The effect of mouth visibility and/or gestures on degraded speech comprehension

Kirstin Bleumink
s1024689
30 June 2021
BA Thesis
Bachelor Linguistics

Supervisors
Prof. Dr. Asli Özyürek
Dr. Linda Drijvers

Radboud University

Preface

In front of you, you can find my bachelor thesis on the effect of mouth visibility and/or gestures on degraded speech comprehension. This research has been conducted as a part the Bachelor's programme Linguistics at the Radboud University Nijmegen. After taking some courses in psycholinguistics and neurolinguistics, my interest in this domain was immediately sparked. After doing my internship at the Max Planck Institute on this subject, I decided I wanted to continue working on this subject for my thesis.

I hope you will read this thesis with great pleasure and that afterwards you will know more about how people not only perceive speech when they hear, but also incorporate visual information.

This thesis would not have been the same if it were not for my supervisors Linda Drijvers and Asli Özyürek. I am grateful for their enthusiasm on this subject and highly appreciate their wonderful feedback and support during the process of writing this thesis.

Besides my supervisors, I also want to express my gratefulness towards all the members of the Communicative Brain Lab – a research group to which I was introduced during my internship. I absolutely loved your supportive attitude and how you made me feel welcome in the group. I hope we will meet again in the future!

At last, I would like to thank my friends and family. You were always there for me when I needed you the most.

Kirstin Bleumink
Nijmegen, 30 June 2021

Table of Contents

Abstract

Speech comprehension does not only exist of processing the auditory information. Usually, visual information is also integrated with the auditory information, also called audiovisual integration. The most common kinds of visual cues used in audiovisual integration are visible speech and (iconic) gestures. Previous research has studied the contribution of the addition of visible speech to speech without visual articulators, the contribution of the addition of gestures on top of visible speech, and the contribution of the addition of those two visual articulators to speech without visual articulators on degraded speech comprehension (Drijvers and Özyürek, 2017). However, there had not been made a comparison between the added value of visible speech to speech without visual articulators and that of gestures to speech without visual articulators. Furthermore, the enhancement effects of adding visible speech to gesture had not been studied and compared to the enhancement effect of adding gestures to visible speech before. This research is thus aimed at comparing the benefit of visible speech alone and that of gestures alone separately in disambiguating speech in noise and to determine how these benefits relate to the benefit of having both visual articulators in a joint context using the same paradigm as in Drijvers and Özyürek (2017). In an online word identification task, participants were presented with videos that included clear or degraded speech (6-band noise-vocoded) or no speech. Clear and degraded speech occurred in no visible speech & no gesture, visible speech & no gesture, no visible speech & gesture, visible speech & gesture contexts. The no speech condition was presented in visible speech & no gesture, no visible speech & gesture, and in visible speech & gesture contexts.

Three different repeated measures ANOVAs were conducted. The results of the first analysis showed that word identification was better if 1) speech quality was clear compared to degraded, 2) visible speech was present compared to when it was not, and 3) gestures were present compared to when they were not. Next, the enhancement effects on degraded speech comprehension that are due to the addition of the visual articulators – either in isolation or in joint context – were analysed. What is found is that the addition of gesture to speech without visual articulators is more beneficial to degraded speech comprehension than the addition of visible speech to speech without visual articulators. Besides that, the addition of gesture to speech that is already visible also enhances degraded speech comprehension more than the addition of visible speech to speech that is already accompanied by gestures. In short, iconic gestures enhance degraded speech comprehension more by itself and in a joint context than visible speech. In the final repeated measures ANOVA analysis two types of enhancement effects (addition of gesture to visible speech and the other way around) were compared in degraded speech and no-speech conditions.

In the no speech condition, the enhancement effect due to the addition of gestures to visible speech was similar to the enhancement effect of the addition of gestures to visible speech in degraded speech. What did differ however, was that the enhancement of the addition of visible speech to gesture was bigger for the no speech condition than for the degraded speech condition.

Thus, all results indicated that the contribution of iconic gestures to degraded speech comprehension is higher than that of visible speech in isolated or joint contexts.

1. Introduction

During everyday communication people use speech to transmit a message from one person to another. However, people do not only use auditory information to accomplish this. Visual information goes along with speech. The receiver of the information thus listens to the auditory information and sees the visual information and next integrates this to get an entire picture of the message that the speaker tries to convey. When auditory and visual information are integrated, this is called audiovisual integration. Occasionally, it is also referred to as multisensory integration, which is a more generic term for the integration of all kinds of sensory information (thus not specifically auditory and visual information).

This audiovisual integration can be seen in most conversations. The auditory information is the speech, and the visual information for example can be the mouth movements made while uttering the speech and co-speech gestures. Auditory information can be compromised, for example because of background noise or strong accents. Auditory information can be degraded in at least two ways: degradation of the speech signal itself (e.g., by noise-vocoding) and embedding the speech signal in background noise. Both can have different degrees of degradation. For noise-vocoding the less frequency bands are chosen, the less intelligible the speech signal is. For a speech sound that is embedded in background noise the signal-to-noise ratio (henceforth SNR) is used to indicate the amount of noise. An interesting question is when the auditory information is degraded, whether and how much visual information, in specific the mouth movements made when speaking and/or co-speech gestures, will improve speech comprehension in these situations. Also, the enhancements from visual articulators need to be studied more profoundly to see whether the enhancement because of mouth movements differs from that of gestures and if this differs from the enhancement of having both visual articulators and what each articulator contributes to enhancement of speech comprehension. Both visual articulators differ in that they give the listener information about the speech on a different level: the mouth movements give information about the phonological features of the word pronounced, while co-speech gestures give information about the semantics of the pronounced word. The present thesis aims to disentangle single versus joint contribution of visible speech and gestures to degraded speech comprehension.

Below previous research on the effect of the visibility of mouth movements on speech comprehension (section 1.1 Speech and Visible speech), on the effect of gestures on speech comprehension (section 1.2 Speech and Iconic Gestures), and on the effect of both visual articulators (section 1.3 Speech and Multiple Visual Articulators) are described regarding clear and/or degraded speech comprehension before giving more detailed information about the aim and hypotheses of the present study (section 1.4).

1.1 Speech and Visible Speech

When the articulatory movements of the mouth are visible, either when there is speech present or there is no audio, this is called presence of visible speech. Visible speech (also referred to as visual speech) does not give new phonological information on top of the auditory information, but in similar speech sounds it can play a role in disambiguation (Peelle & Sommers, 2015). Visible speech can influence speech perception.

One of the first examples of visible speech influencing what is heard, is the McGurk effect (McGurk & MacDonald, 1976). When people were presented with the syllable [ba] in audio, but saw the lips making the syllable [ga] (or vice versa), people responded to hearing the syllable [da]. When they excluded one of the two modalities, either by covering ears or eyes, they reported the correct syllable. This fused syllable is the result of trying to integrate the auditory with the visual information. The speech quality can also influence audiovisual integration. For example, the McGurk effect can be influenced by degradation of the auditory

or visual stimuli: when adding noise to the auditory stimulus, the McGurk effect increased, and when blurring the lips, the McGurk effect decreased (Stacey et al., 2020).

The presence of visible speech can not only influence the perceived speech, but it can also improve the detection of speech (sentences) in noise as shown in Grant and Seitz (2000). In one of their experiments, participants were presented with three target sentences which they had to detect in noise under three conditions: auditory information only, auditory information plus congruent visible speech, and auditory information plus incongruent visible speech. When the auditory information was combined with congruent visible speech, the auditory detection of the target sentences in noise was better than when only auditory information was present. This implies that speech was perceived louder when the auditory information is accompanied by visual information. Similar results were also found in Bernstein et al. (2004).

Audiovisual information also enhances the intelligibility of speech in noise (e.g., Schwartz et al., 2004). This indicates that speech in a noisy environment can be best understood when the speech is also visible. This result was also found in a study by Blackburn et al. (2019), in which participants watched videos in which a sentence was uttered that had to be repeated afterwards. The SNR was varied depending on the quality of the repetition of the sentence. When at least three out of the five keywords were correct the SNR was reduced and if this was not the case the SNR was increased again. The target sentences existed of either clear speech or degraded (vocoded) speech, while in the background babbling of three different amounts (1, 2 or 16) of talkers was heard. The most important result was that visible speech enhanced speech comprehension more when the target sentence was noise-vocoded than when it was clear. Next to that, they also investigated whether the influence of audiovisual integration differed depending on the intelligibility of the speaker, since not all speakers are always equally intelligible. Hence, the sentences in the conditions were uttered by four (in clear speech condition) and three (in vocoded speech condition) different speakers. Based on the performance of the participants in the audio-only condition in clear speech, the speakers were ordered to intelligibility. What was shown was that the intelligibility of the different target talkers resulted in different amounts of visible speech benefit, in the way that a target talker with low intelligibility had a larger visible speech benefit.

Similar to Blackburn et al. (2019), Sumby and Pollack (1954) have also concluded that visible cues can provide a higher absolute gain in speech comprehension of words in noisy environments than in environments in which the speech is very clear. This is because in clear speech conditions the room for improvement is smaller than for noisy speech, because the intelligibility of clear speech is already high in contrast to degraded speech.

Ross et al. (2007) have drawn an even more balanced picture about the intelligibility in noisy environments, after an experiment in which participants had to report the word that was said in the video with differing levels of noisiness (as indicated by differing SNR). The more negative the SNR, the noisier the environment, thus meaning that an SNR of -24 dB indicates a severe noisy environment, while an SNR of 0 dB indicates clear speech. They concluded that the previously suggested inverse effectiveness for visible speech (Stein & Meredith, 1993, in Callan et al., 2003) – which would implicate that the gain from visual speech would be highest in the most degraded speech with the lowest intelligibility – was not nuanced enough. In fact, while the results of Ross et al. (2007) showed that in severe noisy environments the speech comprehension benefits substantially from the visibility of speech, it showed that in intermediate (SNR of -12 dB) noisy environments the benefit was even bigger than what was expected based on the inverse effectiveness principle. This was seen as an indication that in intermediate, and not severe noise the speech comprehension benefits most from visible speech. They explained this by the fact that at the extremes the focus would be on either the visible speech (in a severe noisy environment – SNR of -24 dB) or on the auditory information (in clear speech – SNR of 0 dB). However, at an intermediate (SNR of -12 dB) level of noise the

visible speech and auditory information would be maximally integrated and thus the enhancement of visible speech will be maximal.

1.2 Speech and Iconic Gestures

Visible speech is not the only visual articulator available in everyday conversations. Many speakers also use gestures while speaking, in which case they are also called co-speech gestures. One type of gestures that can co-occur with speech is that of iconic gestures. An iconic gesture, as defined by McNeill and Levy (1982), 'is a formed gesture which depicts in its form or manner of execution aspects of the event or situation being described verbally' (p. 275). A clear example of an iconic gesture would be the forming of the hands as if they hold on to a steering wheel and moving these as would be the case in taking a turn on the road to represent the action verb 'to drive'. Iconic gestures not only co-occur with speech, but they also influence each other. Since gestures are generally seen as a part of natural communication, their influence has been mostly studied when participants hear and see a speaker uttering the sentence. Some of the influence of gestures might therefore be biased because of the presence of visible speech. There is only a limited amount of research that studied the influence of gestures in isolated form.

Kelly et al. (2010) performed a priming experiment, in which speech and gesture was congruent and related to an action prime or either the speech or the gesture was incongruent (weakly or strong) whilst the other was related to the action prime. Participants had to indicate if a part of the target was related to the prime. Only the upper torso and arms of the actors who produced the action prime and the gesture were visible in the stimuli used during this experiment. Therefore, there was no influence from visible speech possible. The results showed that when gesture and speech were incongruent, more errors were made, and their reaction time was slower than when speech and gesture were congruent. In incongruent conditions the gesture interfered with the speech, which caused people to feel like something different was being said than what in fact was conveyed in a way similar to the earlier described McGurk effect (McGurk & MacDonald, 1976). This integration of gesture and speech also seemed to be obligatory; people took in visual information from the gestures, even when they were told to focus on the speech. Both findings were seen as support for the integrated-systems hypothesis. Iconic gestures thus play a role in the comprehension of speech (Özyürek, 2014).

In the study of Obermeier et al. (2012), using the material of Holle and Gunter (2007), participants looked and listened to videos in which an actress uttered complex sentences that consisted of a main clause that contained a homonym (with two meanings differing in dominance) which was disambiguated by a gesture. Then in a sub-clause a target word was uttered that either supported the dominant or the subordinate meaning. This was used to indicate how well the gesture had disambiguated the homonym. In these videos the face of the actress uttering the sentences was covered, thus eliminating the influence of the presence of visible speech. The results of the EEG-experiments showed that in environments with degraded speech (either by background noise or in hearing impairments) the gestural information was integrated with the homonym, while this was not the case in clear speech. Holle et al. (2010) also looked into the effect of iconic gestures (in isolated form because of the actress being masked) on speech comprehension in speech with differing SNRs. They found that – similar to the enhancement of visible speech in noisy environments – the gestural enhancement is also higher in intermediate noisy environments (background noise 4 dB louder compared to the signal) than when it is a clear or severely noisy environment.

The research summarized above on gestures however did not compare influence of visible speech versus the iconic gestures or their joint contribution on disambiguating noise.

1.3 Speech and multiple visual articulators

Despite previous research on either visible speech or gestures not much research has been done on the presence of multiple visual articulators, such as both visible speech and gesture, as both are common in everyday face-to-face communication. Drijvers and Özyürek (2017) researched for the first time in what way iconic gestures and visible speech together enhance speech comprehension in degraded speech. More specifically, they investigated whether adding a gesture to speech which was visible resulted in an even larger enhancement of degraded speech comprehension than the enhancement of having visible speech present. To investigate this, they showed videos to the participants in which an action verb was uttered. The speech quality differed from clear, moderately degraded, severely degraded, to no audio. The degraded speech was made using noise-vocoding (for more information see Drijvers & Özyürek (2017) or section 2.2 Stimuli); they used 6-band noise-vocoding for moderately degraded speech and 2-band noise-vocoding for severely degraded speech. There were three conditions regarding the visual articulators: no visual articulators (lips blurred), visible speech, and visible speech + iconic gesture. Their results showed that speech comprehension in degraded speech was better when both visual articulators were present than when only visible speech was available or when there was no visual articulator present. Even if a visual articulator was already present, in this case visible speech, adding a second visual articulator – in this case gesture – did still improve the comprehension of degraded speech even more. The addition of gesture to speech that is visible led to a larger enhancement effect than the addition of visible speech to speech without any visual articulators in degraded speech. Double enhancement (addition of visible speech and gestures to speech without any visual articulators) was then again larger than both of those enhancement effects.

Further research on this with non-native speakers has shown that the enhancement effects of the visual articulators remain highest in moderately degraded speech, but the enhancement effects of the visual articulators (or the combination of both) is lower for non-native speakers than for native speakers (Drijvers & Özyürek, 2020). Schubotz et al. (2020) looked into the effects of the visual articulators (presence of visible speech and iconic gestures) in younger and older people. For this investigation they used comparable stimuli to Drijvers and Özyürek (2017), but instead of noise-vocoding to create degraded speech, they used speech-in-noise, (SiN) where the signal was embedded in multitalker babble (two different SNRs). This study demonstrated that both groups benefit from having multiple visual articulators (gesture on top of visible speech) available, even though elderly did less.

However, when multiple visual articulators are present, it is not necessarily the case that both visible speech and gesture are equally important audiovisual cues for natural language comprehension. Zhang et al. (2020) also looked into the influence of having multiple visual articulators (and besides that also word predictability and prosody) in a joint context on language comprehension and to the importance of each multimodal cue in language comprehension. As stimuli they used videos of an actress uttering sentence pairs. In these sentence pairs the second sentence had to contain a verb that could be easily paired with a gesture. In half of the videos presented to the participant a gesture was in fact realised. Per video the word predictability and multimodal cues (prosodic information, gestures, and mouth informativeness) that were present were quantified. Word predictability was measured by surprisal; which can be calculated by taking the negative log of the odds of finding this exact word based on the words before. As a measurement for the prosodic information the mean F0 was used. The gestures made were divided into two categories: meaningful gestures (encloses iconic gestures, but also deictic gestures) and beat gestures. For determining the mouth informativeness an additional experiment was conducted in which participants had to indicate what was being said when no audio was available. The mean percentage correct was then used as a measurement of mouth informativeness. The results of Zhang et al. (2020) showed that

prediction of words, as measured by N400 amplitude, during sentence processing is modulated by multimodal cues, especially by prosodic information and gestures. Mouth movements only participated in interaction effects with other multimodal cues and the surprisal of words. Though multimodal cues modulate the prediction of words, not all multimodal cues presented were always equally important. For example, when both visual articulators were present, the benefit of a gesture was smaller when the visible speech was already highly informative than when visible speech was less informative. This showed that the influence of these visual articulators was therefore not always constant. Despite this dynamic nature they still did find a hierarchical structure of the importance of all multimodal cues. After prosodic information, gestural information was considered to be more important for language comprehension than visible speech. This research however did not look at the weighing of different multimodal cues in degraded speech or noisy environments.

1.4 Present study

As the previous research described above shows, the benefit of the presence of visible speech on degraded speech comprehension has been mostly studied in an isolated form, so the mouth movements as the only visual information available (Stacey et al., 2020; Grant & Seitz, 2000; Schwartz et al., 2004; Sumby & Pollack, 1954; Ross et al., 2007). Meanwhile, some other research has looked at the benefit from gestures on degraded speech comprehension in isolation as well. Obermeier et al. (2012), Holle et al. (2010), and Kelly et al. (2010) are a few studies in which gestures were studied while visible speech was (made) invisible. However, these studies only focused on either one of the visual articulators without making a direct comparison between these two articulators or the joint contribution of the two, especially regarding degraded speech comprehension.

Only recently a study has looked at the joint effect of both visible speech and gestures. The study by Drijvers and Özyürek (2017) investigated the enhancement of these visual articulators in a joint context. They investigated the benefit of the different numbers of visual articulators (0 – speech; 1 – speech + visible speech; 2 – speech + visible speech + gesture) on degraded speech comprehension, and thus made a direct comparison between the enhancement of visible speech in isolated form and that of visible speech and gesture in a joint context. However, since no condition was present with gestures without visible speech, they have not looked into the effect of having only gesture present in comparison to the effect of only visible speech present and the effect of joint context on either single context. Zhang et al. (2020) did make a comparison between the impact of gesture or visible speech as a multimodal cue when both were present. Nevertheless, it was only studied from the perspective of them co-occurring. Besides that, this study was not conducted in degraded speech comprehension. Furthermore, even though Drijvers and Özyürek (2017) did investigate the effect of adding gesture as a second visual articulator to speech which was also visible, they could not look into the effect of adding visible speech as a second visual articulator to speech accompanied with gestures because of the missing gesture only condition.

This current study is aimed at comparing the benefit of either visible speech or iconic gestures on degraded speech comprehension and to determine if – and in case it does in what way – this benefit differs from the benefit of the two visual articulators in a joint context. The present study will look at speech comprehension in the following conditions: visible speech + gesture, no visible speech + gesture, visible speech + no gesture, no visible speech + no gesture in both clear and degraded speech and will look into the first three conditions in a no speech condition as well. The visible speech + gesture, visible speech + no gesture, and no visible speech + no gesture conditions were also used by Drijvers and Özyürek (2017), but the no visible speech + gesture condition is an addition to their study. While in Drijvers and Özyürek (2017) there were also two levels of degradation (severe – 2-band noise-vocoding and

intermediate – 6-band noise-vocoding), in this study only intermediate degradation is used as highest enhancement effects for visual articulators were found there. The no speech condition is present in order to provide insight into the amount of information that can be extracted from the visible speech or the gestures alone. A more detailed and graphically illustrated description of the conditions used can be found in section 2.3. Addition of this gesture-only condition will provide a more complete picture on the effect of the two visual articulators gesture and visible speech, including the two of those in a joint context. Besides that, in current time of COVID-19 pandemic, especially understanding the effect of solely adding gesture to speech when visible speech is not present and in speech degradation due to wearing a mask might be of high societal relevance because of the frequent use of mouth masks.

Based on previous research (Schwartz et al., 2004; Sumby & Pollack, 1954; Ross et al., 2007; Drijvers & Özyürek, 2017) it is to be expected that speech comprehension in degraded speech is better when the speech is accompanied by visible speech than when there is no visible speech present. Besides that, it is also to be expected that speech comprehension will be better when the speech is accompanied by iconic gestures than when there are no gestures present (Kelly et al., 2010; Drijvers et al., 2018; Obermeier et al., 2011; Holle et al. 2010). Which visual articulator provides the largest enhancement when it is the only available visual articulator is not known yet. However, since Drijvers and Özyürek (2017) found a larger enhancement for their 'gestural enhancement' (addition of gesture to speech that is already visible) than for their 'visible speech enhancement' (addition of visible speech to speech without visual articulators), it is expected that the enhancement of the addition of gesture to speech without visual articulators is larger than that of the addition of visible speech to speech without visual articulators. It is also expected that having both articulators will enhance degraded speech comprehension more than any one of the articulators.

Based on Drijvers and Özyürek (2017) it is expected that the enhancement effect of addition of gesture to visible speech is bigger than the enhancement effect of addition of visible speech to speech without visual articulators. Zhang et al. (2020) showed that when both visual articulators are present, the influence of visible speech is smaller than that of gestures. Therefore, it is also expected that there is a larger enhancement effect for addition of a gesture to visible speech than for addition of visible speech to gestures. In replication of Drijvers and Özyürek (2017) it is expected that also in the present study the biggest enhancement effect will occur when both visible speech and gestures are added to speech without any visual articulators (double enhancement).

Though not explicitly the topic of the current research, it is also of interest whether all results found in the lab experiment of Drijvers and Özyürek (2017) – that are directly comparable to the present study – will be replicated in the present comparable online study. This will give insight into the effects of the similarities and differences between online and in-person testing on (degraded) speech comprehension research.

2. Method

2.1 Participants

In this study 43 native Dutch speakers between eighteen and thirty-five years old participated. Two of them were excluded because they reported to have autism, which might have influenced their audiovisual speech integration and lipreading abilities (Smith & Bennetto, 2007), their speech-and-gesture integration (Silverman et al., 2010) and audiovisual multisensory integration (Feldman et al., 2018). This could thus have influenced their performance in this experiment. One of the participants reported having ADD. Since research has shown that adults with ADHD tend towards a faster response pattern, which might indicate some impulsivity, and more importantly show different brain patterns of multisensory integration than people without

attention deficit (hyperactivity) disorder (McCracken et al., 2019), this participant was excluded from this study. Two other participants were also excluded, because they did not perform the task as was instructed (neglected to make a guess in two or more entire conditions – see 2.4 Procedure).

The remaining participants (n = 38), who had a mean age of 24 years and 9 months (SD = 4,24). No neurological, language related or behavioural problems were reported by the participants. None of them had any hearing impairments and they all reported to have normal or corrected-to-normal sight. The majority of the participants was female (5 male and 33 female) and highly educated. All participants were recruited via the Max Planck Institute subject database. Prior to starting the experiment, all participants gave their consent. After finishing the experiment, they received within-standards financial compensation for their participation. The present study was approved by the Ethics Committee of the Faculty of Social Sciences at Radboud University (ECSW-2020-049).

2.2 Stimuli

The stimuli used were a subset (in total 176 videos) of the videos used in Drijvers and Özyürek (2017). In each video there was a Dutch actress who uttered one Dutch high-frequency action verb; this actress remained the same during all videos. A JVC GY-HM100 camcorder was used to record the videos and each video lasted for approximately two seconds. The average speech onset after starting the video was 680 ms. The setting of the video remained consistent throughout: the actress was shown from head to knees while she was wearing clothes in a neutral colour in front of a grey background.

For the videos containing a gesture the actress was told to make a gesture which she thought that fitted the action verb while pronouncing that action verb. Besides the action verb, the actress did not receive any instructions about the gesture she should make, the gestures were thus created spontaneously in order to keep the gestures as ecologically valid as possible. The actress also did not receive any feedback on the gestures made. The starting position of the actress was the same as in the videos without gestures, namely with the arms besides the body. 120 milliseconds after the start of the video the preparation of the gestures started, and the execution of the stroke concurred with the uttered verb. After recording these videos, it was determined whether the gesture made in the video did indeed correspond to the verb that was intended by means of a pre-test (Drijvers & Özyürek, 2017). The pre-test participants were screened for visual, hearing, and neurological limitations similar to the participant criteria for their main experiment. In the pre-test the participants saw a video and had to fill in which verb they associated with this video. After that, they were presented the intended verb and had to indicate if the verb fitted the movement. The six videos that were excluded in this research because of the low score in how well the verb suited the movement, were also excluded in the present study.

The mouth blur that was present in the videos belonging to the no visible speech condition was created using Adobe Premier Pro. Only the lips are blurred, using a semi-opaque blur. The participants could thus see the colour of the skin underneath, but not the exact movements of the mouth.

The intensity of the sound in the video was set to 70 dB and Praat (Boersma & Weenink, 2015) was used to denoise the audio. For the conditions that contained clear speech this improved sound file was recombined with the video. For the conditions that contained degraded speech, noise-vocoding was applied to the sound of the videos. Noise-vocoding is an artificial distortion of speech where the fine structure of the signal is replaced with noise. More detailed this means that first the sound was filtered into a certain number of frequency bands, then the smoothed amplitude profile was extracted, next noise was modulated in each band and at last the frequency bands were recombined again (Sheldon et al., 2008). The higher the number of

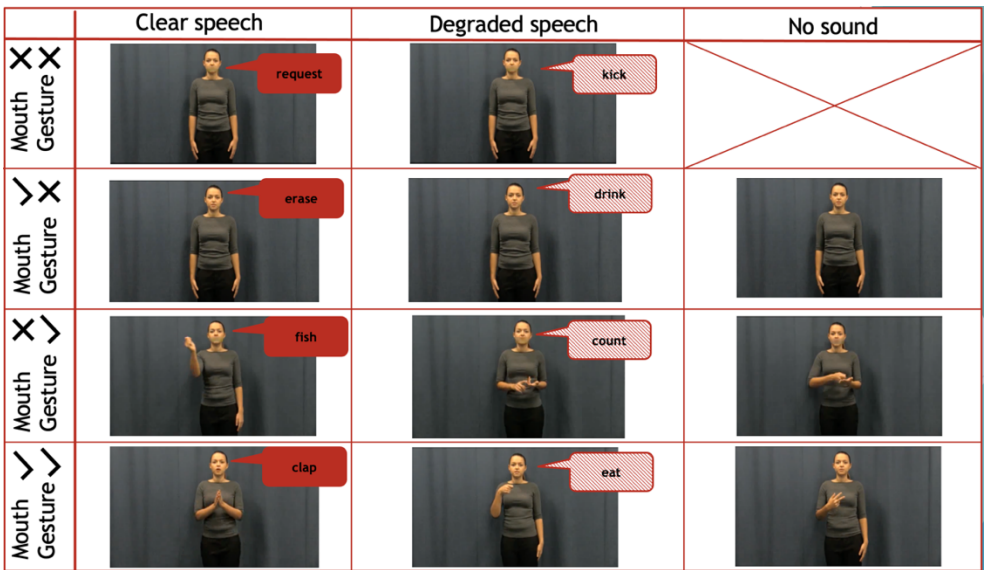
frequency bands in which the sound is filtered, the higher the intelligibility of the noise-vocoded speech. In this experiment 6-band noise-vocoding, moderate degradation, has been applied by using a custom-made script in Praat (Boersma & Weenink, 2015). 6-band noise-vocoding was chosen because at intermediate degradation, auditory and visual information is maximally integrated. This resulted in a more pronounced benefit of visual articulators in comparison to severely degraded or clear speech (Drijvers & Özyürek, 2017). For creating the noise-vocoded speech file, first of all the cut-off frequencies for this noise-vocoding had to be determined. In order to determine these frequencies, the sound file was band-pass filtered between 50 and 8000 Hz. Next, between these frequencies the signal was divided into logarithmically spaced frequency bands. The cut-off frequencies that resulted from this procedure were 50 Hz, 116.5 Hz, 271.4 Hz, 1473.6 Hz, 3433.5 Hz and 8000 Hz. White noise was then filtered by using these frequencies in order to obtain six noise bands. The amplitude envelope of each band was extracted using half-wave rectification. Next, in order to complete the noise-vocoded audio files for the degraded speech conditions, the amplitude envelope was multiplied with the noise bands and finally the bands were recombined again. For the no speech condition, all audio was removed from the video.

2.3 Design

To look into the effect of mouth coverage on speech and gesture comprehension in clear and degraded speech a repeated measures design has been used. In total there were eleven conditions, which resulted from combinations of three factors: speech quality, visible speech, and gestures. The speech could be clear, degraded or there could be no speech present at all. Furthermore, there could be gestures present or not and the mouth was either visible or blurred. The condition in which there would be no speech, no gestures and no mouth visible is left out, since participants then would have no information about the speech uttered. An overview of these conditions is presented in Figure 1. In total the participants saw 176 videos, 16 per condition. All the participants saw the exact same videos, but in a different order. All videos contained a different action verb. The participants had to perform a word identification task (further explanation in 2.4 Procedure). The percentage of correctly named verbs is the dependent variable.

Figure 1

Overview of conditions used in the experiment



2.4 Procedure

This experiment is conducted online by using Qualtrics (2005). Before the experiment started, the participants received information about the tasks in the experiment, were told that it took approximately 45 minutes and how much financial compensation they would receive for their participation. They were also told that they could do the experiment on a laptop or computer, but not on any mobile device and that they should use headphones during the experiment. The experiment should also be done in quiet surroundings. If then consent was given, they would automatically proceed to the experiment.

First participants were asked to answer a few general questions about gender, age, education and neurological, behavioural, language-related, vision or hearing problems. After that they watched a video (also scaled at 70 dB using Praat) in which a sentence was uttered. To check whether their sound worked properly the participants had to answer a question to which they only knew the answer if they watched and comprehended the video. The participants were also instructed to use the video to put their volume on a comfortable level. They were then told not to make adjustments to the volume during the experiment.

Next, they received further detailed instructions about the experiment. The participants were informed that not all videos will look the same, sometimes the speech will be clear or degraded or there might not be speech at all. They were also told that in some videos they can see the mouth and in others not and you sometimes see a hand gesture. The instruction also indicated that for each video the participants see, they have to type the verb (infinitive) they think is conveyed by the actress in the corresponding text box (word identification task). The participants were encouraged to fill in their best guess, but also told to fill in an 'X' when they were unable to make a guess. The participants were also requested to keep a distance of approximately 70 to 80 centimetres of their laptop or computer screen, sit straight in front of the screen (eye-height) and finish the experiment in one go within 60 minutes.

The 176 videos the participants saw, were randomized; each participant saw the videos in a different order. The participants could see only one video at a time, after which they had to click on a 'next'-button to proceed to the next video. The video was started when the participant clicked on the 'play'-button; each video could only be played once. After every 40 items the participants could take a break of a maximum of two minutes, then the screen automatically proceeded to the next video. The break could also be skipped by the participant if desired.

After having seen all videos, the participants were asked if they experienced problems with playing the videos or if they encountered other particularities. There were also a few questions about the duration of the experiment and whether the participants took any unplanned breaks.

2.5 Data-coding and analysis

For the data-coding the same rules were applied as stated in Drijvers and Özyürek (2017). This means that when participants wrote down the verb that was intended or made some minor spelling errors (e.g., 'aankruizen' instead of 'aankruisen') which did not alter the meaning of the verb, their answer was coded as correct. Their answer was also coded as correct if they wrote down the noun that is derived from the intended verb (e.g., 'basketbal' instead of 'basketballen'). Their answer was coded as incorrect if they gave a synonym (e.g., 'poetsen' instead of 'schoonmaken'), a category-related verb (e.g., 'schroeven' instead of 'boren') or a (supposed-to-be) compound verb in which the preposition was incorrect, added or deleted (e.g., 'verschuiven' instead of 'wegschuiven').

To analyse the data for accurate responses, first a repeated measures ANOVA was performed to determine the main effects of speech quality, gestures and visible speech and

possible interaction effects (see section 3.1). In this repeated measures analysis, the no sound condition was excluded because otherwise an unbalanced analysis would arise.

Thereafter, the enhancement effects were calculated of adding a certain visual articulator as a first or as a second visual articulator and of adding both visual articulators to speech. This is analysed as well in a repeated measures design to find out which visual articulator enhanced speech comprehension most (see section 3.2).

Finally, a repeated measures analysis was conducted with the available enhancement effects between the no speech and degraded speech conditions, to see in what way visual articulators enhanced comprehension when there was no audio available and if this differed from the enhancements in degraded speech (see section 3.3).

3. Results

Below the results of the analyses can be found. It is good to note that not always all types of speech quality were used in the analyses. While in section 3.1 accuracy results are analysed for clear and degraded speech, the analyses for all the possible enhancement effects were only conducted in degraded speech (section 3.2). Finally, the enhancement effects that could be found in the no speech condition are compared to their counterparts in degraded speech (section 3.3).

3.1 Analysis of the percentage correct answers regarding speech quality, visible speech and gesture presence in clear and degraded speech

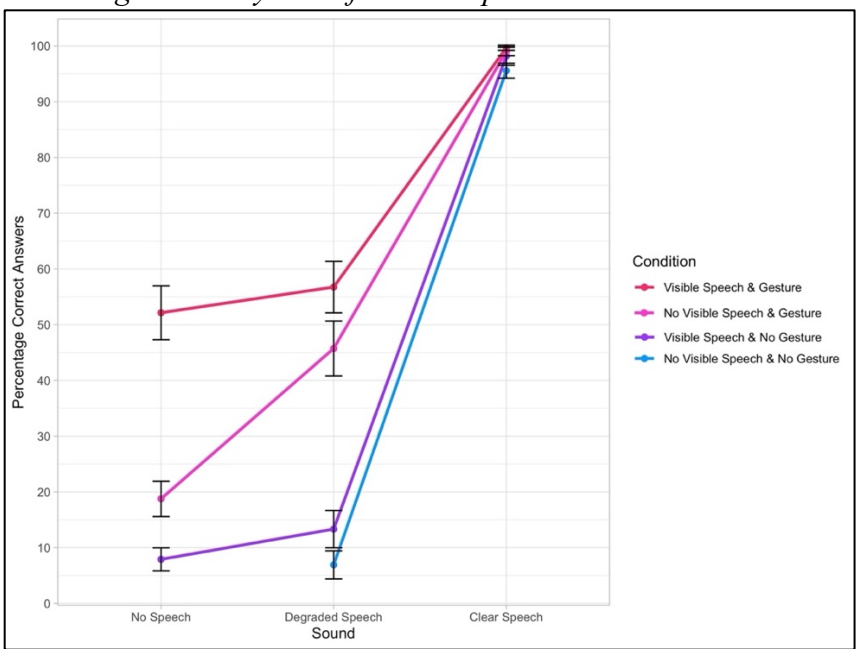
After collecting the data, the mean percentage of correctly given verbs per condition was calculated. The results are presented in Figure 2. The percentage of correct answers was highest in the conditions which contained clear speech with some minor differences depending on whether the mouth was visible (visible speech) or blurred (no visible speech) and if gestures were present or not. When both visual articulators, visible speech and gesture (red line in Figure 2), were present in the clear speech condition the percentage of correct answers almost reached ceiling (M = 99.67%, SD = 1.41). When only one visual articulator was present in the clear speech condition, so either visible speech without a gesture (purple line in Figure 2; M = 98.19%, SD = 5.01) or a gesture without visible speech (pink line in Figure 2; M = 99.01%, SD = 2.31), the averages were still almost at ceiling. When no visual articulator (blue line in Figure 2) was present in the clear speech condition, the average percentage correct was lowest of all clear speech conditions (M = 95.55%, SD = 4.09).

In the conditions with degraded speech the mean percentage correct was highest when both visual articulators were present (M = 56.74%, SD = 14.04). When only one visual articulator was present the average when only visible speech was available (M = 13.32%, SD = 10.19) was lower than when only gesture was available (M = 45.72%, SD = 14.97). The average percentage correct was lowest when no visual articulators were present (M = 6.91%, SD = 7.66).

In the conditions without sound, the mean percentage of correct answers was lowest. When both visual articulators were present the percentage correct was higher (M = 52.14%, SD = 14.70) than when there was only visible speech (M = 7.89%, SD = 6.28) or a gesture present (M = 18.75%, SD = 9.64). Note that there was no condition containing no speech + no visible speech + no gesture, since participants then would get no information (either visual or auditory) at all.

Figure 2

Percentage correctly identified verbs per condition



Note. The error bars represent the 95% CI-interval.

To look into the effect of each visual articulator and the speech quality on the mean percentage of correctly identified verbs a 2x2x2 repeated measures ANOVA was executed. The within-subjects factors were speech quality (clear speech versus degraded speech), visible speech (mouth visible versus mouth blurred) and the presence of a gesture (gesture versus no gesture) and the dependent variable was the percentage of correctly identified verbs. In this analysis the No Speech conditions were left out, since these would cause an unbalanced analysis because of the missing no visible speech & no gesture condition. This repeated measures ANOVA was conducted despite a violation of the assumption of normality of the residuals, because ANOVA is quite robust for violations of the normality when the groups are equal, which is the case since every condition has the same number of participants (n = 38) and video's (Field, 2013), and to maintain the ability to compare the results of the present study with the results from Drijvers and Özyürek (2017).

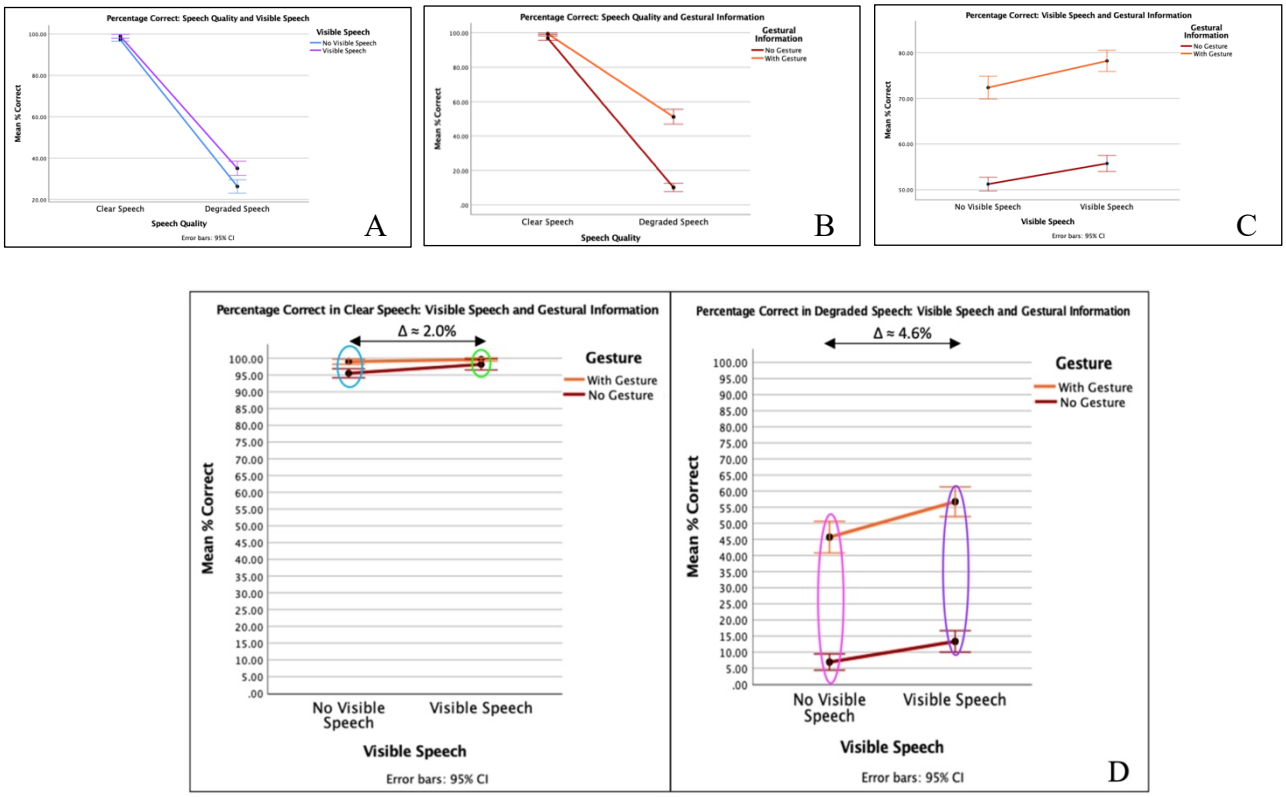
The results of the analyses show that, as hypothesized, there was a large main effect of speech quality (regardless of the presence of visible speech or gestures), $F(1,37) = 1973.412$, $p < 0.001$, $\eta^2 = 0.982$, which means that the percentage of correct answers was higher when the speech was clear then when this was degraded. Besides that there was a main effect of visible speech (thus regardless of speech quality or gesture), $F(1,37) = 2039.988$, $p < 0.001$, $\eta^2 = 0.560$. There was also a main effect of gesture (thus regardless of speech quality or visible speech), $F(1,37) = 36094.264$, $p < 0.001$, $\eta^2 = 0.939$. These indicate that when there was visible speech present the percentage of correct answers was higher than when there was no visible speech present. Also, the percentage of correct answers was higher when there was a gesture present than when there was no gesture present.

Next, the interaction effects were studied, these are also presented in Figure 3. As can be seen in Figure 3A, there was an interaction effect between speech quality and visible speech ($F(1,37) = 21.220$, $p < 0.001$, $\eta^2 = 0.364$). This means that the difference in the percentage correct answers between visible speech and no visible speech was larger in degraded speech than in clear speech. In Figure 3B, it can be seen that having a gesture present increased the percentage of correct answers much more compared to when it was absent in degraded speech than in clear speech, $F(1,37) = 409.812$, $p < 0.001$, $\eta^2 = 0.917$. The percentage of correct

answers was higher when there was a gesture present compared to when there was only visible speech present. However, there was no evidence that adding a gesture might provide even an additional benefit when it is accompanied by visible speech, since there was no significant interaction effect between visible speech and gesture ($F(1,37) = 1.180$, $p = 0.284$, $\eta^2 = 0.031$, as can be seen in Figure 3C. There was also a significant three-way interaction effect for speech quality, visible speech and gesture ($F(1,37) = 205.592$, $p < 0.001$, $\eta^2 = 0.156$ – also graphically presented in Figure 3D). This means that the difference between the mean percentages correct with gesture and no gesture in the visible speech condition (1.5% - green circle) and the no visible speech condition (3.5% - blue circle) was approximately 2.0% in the clear speech condition, which was significantly different from the difference of approximately 4.6% between the mean percentages correct of gesture and no gesture in the visible speech condition (43.4% - purple circle) and the no visible speech condition (38.8% - pink circle) in degraded speech. Figure 3D indicates that the percentage correct in clear speech is a bit higher for the no visible speech condition when there is a gesture present, while this does not seem true for the visible speech condition. The percentage correct in degraded speech is, independent of the visibility of the speech, always higher when there is a gesture present. The statistical significance of this three-way interaction effect indicates that while in both clear and degraded speech the percentage correct is higher when a gesture is present, in clear speech the benefit of having a gesture is less dependent of the presence of visible speech than in degraded speech.

Figure 3

Graphical representation of Interaction Effects



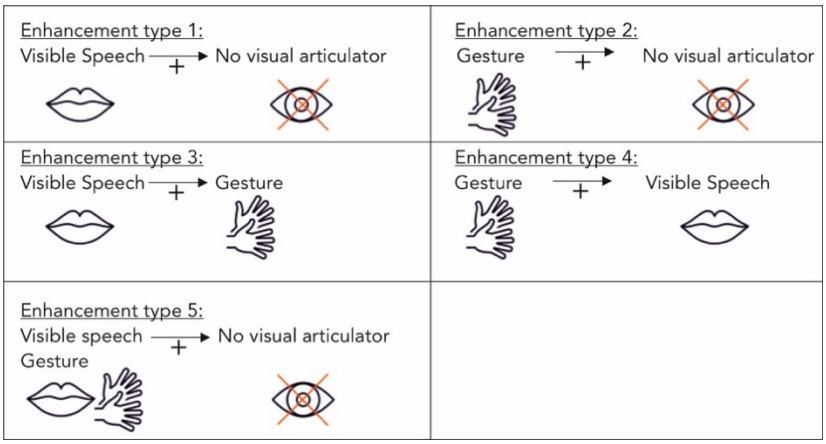
Note. A) Speech Quality & Visible Speech, B) Speech Quality & Gesture, C) Visible Speech & Gesture D) Speech Quality & Visible Speech & Gesture (Δ represents the difference between the two circled differences).

3.2 Analysis of enhancement effects in degraded speech

Next, it was studied how much adding a visual articulator would enhance the percentage of correct answers by looking at the enhancement effects. In total there are five types of enhancements for which an overview is presented in figure 4: 1) adding visible speech to speech without visual articulators, 2) adding a gesture to speech without visual articulators, 3) adding visible speech to speech which is already accompanied by a gesture, 4) adding a gesture to speech which is already accompanied by visible speech, and 5) adding visible speech and a gesture to speech without visual articulators (also named double enhancement).

Figure 4

Overview of enhancement types



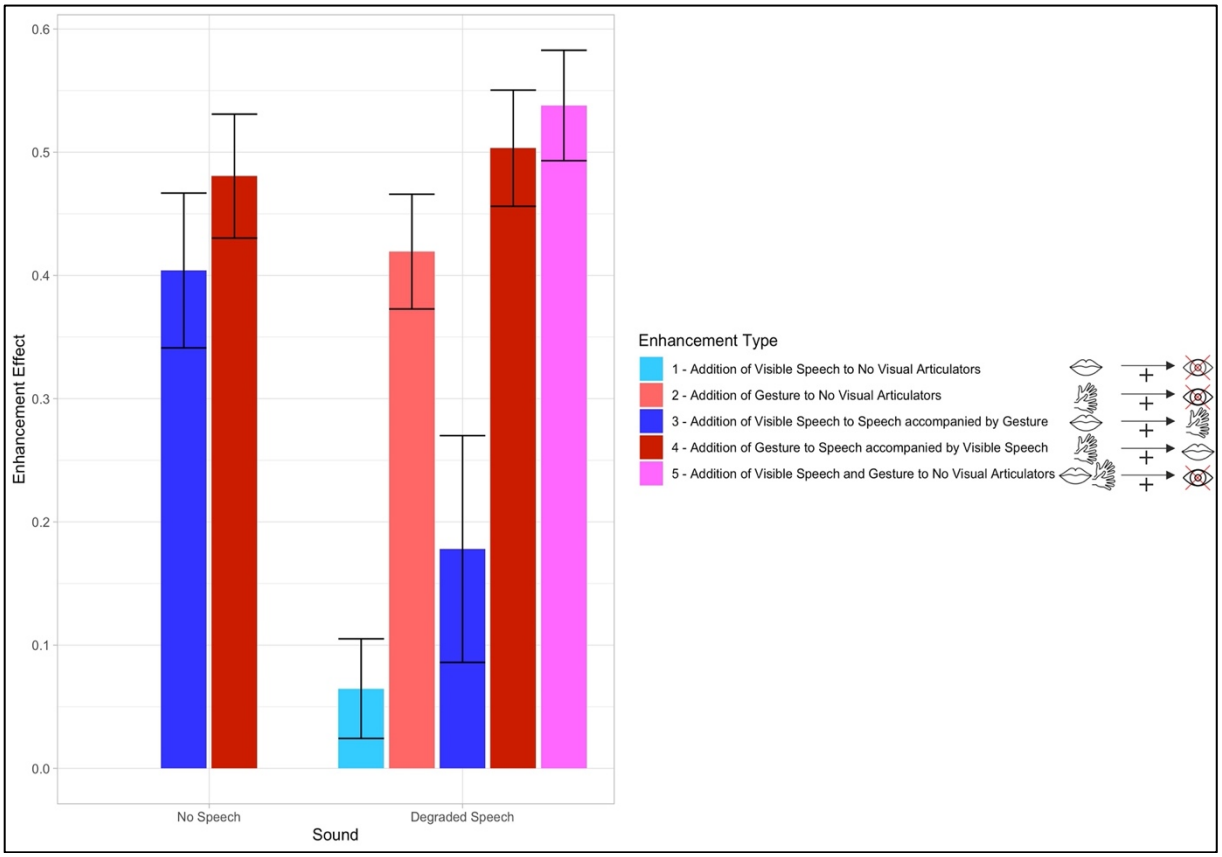
When for the enhancement effects two conditions are compared, possible floor effects must be taken into account since the more intelligible the speech is, the less it can possibly benefit from adding visual information (inverse relation – Grant & Walden, 1996). In order to ensure that the enhancement effects are comparable, they are looked at in a more relative manner. For the exact calculation of the enhancement effects the following formula was used: enhancement effect = condition A – condition B / 100 – condition B (Drijvers & Özyürek, 2017; Sumby & Pollack, 1954). The enhancement effect is then a measure of how much of the maximum gain in percentage is in fact realized. In this formula condition A is the condition after adding the visual articulator and condition B is the condition before adding the visual articulator. For example, when calculating the enhancement effect of enhancement type 5 – the double enhancement – the percentage of correct answers of the condition visible speech & gesture [A] will be subtracted from the percentage of correct answers of the condition no visible speech & no gesture [B] and divided by 100 minus the percentage of that last condition. The only problem is then that in the clear speech trials participants often have a perfect score within that trial, which causes a zero in the denominator. Following Schubotz et al. (2020) the clear speech conditions were therefore excluded of further analysis.

The enhancement effects were plotted against the enhancement type in Figure 5. In this figure can be seen that in the degraded speech condition the enhancement effect of enhancement type 1 (adding the visible speech to no visual articulators) was the smallest enhancement of all enhancement types (M = 0.06, SD = 0.12). Adding the visible speech to speech which is already accompanied by a gesture (enhancement type 3) also caused a very small enhancement effect (M = 0.18, SD = 0.28). The enhancement effects that resulted from adding a gesture to either speech without a visual articulator (enhancement type 2; M = 0.42, SD = 0.14) or to speech with visible speech (enhancement type 4; M = 0.42, SD = 0.14) were higher in comparison to

addition of visible speech. The biggest enhancement effect resulted from adding a gesture and visible speech to speech without any visual articulators (M = 0.54, SD = 0.14).

Figure 5

Enhancement effect per Enhancement Type in the No Speech and Degraded Speech condition



Note. The error bars represent the 95% CI-interval.

To check whether the enhancement type had a significant effect on the enhancement and whether the different enhancement types differed in their influence on the enhancement, a 1x5 Repeated Measures ANOVA was conducted[1] using enhancement type as within-subjects factor. Note that in this analysis again the no speech condition was left out, because of the otherwise unbalanced analysis. Mauchly's Test showed that the assumption of sphericity was violated for the main effect of Enhancement Type, χ^2 (9) = 182.050, $p < 0.001$. Therefore, a Greenhouse-Geisser adjustment was executed. There was a significant main effect of enhancement type, $F(1.673,61.911) = 68.631$, $p < 0.001$, $\eta^2 = 0.650$, which means that at least one enhancement type differed significantly from another enhancement type regarding the caused enhancement effect. A pairwise comparison (Bonferroni-corrected) was then executed to investigate which enhancement types differed. The results of this comparison are summarized in Table 1 (Appendix 1). In this table it can be seen that almost every enhancement type had a statistically different enhancement effect from another enhancement type. Since the enhancement effect of the addition of two visual articulators differed significantly from the enhancement of the addition of only visible speech or only gestures, it can be concluded that the addition of two visual articulators causes a larger enhancement than when only one visual articulator is added. The only comparison that was not significant was that of Type 1 versus Type 3, which means that there was no larger enhancement effect when visible speech was added to speech which was already accompanied by a gesture than when visible speech was added to speech without

[1] Repeated Measures ANOVA was executed despite violating the assumption of normality of the residuals, because of reasons as elaborated on in section 3.1.

visual articulators. These results support the hypotheses that the enhancement of adding a gesture is bigger than the enhancement of adding visible speech, either to speech without visual articulators or speech which is already accompanied by one visual articulator. Next to that, it also confirmed that adding both visual articulators to speech without visual articulator gave the highest enhancement effect.

3.3 Comparisons of Enhancement Effects in Degraded and No Speech condition

At last, a comparison was made between the enhancement effects in the degraded and no speech condition, to see how much information could be derived from the visual articulators alone (without speech). In order to make sure that the analysis remained balanced the factor enhancement type only contained the levels enhancement types 3 and 4. The enhancement effects for the no speech conditions were calculated the same way as described in section 3.2. The mean enhancement effect of adding visible speech to speech with a gesture in the degraded speech condition was lower ($M = 0.18$, $SD = 0.28$) than in the no speech condition ($M = 0.40$, $SD = 0.19$). The mean enhancement effect of adding a gesture to speech that is visible in the degraded speech condition ($M = 0.50$, $SD = 0.14$) did not appear to differ from the no speech condition ($M = 0.48$, $SD = 0.15$). The results of the 2x2 Repeated Measures ANOVA[1] showed that there was a main effect of sound, in the way that the mean enhancement effects were higher in the no speech than in the degraded speech condition ($F(1,37) = 8.527$, $p = 0.006$, $\eta^2 = 0.187$). There was also a main effect of enhancement type, in the way that adding a gesture to speech accompanied by visible speech (enhancement type 4) led to a bigger enhancement effect than when visible speech was added to speech accompanied by a gesture (enhancement type 3), $F(1,37) = 1.534$, $p < 0.001$, $\eta^2 = 0.689$. There was also an interaction effect between enhancement type and sound, which indicates that the difference between enhancement type 3 and 4 in the no speech condition differed of that in the degraded speech condition ($F(1,37) = 30.170$, $p < 0.001$, $\eta^2 = 0.449$). Post hoc paired t-tests showed that adding visible speech to speech that was already accompanied by a gesture gave a much lower enhancement in degraded speech than in the no speech condition ($t(37) = -4.338$, $p < 0.001$), while the addition of a gesture to speech that was already visible did not result in a different enhancement effect in degraded and no speech ($t(37) = 0.836$, $p = 0.409$).

4. Discussion

This research was aimed at investigating the role of co-speech iconic gestures and visible speech in speech comprehension in clear and degraded speech. More specifically, the goal was to compare the benefit of adding a gesture to speech without any visual articulators and the benefit of adding visible speech to speech without any visual articulators. Besides that, it was also of interest whether the benefit from one of the visual articulators differed from the benefit of having the two visual articulators in a joint context.

The present study showed that speech quality, mouth visibility and iconic gestures all influence speech comprehension in general. In the clear speech conditions the accuracy in this word identification task was significantly higher than in the degraded speech condition. The visibility of the mouth and presence of a gesture both have a positive effect on the accuracy as well, but more in the degraded speech than in the clear speech condition. Accuracy appeared to be highest when gestures and visible speech were both present as visual articulators in a joint context (Figure 2). Thereafter, accuracy appeared to be higher when gestures were present as the only visual articulator than when visible speech was the only visual articulator. Lowest accuracy appeared to be achieved when neither visual articulator was present. Noteworthy is that the benefit of having a gesture present is in general not dependent of the mouth visibility, although there seems to be a bit larger dependence in the degraded speech condition than in the

[1] Repeated Measures ANOVA was executed despite violating the assumption of normality of the residuals, because of reasons as elaborated on in section 3.1.

clear speech condition. This could possibly indicate that the more speech signal is lost, the more lips are needed to benefit from gesture, though these differences and effect sizes are small.

When studying the enhancement effects, the increase in percentage correct when adding one (or more) visual articulators shows that the addition of visible speech, either as a first or as a second visual articulator, has the smallest enhancement effect. The enhancement effect of adding visible speech to speech without any visual articulators (type 1) does not differ from the enhancement effect of adding visible speech to speech with iconic gestures (type 3). The enhancement effects of addition of gesture to speech without visual articulators (type 2) and to speech with visible speech (type 4) are both higher than the enhancement effects of addition of visible speech to speech without visual articulators (type 1) and to speech that is already accompanied by gestures (type 3). In other words, gesture alone contributes more than visible speech alone and the added contribution of gesture to visible speech is higher than the added contribution of visible speech to gesture. Gesture thus contributes more by itself and in joint context than visible speech does in degraded speech comprehension. However, the highest enhancement effect is reached when both visible speech and gestures are added to speech without visual articulators (type 5). While the enhancement caused by adding gesture to visible speech is comparable in the degraded and the no speech condition, the enhancement caused by adding visible speech to speech with co-speech gestures (type 3) is bigger in the no speech condition than in the speech condition. This could possibly indicate that a gesture has a consistent impact on speech comprehension, whilst the impact of visible speech varies more in the sense that people do not use visible speech in degraded speech as much as in the no speech condition. The more degraded the speech is, the harder it is to pair the phonological information from the visible speech. However, in the no speech condition people are forced to rely on visible speech more and are not hindered by degraded auditory information. This could also indicate that for people who are deaf, visible speech is of higher importance than for people with hearing impairments.

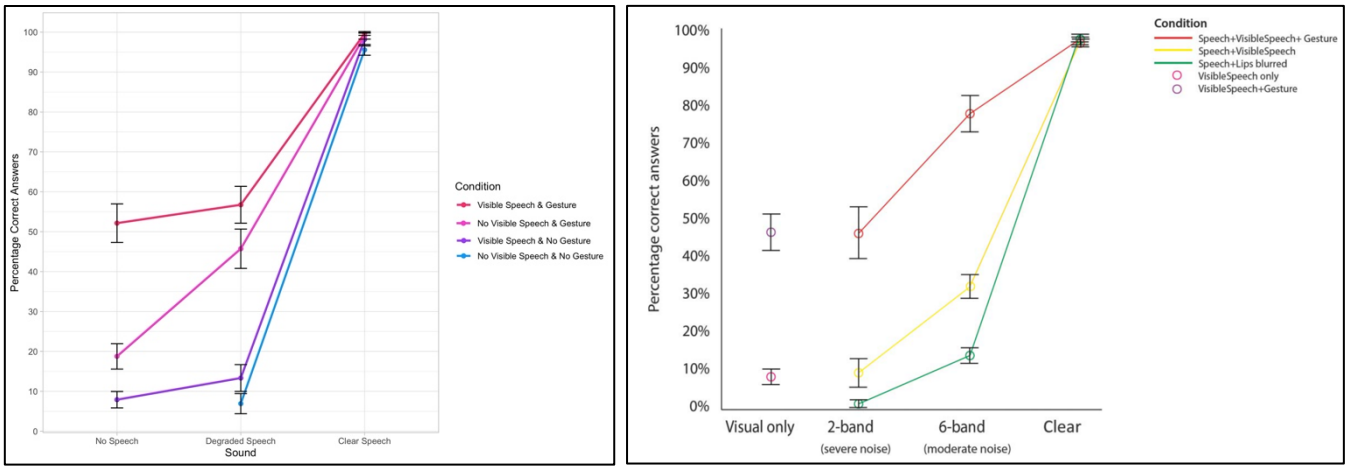
Most results were in line with the expectations that were made based on previous literature. That the addition of gesture to speech without visual articulators (enhancement type 2) leads to such a larger enhancement effect than that of the addition of visible speech to speech without visual articulator (enhancement type 1) shows that the difference found in Drijvers and Özyürek (2017) between their 'gestural enhancement' (enhancement type 4) and 'visible speech enhancement' (enhancement type 2) cannot be explained by a possible interaction between gesture and visible speech only. This research thus provides new evidence that the added value of gesture is higher than that of visible speech. This also fits the results from Zhang et al. (2012) who concluded that even when visible speech and gestures are used in a joined naturalistic context, the brain also uses gestures more than visible speech for speech comprehension. That the double enhancement leads to the most optimal enhancement for degraded speech comprehension does not come as a surprise, since the same result has been found in Drijvers & Özyürek (2017) and their individual counterparts have proven to be related to an enhancement in degraded speech comprehension as well (visible speech: e.g., Schwartz et al., 2004 – gesture: e.g., Holle et al., 2010). Even though this study used online testing as a method instead of in-person testing, this does not seem to have influenced the replication of the results in Drijvers and Özyürek (2017). Online testing thus appeared to be a suitable alternative to in-person testing for these kinds of robust effects.

What is also noteworthy to mention, though not explicitly researched in this study, is that while the effects found in Drijvers and Özyürek (2017) are replicated, the percentage of correct answers in the degraded speech are much lower in the present study compared to those found in Drijvers and Özyürek (2017). For example, in 6-band noise-vocoded speech in the condition visible speech & gesture the percentage correct in the present study is only ± 57%, while that was approximately 20% higher in Drijvers and Özyürek (2017). The percentages

correct in this study might resemble the results in 2-band degradation more than those in 6-band degradation (Figure 6). This might indicate that in online studies degraded speech sounds more degraded to participants than it would in a controlled lab environment.

Figure 6

Comparison of percentage correct answers in different conditions in present study and in Drijvers and Özyürek (2017)



Note. The present study (left) existed of an online word identification task, while Drijvers and Özyürek (2017) used in-person testing for their word identification task (right).

Since this study is conducted online because of the current COVID-19 pandemic, this brings some possible limitations for controlling external factors. For example, even though people were instructed not to adjust their volume after the test item, there is no way of checking whether they actually did this or not. Besides that, where in a controlled lab environment the volume can be set to 70 dB for every participant, the volume of the devices of the participants cannot be controlled that way. Participants also use different kinds of hear-/earphones, which might have led to different amounts of distortion to the sound. Sound quality might have therefore been a little different for every participant. Next to that, participants might have been disturbed during the experiment by other people or sudden background noise. Doing the experiment without supervision of a researcher or the external pressure of being in a lab might also have influenced the effort that the participants take for the experiment. In this case it might have influenced how often the participants did not even try to guess, which in extreme cases might have caused an underrepresentation of speech comprehension.

What might have possibly influenced the results are the individual items. All action verbs have only been used once in this study, to prevent possible repetition effects. Although all are considered to be high-frequency action verbs they might differ in how high-frequent they are. The action verbs for example vary from 'eten' (*to eat*) and 'fietsen' (*to cycle*) to 'faxen' (*to fax*) and 'parfumeren' (*to perfume*). This influence might be accounted for by performing a more complex data-analysis, which was outside the scope of the present study. When performing a more complex data-analysis the order of the videos as shown to the participants might also be added as a factor to account for possible order effects.

It would also be highly recommended to investigate the use of these two visual articulators in a more ecologically valid context. Research has shown that people tend to adjust the intensity of their speech, their visible speech (though not a strategy adapted by everyone), but also gesture kinematics in noise (Trujillo et al., 2020). This might indicate that in real life

the gestures and (in some cases) visible speech are more pronounced when the speech is hard to understand. This was not taken into account since speech degradation was only performed after the recording of the videos. When gestures and (for some instances) visible speech are more pronounced, this might lead to bigger enhancements of gesture and possibly visible speech. Thus, testing in a more ecologically valid context might influence the enhancement effects found in the present study.

For future research it might be of interest to look into other, not in the present study investigated, visual articulators as well, since there might be interactions between these and visual articulators and/or visible speech. This could mean that the enhancement effects found in this study are (possibly) influenced by other visual articulators. For example, the upper head has been related to a possible interaction between the effect of movements of the head and with that of visible speech (Davis & Kim, 2006; Munhall et al., 2004). Head movements might then take up the role of visual prosody, which can support extracting information from visible speech. The individual contribution of visible speech in the present study might possibly contain an extra enhancement because of this possible interaction between head movements and visible speech. This could possibly mean that the isolated enhancement due to visible speech might be a bit lower when the effect of interaction with other visual articulators is excluded. There is however little research on multiple visual articulators in a joint context apart from the literature discussed about visible speech and gestures. This would thus be interesting to get an even more complete picture on multimodal communication.

It will be interesting to investigate the results found in this study in a larger context in the future. Instead of a word recognition task a sentence recognition task might be used, since this provides context and enables prediction while processing speech. Zhang et al. (2020) have already shown that in clear speech there are some interactions between word surprisal and the benefit of visible speech and iconic gestures. However, in what way word prediction (and thus surprisal) influences the enhancement effects found in the present study in degraded speech remains to be investigated.

5. Conclusion

This study has shown that visible speech can enhance degraded speech comprehension and that iconic gestures can enhance degraded speech comprehension. The benefit from gestures alone is higher than that of visible speech alone. Similarly, speech comprehension is also more enhanced when gestures are added to visible speech than when visible speech is added to speech that is already accompanied by gestures. The enhancement effects of visible speech to speech without visual articulators or to speech with gestures do not differ. The highest enhancement is reached when both visible speech and gestures are added to speech without any visual articulators.

What the results of the current study mean for the current COVID-19 pandemic, is that maximum enhancement usually cannot be achieved since wearing a mouth mask makes the speech invisible and it makes the speech usually harder to hear and comprehend. Our results show that when lips are not visible and speech is degraded the comprehension is very low. However, this can be enhanced significantly with gesture – even when speech is degraded and lips are not visible. Therefore, it is of utmost importance that when wearing a mouth mask people use iconic gestures as support for speech comprehension whenever possible. Another interesting finding was that visible speech on top of gesture enhanced comprehension more in no speech condition than in degraded speech condition, showing that for deaf people seeing lips is more important and thus clear masks are advised.

References

Bernstein, L. E., Auer Jr, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*(1-4), 5-18. https://doi.org/10.1016/j.specom.2004.10.011

Blackburn, C. L., Kitterick, P. T., Jones, G., Sumner, C. J., & Stacey, P. C. (2019). Visual speech benefit in clear and degraded speech depends on the auditory intelligibility of the talker and the number of background talkers. *Trends in hearing*, *23*, 1-14. https://doi.org/10.1177/2331216519837866

Boersma, P., & Weenink, D. (2015). *Praat: Doing phonetics by computer* (Version 6.0.19). 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), 2213-2218.

Davis, C., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition*, *100*(3), B21-B31. https://doi.org/10.1016/j.cognition.2005.09.002

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212-222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Drijvers, L., & Özyürek, A. (2020). Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language and speech*, *63*(2), 209-220. https://doi.org/10.1177%2F0023830919831311

Feldman, J. I., Dunham, K., Cassidy, M., Wallace, M. T., Liu, Y., & Woynaroski, T. G. (2018). Audiovisual multisensory integration in individuals with autism spectrum disorder: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *95*, 220-234. https://dx.doi.org/10.1016%2Fj.neubiorev.2018.09.020

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4de ed.). SAGE Publications.

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

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. https://doi-org.ru.idm.oclc.org/10.1121/1.417950

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://www.mitpressjournals.org/doi/pdfplus/10.1162/jocn.2007.19.7.1175

Holle, H., Obleser, J., Rueschemeyer, S. A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *Neuroimage*, *49*(1), 875-884. https://doi.org/10.1016/j.neuroimage.2009.08.058

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260-267. https://doi.org/10.1177/0956797609357327

McCracken, H. S., Murphy, B. A., Glazebrook, C. M., Burkitt, J. J., Karellas, A. M., & Yielder, P. C. (2019). Audiovisual multisensory integration and evoked potentials in young adults with and without Attention-Deficit/Hyperactivity Disorder. *Frontiers in Human Neuroscience*, *13*, 1-11. https://doi.org/10.3389/fnhum.2019.00095

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

McNeill, D., & Levy, E. (1982). Conceptual representations in language activity and gesture. *Speech, Place, and Action*, 271-295.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological science*, *15*(2), 133-137. https://doi.org/10.1111%2Fj.0963-7214.2004.01502010.x

Obermeier, C., Dolk, T., & Gunter, T. C. (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex*, *48*(7), 857-870. https://doi.org/10.1016/j.cortex.2011.02.007

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*. http://dx.doi.org/10.1098/rstb.2013.0296

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

Qualtrics (2005). *Qualtrics* (June, 2021). https://www.qualtrics.com

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

Schubotz, L., Holler, J., Drijvers, L., & Özyürek, A. (2020). Aging and working memory modulate the ability to benefit from visible speech and iconic gestures during speech-in-noise comprehension. *Psychological Research*, 1-15. https://doi.org/10.1007/s00426-020-01363-8

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

Sheldon, S., Pichora-Fuller, M. K., & Schneider, B. A. (2008). Effect of age, presentation method, and learning on identification of noise-vocoded words. *The Journal of the Acoustical Society of America*, *123*(1), 476-488. https://doi.org/10.1121/1.2805676

Silverman, L. B., Bennetto, L., Campana, E., & Tanenhaus, M. K. (2010). Speech-and-gesture integration in high functioning autism. *Cognition*, *115*(3), 380-393. https://doi.org/10.1016/j.cognition.2010.01.002

Smith, E. G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology and Psychiatry*, *48*(8), 813-821. https://doi.org/10.1111/j.1469-7610.2007.01766.x

Stacey, J. E., Howard, C. J., Mitra, S., & Stacey, P. C. (2020). Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect. *Attention, Perception, & Psychophysics*, *82*, 3544-3557. https://doi.org/10.3758/s13414-020-02042-x

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of America*, *26*(2), 212-215. https://doi.org/10.1121/1.1907309

Trujillo, J., Özyürek, A., Holler, J., & Drijvers, L. (2020). Evidence for a Multimodal Lombard Effect: Speakers modulate not only speech but also gesture to overcome noise. Article in preparation.

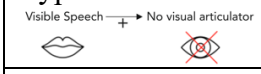
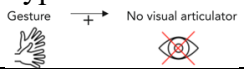
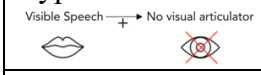
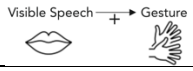
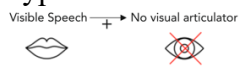
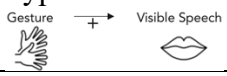
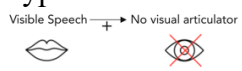
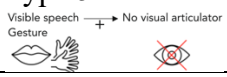
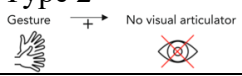
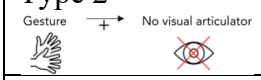
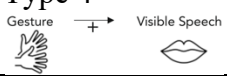
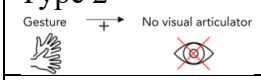
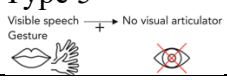
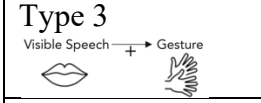
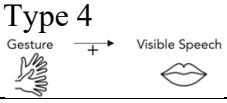
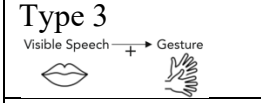
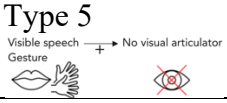
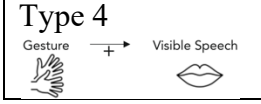
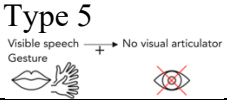
Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J. I., & Vigliocco, G. (2020). More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension. Article in preparation.

Appendix

Appendix 1 – Comparisons of the enhancement effects of different enhancement types

Table 1

Comparisons of the enhancement effects of different enhancement types

Compared Enhancement Types left Type versus right Type		Mean Difference	Significance	95% Confidence Interval
Type 1 	Type 2 	-0.355	< 0.001	[-0.442 – -0.267]
Type 1 	Type 3 	-0.113	0.208	[-0.253 – 0.027]
Type 1 	Type 4 	-0.439	< 0.001	[-0.537 – -0.341]
Type 1 	Type 5 	-0.473	< 0.001	[-0.551 – -0.396]
Type 2 	Type 3 	0.241	0.002	[0.067 – 0.416]
Type 2 	Type 4 	-0.084	0.004	[-0.148 – -0.020]
Type 2 	Type 5 	-0.119	< 0.001	[-0.180 – -0.057]
Type 3 	Type 4 	-0.325	< 0.001	[-0.452 – -0.198]
Type 3 	Type 5 	-0.360	< 0.001	[-0.482 – -0.238]
Type 4 	Type 5 	-0.035	0.007	[0.007 – 0.063]