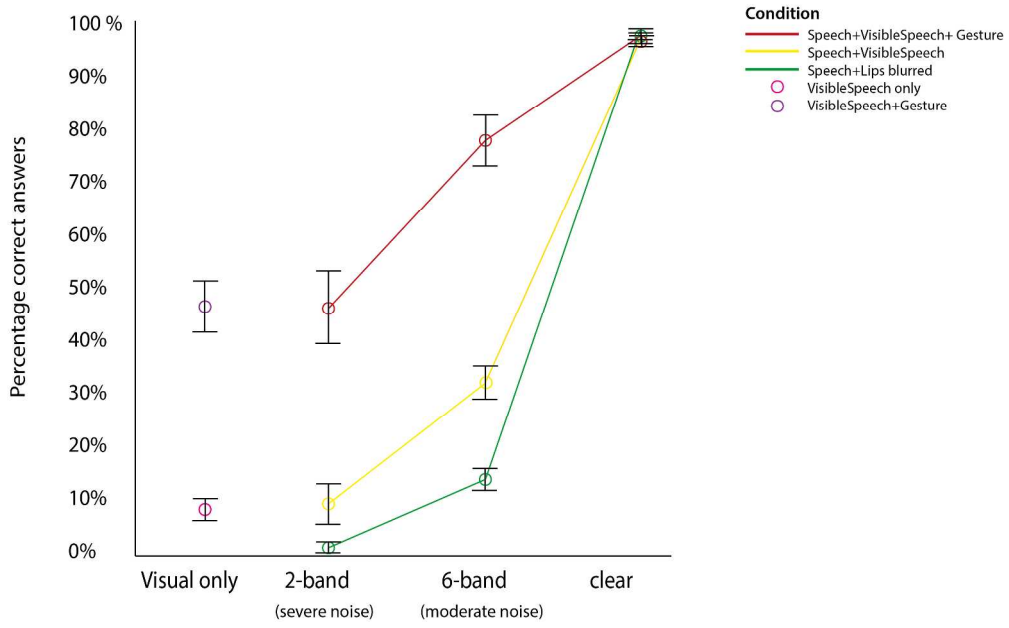
Figure 1: *Overview of the design and conditions used in the experiment.*

Figure 2: *Percentage of correctly identified verbs (% correct) per condition. Error bars represent standard deviations.*

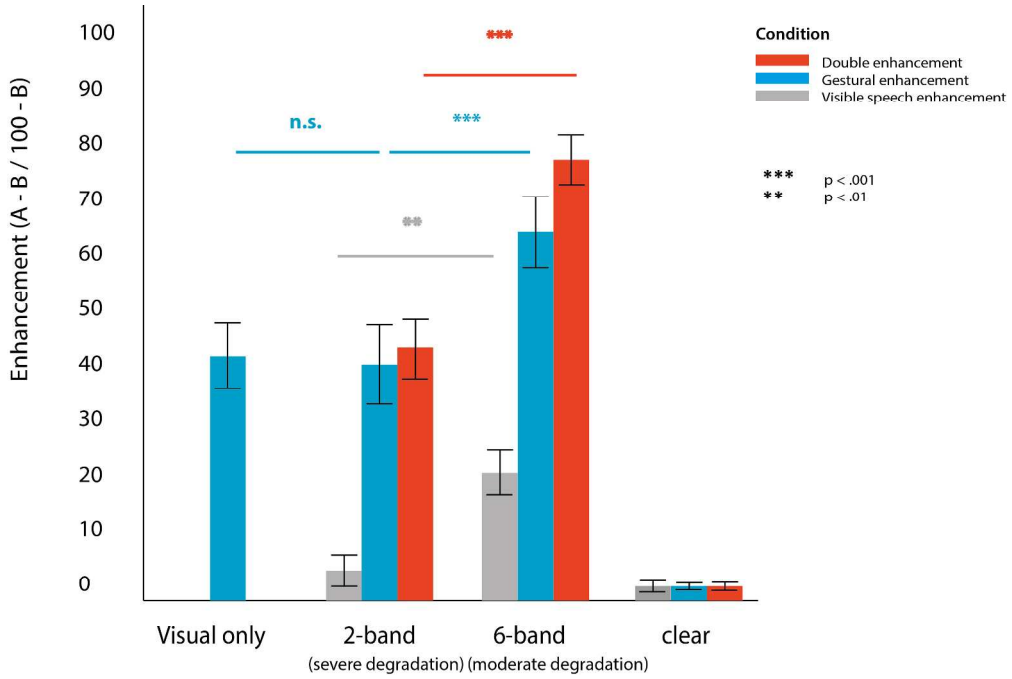
Figure 3*: Enhancement effect (A-B/100-B) corrected for floor effects. Error bars represent standard deviations.*



Overview of the design and conditions used in the experiment
365x282mm (300 x 300 DPI)



Percentage of correctly identified verbs (% correct) per condition. Error bars represent standard deviations.
268x220mm (300 x 300 DPI)



Enhancement effect (A-B/100-B) corrected for floor effects. Error bars represent standard deviations.
256x170mm (300 x 300 DPI)