Conveying a message through noise

A study on speech, gesture and multimodal productions in noisy environments

A thesis for the degree of

Master of Arts

by

Emma Berensen

Radboud University Nijmegen – faculty of Arts

Max Planck Institute for Psycholinguistics

August 2019

Supervisors:

James Trujillo & Dr. Linda Drijvers

Table of contents

i. Acknowledgements

I would not have been able to write this thesis without the members of the MLC lab of the Max Planck Institute. I would like to thank James Trujillo, Dr. Linda Drijvers, Prof. Dr. Asli Özyürek and Dr. Gerardo Ortega.

James and Linda – first of all thank you for letting me work on this great project. I have really enjoyed working this Lowlands project and data. I am very grateful for your patience, and for your advice and guidance through each stage of the process. James – thank you for your patience when I was struggling with the statistics and R, and for your advice, time and efforts to help me even when you were busy writing your PhD thesis. Linda – thank you for your time, advice and efforts, and for all your feedback, despite your busy schedule.
I have learned so much this past half year from both of you – you are great teachers, and I'm glad to have had you as my supervisors.

Asli – thank you for providing me with this fantastic opportunity. I wouldn't have been able to spend a whole year in the MLC lab, with my internship and then thesis, if it weren't for you.

Gerardo – thank you for introducing me to the MLC lab, and opening doors for me to start my internship there. If it weren't you I would not have been able to do an internship and subsequently write my thesis at the MPI.

ii. Abstract

In this paper, we investigate to what extent different noise levels influence the productions of speech and gestures, as well as what differences can be found in the different communicative attempts that were produced by the participants. An experiment was conducted in which directors were asked to convey twenty Dutch action verbs to the matcher in three noise conditions: either no noise, 4-talker babble noise or 8-talker babble noise. The results showed that the change in noise level had no significant influence on the production of gestures, nor on the production of speech. A possible explanation for this is that the noise levels changed too rapidly, and the director could not adjust their communicative strategy in time.

Comparing the first and second attempt, irrespective of noise level, the results showed that there was a significant change between the two communicative attempts only for the variables strokes, change in referent and change in hand: strokes were produced more often in the second attempt as compared to the first, the other two variables were produced less often. Noise level thus did not significantly influence the production of gestures, the number of attempts did. Directors thus produced more gestures in attempt 2, but less of these gestures are characterised by change of referent. This implies that the directors either produce other gesture feature changes or produce the exact same gestures. However, none of the other gesture feature changes are produced significantly more often in the second attempt than in the first. Coding all the attempts (instead of only the first two) in future research might give more results in the director's adjusted communicative behaviour over time.

1. Introduction

Communication is an important, everyday phenomenon. Often, communication that contains spoken utterances is accompanied by co-speech hand gestures (Kita, 2000; McNeill, 1992, 2005). These hand gestures can be important, if not crucial, to reach mutual understanding. It is known that gestures contribute to the communicative message, and that the listener attends to the information that is conveyed in gesture. For example, Beattie and Shovelton (1999) investigated the uptake of gestures by showing participants a cartoon narration in either an audio-visual or an audio-only condition, and found that the participants who could both hear the speech and see the gestures were more accurate in retelling the relative position and the size of objects. They concluded that iconic gestures relating to particular semantic features add to the linguistic message (Beattie & Shovelton, 1999, p. 27).

Cassell, McNeill & McCullough (1999) found that listeners attend to co-speech gestures not only when it makes a contribution to the message that is conveyed with speech, but even when the gesture contradicts the speech. They furthermore found that the listener integrates gestures together with the speech into a single linguistic representation, suggesting that speech and gesture are integrated systems. A similar result was found in Kelly, Özyürek & Maris (2010). In their first experiment, they presented participants with a short video clip of an action being performed, the prime, followed by matching or mismatching speech-gesture target pairs. Participants were quicker to correctly identify the target in matching speech-gesture conditions and produced fewer errors than in mismatching conditions. When comparing the weak and strong mismatching conditions, fewer errors were produced in weak mismatching conditions (i.e. speech: “chop”, gesture: cut) than strong mismatching ones (i.e. speech: “chop”, gesture: twist). In their second experiment, the stimuli consisted of only the speech-target conditions, and participants were asked to focus on the speech content in the task, thus not putting focus on gesture. Results showed that the participants still paid attention to the gestures, and that gesture and speech are integrated. Moreover, it has also been shown that gestures positively influence sentence memory by listeners (Feyereisen, 2006), that questions accompanied by gestures get faster responses (Holler, Kendrick & Levinson, 2018), and that they can help to disambiguate speech in case of ambiguous sentences (Holle & Gunter, 2007).

In a noisy environment especially, gestures seem to aid speech comprehension. Rogers (1978) concluded in his paper that the use of gestures can greatly improve speech comprehension, mostly so in lower signal-to-noise ratios, as gestures accounted for 60 to 65%

of the total visual improvement. Similarly, Drijvers & Özyürek (2017) found that iconic gestures enhance speech comprehension in noise, and that there exists a double enhancement, where both gestures and visual speech aid speech comprehension, which is strongest in a moderate noise condition. In an MEG study, Drijvers, Özyürek & Jensen (2018a) have studied gestural enhancement in clear and degraded speech. There was a bigger engagement of the hand-era of the motor cortex, the extended language network, the medial temporal lobe and occipital regions in degraded speech than in clear speech. This larger engagement was found in regions that are involved in the unification of information of different modalities, and in accessing lexical-semantic, phonological, morphological and syntactical information (Hagoort, 2013 in Drijvers et al., 2018a). These engagements of different regions can cause an increased uptake of gestures; the motor cortex might be engaged to extract semantic information from the gesture that help in speech comprehension is degraded speech. The visual areas are engaged to allow visual attention to gestures when the speech is degraded.

Another study focused on speech-gesture integration with matches and mismatches. It was found that the visual regions and the regions involved in unification were more engaged when a gesture mismatched in clear speech in comparison to when it matched. Engagement of the visual regions suggests that a mismatch allows for more visual attention. The engagement of the unification regions are reflective of the larger engagement required to resolve the mismatch between the auditory and visual information. Listeners also engage their motor system more strongly when a gesture mismatched the (clear) speech than when it matched as to 'simulate' the mismatching gesture to see if it fits the auditory signal. The regions were less engaged when the speech was degraded, which might occur because the degraded speech signal hinders integration of gestures with the speech (Drijvers, Özyürek & Jensen, 2018b).

Furthermore, studies have found that speakers tend to produce speech utterances in noise that differ from those in a clear environment. Among other modifications, speakers increase their vocal intensity, as well as the spectral tilt (Castellanos, Benedi & Casacuberta, 1996; Junqua, 1993). It has also been shown that speech produced in noise is more intelligible than that produced in silence (Pittman & Wiley, 2001; Van Summers et al., 1988). Speakers also tend to gesture more when they are not allowed to speak (Goldin-Meadow, McNeill & Singleton, 1996), or when trying to solve linguistic ambiguities (Holler & Beattie, 2003). Moreover, research has shown that gestures are as effective and sometimes even more effective than speech when conveying information about position or size (Holler, Shovelton & Beattie, 2009).

Speakers adjust to the communicative context: in a more communicative context, they tend to make their gestures larger, more complex and with a greater vertical amplitude than in a less communicative context (Trujillo, Simanova, Bekkering & Özyürek, 2018). Furthermore, it was found that speakers make their gestures bigger and more informative when communicating with a child as compared to an adult (Campisi & Özyürek, 2013). Finally, speakers combine several different gestures to describe a single event in order to re-create descriptions (Goldin-Meadow, McNeill & Singleton, 1996), thus changing the feature of the gesture.

So far, research has been conducted in order to gain insight in both the comprehension and the production side in noisy environments, as well as multimodal productions in a face-to-face setting and in noisy environments. However, studies concerning communication in noisy environments have often worked with pre-recorded video stimuli in which words were uttered in either clear or degraded speech. Questions remain on how multimodal productions are *created in* a noisy environment. Thus, this study aims to look at the multimodal communication from the speaker's side in a noisy face-to-face environment. It investigates which modality or modalities speakers produce when they are conveying a message in a moderately and highly noisy environment as compared to a clear surrounding. On top of that, it aims to study the tradeoff relation between gesture and speech. This paper aims to investigate the influence that noise level has on the production of speech utterances and co-speech hand gestures on the communicative strategy of the speakers. Furthermore, it aims to find out what changes are made by the producer in regards to speech, gesture, and gesture features when the communication between the producer and the listener seems to fail. In studying the produced strategies in communicating in noise, we could gain more insight in which way communication works most efficiently in a suboptimal environment. This could then be extended to communication with hearing-impaired individuals, which might influence audio-visual training.

In this thesis, first the characteristics of gestures that are relevant to this paper are discussed. It delves into the different types of gestures, the gesture phrases and their features. Subsequently, previous research concerning communication in a noisy environment as well as gesture and speech production during communication will be described. In that section, first the role of gesture on comprehension will be discussed, after which the focus will be on the production side during communication, of both speech and gesture. This is followed by communicative intent, communicative failures and then by two prevailing hypotheses

concerning the interaction between gesture and speech: the tradeoff hypothesis and the hand-in-hand hypothesis. In the subsequent section, information concerning the current study is given, and the research questions and the hypotheses will be introduced.

In the method chapter, a detailed description of the experiment set-up and procedure will be given, after which the results will be shown in the following chapter. In the discussion, the results will be interpreted, and the results will be linked to the discussed literature. Furthermore, the implications of the results of the experiment are treated, as well as its limitations. In addition, suggestions for future research will be given.

2. Literature review

2.1 Gestures

Gesturing can be done in silence by using emblems, which are hand movements that have a meaning of their own (for example an OK sign, or a thumbs up) (Obermeier, Dolk Gunter, 2012). Though more frequently, gestures occur in combination with speech (*ibid*.). It is these co-speech gestures that can enhance language processing and therefore have been the focus of many studies (for example Holler et al., 2014; Kelly, Özyürek & Maris, 2010; McNeill, Cassell & McCullough, 1994).

The focus of this paper is on the use of gesture and speech in communicative productions, and the aim is to discover which modality or modalities the producer uses to convey a message in a noisy environment. McNeill (2005) has argued that important characteristics of gestures are that they carry meaning, and that the gesture and the accompanying speech are co-expressive and simultaneous, but not redundant. In other words, he states that gesture and speech can convey the same message at the same time, but do so in their own way (*ibid*.).

So even though co-speech gestures are meaningful on their own, they do not replace speech. As McNeill (2005) showed in his paper, gestures, rather than replacing speech, do follow speech and vice versa: when speech diminishes, then so do gestures. When speech increases again, the gestures increase as well. Similarly, McNeill reports that when a speaker gets confused when telling a story, the gestures lose complexity, but gain it again when the speaker comes back to it.

2.1.1 Gesture types

Co-speech gestures can be divided into four gesture type groups, as proposed by McNeill (1992, 2005). These are iconic, metaphoric, deictic and beat.

1. Iconic gestures are gestures that depict a feature of concrete entities, events or actions. McNeill (2005) describes them as "*gestures in which the form of the gesture and / or its manner of execution embodies picturable aspects of semantic content (aspects of which are also present in speech)*" (p. 39). These gestures can refer to the form of an object, an action that is performed, the handling of an object or the trajectory an object covers. An example of an iconic gesture is given by Cassell, McNeill & McCullough

(1999): "*he climbed up the pipe*", which is accompanied with a hand gesture that goes upwards.

2. Metaphoric gestures are like iconic gestures in that they represent a concept. Metaphoric gestures however do not depict any concrete actions, but abstract ones, for example when a speaker is presenting an idea in his hand as if to hold a concrete object. Cassell et al. (1999) give an example of a speaker uttering the sentence "*the meeting went on and on*", which co-occurs with a rolling hand gesture (p.3). Metaphoric gesture present "*images of the abstract*" (McNeill, 2005, p. 39).
3. Deictic gestures are mostly thought of as pointing gestures, a hand with an extended index finger. Though deictic gestures are not only used to point at objects: they can also be used to locate something in the physical space in front of the speaker, compared to a reference point. Pointing gestures can thus be made at concrete objects, but also abstract ones. This abstract pointing is considered part of metaphoric gestures, as it spatializes locations for abstract concepts. An example given by McNeill (2005) is

   "when the speaker said, "they're supposed to be the good guys" and pointed to the central space; then said,"but she really did kill him" and pointed to the left space; next, "and he's a bad guy" and pointed again to the central space; and finally, "but he really didn't kill him" and pointed left. The difference between the central space (attributed morality) and the lefts pace (actual morality) became the speaker's metaphor, a temporary one, for the appearance / reality contrast" (McNeill, 2005, p. 40).

   The difference between concrete and abstract pointing is that the first creates new references, where the latter find references in it (McNeill, 2005; McNeill, Cassell & McCullough, 1994).
4. Beat gestures are small moving gestures, taking the form of a hand beating time (McNeill, 2005; p. 40). It has been observed that beat gestures tend to co-occur with the stressed syllable in multisyllabic words (McClave, 1994). Beat gestures can signal something the speaker thinks is important in the conversation (McNeill, 1992, 2005), and increasing the frequency with which beat gestures are produced enhance the salience of the information (Zappavigna et al., p. 229).

2.1.2 Gesture phrases

A gesture is made up of a series of phases which all have their own role in the gesture (McNeill, 2005). Kendon (1972, 1980) distinguishes gesture units, gesture phrases and gesture phases. A gesture unit starts when the limb leaves its resting position and ends when it moved back to a resting position (Kita, Van Gijn & Van der Hulst, 1997). This gesture unit can contain one or several gesture phrases. A gesture phrase is what people intuitively would call a gesture (McNeill, 2005). This gesture phrase then can contain several phases (without an "r"). A gesture phrase can consist of the phases called *preparation*, *pre-stroke hold*, *stroke*, *stroke hold*, *post-stroke hold* and *retraction*.

The *preparation phase* starts when the arms start moving from the resting position into the gesture space where the stroke can be produced. The start of this preparation phase also depicts the moment at which the visuospatial content of the stroke starts to unfold in the cognitive experience of the speaker (McNeill, 2005; Kita, Van Gijn & Van der Hulst, 1997).

A *pre-stroke hold* occurs when the movement of the limb stops temporarily before the stroke. If a speaker holds a gesture, it suggests that the speech and gesture are not aligned. The arm usually stays in this position until the speech utterance reaches the point which co-occurs with the gesture. Pre-stroke holds are thus a period in which the gesture waits for the speech so that cohesion can be established (Kita, 1990 (in Kita et al., 1997); McNeill, 2005).

The *stroke* is the heart of the gesture phrase: it is the phase with meaning and the only phase that is mandatory in a gesture phrase (McNeill, 2005). Kita et al. (1997) define a stroke as follows:

"A phase, in which more force is exerted than neighbouring phases, is a stroke. Note that acceleration (and deceleration) are good indicator of the exerted force, but sometimes a downward retraction has bigger acceleration than a stroke because of the gravity." (p. 8).

Stroke phases are crucial to a gesture phrase: without a stroke, a gesture is considered not to occur (McNeill, 2005). The majority of strokes (90%) of a gesture are synchronous with their accompanying speech. It is thought that, when a stroke and speech are not synchronous, the speech follows the stroke. The opposite, that strokes follow the speech, seldom happens (Kendon, 1972; McNeill, 2005; Nobe, 2000; Valbonesi et al., 2002 (in McNeill, 2005)).

A *stroke hold* is a stroke where the hands do not move. Stroke holds are "*strokes in the sense of meaning and effort but occur with motionless hands*" (McNeill, 2005, p.32). An

example would be when a speaker is depicting a specific form with his hands. Kita et al. (1997) differentiate between an *independent hold* and a *dependent hold*. The former refers to a stroke hold, the latter to a pre- or post-stroke hold.

A *post-stroke hold* occurs when the hands freeze in between the stroke and the retraction phase. This phase is optional, and can arise when a stroke phase has already ended, but the speech utterance is still ongoing. It was proposed that "*a post-stroke hold was a way to temporally extend a single movement stroke so that the stroke and the post stroke hold together will synchronize with the co-expressive portion of speech*" (Kita et al., 1997, p. 4, idea first put forward by McNeill, 1989).

The *retraction phrase* finally is when the hands go back to their resting position, which is not necessarily the same as the starting position. Kita, et al. (1997) also discuss a *partial retraction*: an interrupted retraction during which "*the hand makes a non-stroke movement toward a potential resting position, but before reaching the resting position shifts to a preparation of another stroke*" (Kita et al., 1994, p. 8; McNeill, 2005).

2.1.3 Gesture functions

Gestures can have several functions. For one, (iconic) gestures can specify the manner in which an action is performed. These gestures can hold information that has not been conveyed in the speech. An example of this is given by Cassell, McNeill & McCullough (1999, p. 4): when retelling a cartoon, a participant said "*he went back and forth*", but made a gesture with the index and middle fingers pointing down and wiggling as if the person was walking, indicating that the character was walking back and forth. Gestures can also be combined in order to re-create descriptions. This was shown by Goldin-Meadow, McNeill & Singleton (1996). In their study, participants were shown a video containing small dolls that moved and interacted with objects. They were asked to describe the video, and were either allowed or disallowed to speak. Results showed that the participants not only gestured much more when they were not allowed to speak, but also that several gestures were combined to describe an event. McNeill (2005) has described an example of this experiment:

"*For example, a scene in which a small doll is shown somersaulting through the air into an adjacent ashtray (the ashtray proportionately the size of a sandbox to the doll) was rendered thus: First, the subject used two hands to form a circle: the ashtray; next, she formed a small vertical space between the forefinger and thumb of her right hand: the doll; then, still holding*

this posture, her hand rose up, circled in mid-air, and dropped down into the space where the ashtray-gesture had been: the somersault arc landing in the ashtray" (p. 29).

The order in which this description was created was thus a stationary object first, followed by the moving object, and then the action. The three actions all contributed to the description of the action being performed by the doll. These can be seen as changes in referent: every gesture described a different referent or action than the previous one.

Iconic gestures may also specify the viewpoint from which the action is seen. Viewpoint is described as "*the locus of consciousness for model of the world*" (Parrill, 2009, p. 272). These gestures can be performed in an external representation in a third-person viewpoint, and an internal representation in a first-person viewpoint. These external gestures are also called *observer viewpoint gestures*, and the internal gestures are called *character viewpoint gestures*. (Cassell et al., 1999; McNeill, 2005; Parrill, 2009).

In addition, gestures serve to solve ambiguities. Holler & Beattie (2003) found that speakers, when producing sentences that contain homonyms, give disambiguating information in the gesture, but not in the speech. It seems that the speakers tend to rely only on the gesture to provide the information needed to disambiguate the sentence. Two examples given in Holler & Beattie's paper (p. 140) are given below.

Table 1: examples of cases in which gestures are used to disambiguate the sentence.

Speech	Gesture
1) 'first a ring came [into my mind]'	[thumb and index finger of the right hand slide up and down the middle finger of the left hand]
2) 'um…arms in […arms or]…weapons'	[right hand touches the right upper arm and the left hand the left one]

Other gesture functions as described by Cassell et al. (1999) are deictic gestures that locate characters in space and describe the spatial relations between them, beat gestures that signal that the linguistic information does not contribute to the advancement of the story, and

metaphoric gestures which can serve as an indication that a new segment in the narration is starting.

2.2 The role of gestures in comprehension

The reception and comprehension of listeners has been the subject of several studies in both clear and suboptimal environments. It has been argued in both behavioural and neuroscientific studies that iconic gestures have an impact on language comprehension (Beattie & Shovelton, 1999, 2002; Drijvers, 2019; Drijvers & Özyürek, 2017; Drijvers, Özyürek & Jensen, 2018a; Holler, Shovelton & Beattie, 2009; Holle & Gunter, 2007; Kelley, Healey, Özyürek & Holler, 2015; Kelly, Özyürek & Maris, 2010; Obermeier, Dolk & Gunter, 2012; Obermeier, Holle & Gunter, 2011).

2.2.1 The influence of gestures on comprehension

The role of gesture in comprehension has been the focus of extensive research. In the experiment carried out by Holler et al. (2014) a communication set-up between multiple people was created, with video clips of an actress uttering object-related messages being shown to two participants who could not see each other, causing the actress to alternate her eye-gaze between both participants. The authors manipulated the eye-gaze (direct or indirect) and the modality (speech-only or speech + gesture) of the video clips. Participants watched the video clips and were then asked to indicate via a button press which of the shown pictures corresponded to the message of the speaker. Results showed that the participants that were not directly addressed were significantly slower than their addressed counterparts. Importantly, it was also found that these unaddressed participants were faster in responding to the multimodal messages (speech + gesture) than the unimodal one. In other words, when the speaker was not addressed, the processing of speech was influenced by it, but not of gesture.

Holler, Kendrick & Levinson (2018) have studied the influence of bodily signals of comprehension. The authors invited participants in groups of two or three to converse freely, and analysed the question-response sequences. It was shown that questions that were accompanied by a gesture were responded to faster by around 200 ms than those that were not.

Gestures are taken into account also when they do not match with the linguistic message. Cassell, McNeill & McCullough (1999), have studied the influence of gestures on reception of linguistic and non-linguistic information, as well as its underlying representation. Recruited participants were divided into either a narrator group or a listener group. The narrators were shown a stimulus video showing an individual telling a story which contained alternatively speech-gesture combinations that either matched or mismatched in the categories of referent, viewpoint and manner mismatches. The participants in the narrator group were asked to watch a segment of the stimulus video, and then retell the story to the participants in the listener group. The authors argued that, if listeners would not pay attention to the gestures, they would not notice the speech-gesture mismatches, and the retellings of the participants would be the same as the one they saw in the stimulus video.

The results showed that all three types of mismatches resulted in inaccuracies in the retellings, causing the authors to suggest that listeners do pay attention to the semantic relationship between gesture and speech, and that the listeners still took gestural information into account when gestural information contradicted the information conveyed by the speech. Moreover, listeners take into account information that is conveyed only in gesture and try to combine contradicting information from gesture and speech.

The results from these studies showed that listeners do attend to the gestures not only in a natural environment, when the speech and gesture are aligned, but also when gesture conveys information that contradicts the accompanying speech.

2.2.2 Gesture-speech integration in noise

When situated in a suboptimal environment, participants take into account not only the gestural information, but still try to extract information from the auditory input. For example, Holle & Gunter (2007) investigated the role of iconic gestures in ambiguous speech sentences. Participants were shown videos of an actor who uttered a sentence whilst making gestures. Every sentence contained an unbalanced homonym early on in the sentence, which was then disambiguated later on in the sentence. Together with uttering the sentence, the speaker made an iconic gesture that depicted either the optimal meaning of the sentence or the suboptimal meaning. Measuring the time-locked event-related potentials, the authors found that there was a smaller N400 after a congruent gesture in comparison to an incongruent one. Furthermore, the participants showed a bigger N400 with the suboptimal target words when

the dominant gesture was produced and smaller when the matching suboptimal gesture was produced.

In their paper, Drijvers & Özyürek (2017) have conducted an experiment investigating the influence of gestures on top of visible speech (i.e. information from lip movements, tongue movements and teeth) in a noisy environment. The stimulus materials consisted of short video clips of an individual who uttered a Dutch action verb. The authors manipulated the noise level in the videos (in addition to clear speech, a highly noisy condition and moderately noisy condition were added), as well as the audio-visual information from the video. In other words, the video could consist of speech + lips blurred, speech + visible speech, and speech + visible speech + gesture for every noise level. On top of that, two conditions without sound were added: visible speech only, and visible speech + gesture. During the experiment, participants were presented with these short video clips, and were then asked to type what verb they thought was being conveyed. The results showed that participants benefit most from both visual speech and gesture when perceiving a message in a noisy environment. This double enhancement is optimal at the moderate noise condition, where "*there is an optimal range for maximal multimodal integration where listeners can benefit most from the visual information*" (p. 219). The authors argue that at this noise level the auditory cues were still distinguishable, and that this, together with the information gained form visible speech and iconic gestures results in an "*additive effect of double, multimodal enhancement from visible speech and auditory cues*" (p. 219). Such an additive effect was not found in the highly noisy condition, which suggests that, in severe noise, visible speech is not deemed reliable enough to be matched to the phonological information in the degraded speech signal.

A similar result was found in Holle et al. (2010), also found a pattern of inverse effectiveness: there was a greater neural enhancement for bimodal stimulation in moderate noise than in in clear speech.

Drijvers, Jensen & Özyürek (2018a) have also studied the gestural enhancement in degraded speech comprehension. They presented participants with videos of an actress who uttered an action verb that either was or wasn't accompanied by a gesture. These videos were shown in clear speech or in moderate noise. After each video they saw, the participants were presented with four verbs, of which they had to identify the correct one. These four verbs consisted of the correct verb, a phonological competitor, a semantic competitor and a verb that was unrelated. The results showed that gestural enhancement is largest in degraded

speech (in comparison to clear speech): when speech was degraded and a gesture was present, listeners had a shorter reaction time. The authors also found engagement of the hand-era of the motor cortex, the extended language network, medial temporal lobe and occipital regions; the regions that are associated with gestural enhancement of degraded speech, and simulation of gestures, as well as an increased visual attention to the gestures.

Obermeier, Dolk & Gunter (2012) studied the uptake of gestures in disambiguating speech in noise. They showed participants videos of an individual uttering a sentence in multi-babble noise or not. The sentence contained a homonym, that was disambiguated with a gesture. Later in the sentence a target word was uttered that either referred to the dominant meaning or subordinate meaning of the homonym. Results showed that in noise, gestures were taken into account as a communicative cue and gesture processing was enhanced, but not in the noise-free videos.

2.2.3 Native vs non-native listeners

Drijvers & Özyürek (2018) studied the integration of iconic gestures with speech in clear and noisy environments. Native and non-native speakers of Dutch were exposed to videos of an actress uttering a verb that was accompanied by an iconic co-speech gesture, which could either match or mismatch the speech signal. During the experiment, the EEG of the participants was measured. While both groups showed similar behavioural results – clear speech and gesture-speech matches led to a higher identification rate - EEG results showed that speech is integrated with gestures differently for native listeners than for non-native listeners. Native listeners showed a N400 that was more negative when speech and gesture mismatched than when they matched, More negative N400 amplitudes were also found when the speech was degraded in comparison to when it was clear.

Non-native listeners also showed more negative N400 amplitude for speech-gesture mismatches than for matches when presented in clear speech. A similar pattern was found in degraded speech as compared to clear speech, which suggests that the integration of gesture with speech required more neural resources when the speech was degraded. When comparing the gesture-speech matches and mismatches in degraded speech however, no difference in N400 amplitudes was found between them. Both of these amplitudes did not differ from the amplitude found after the mismatching gesture in clear speech. The authors suggest that non-native listeners cannot fully make use of the semantic cues of gestures when the auditory signal is too difficult to resolve; it could be that more neural resources were required to

resolve the degraded auditory information, which may have caused that the non-natives did not benefit from visual information for comprehension. They stated that non-native listeners were more hindered in coupling the semantic information conveyed by gesture to degraded auditory cues than natives, "*possibly because they need more auditory cues to facilitate access to gestural information*" (p.17).

2.3 Gesture production

Speakers tend to adjust their gestures to their listeners and the communicative environment. In other words, speakers tend to convey important information that is not included in the speech in their gestures. This is known as the cross-modal compensation hypothesis: speakers identify a referent with the use of gestures in particular when the speech does not uniquely specify the referent. Speakers thus use gestures in order to compensate for the lack of specification in their speech (Cohen, 1977; De Ruiter, 2006; Kendon, 1983; So, Kita, & Goldin-Meadow, 2009).

An example of speakers adjusting to their audience is found in the study of Peeters et al. (2015), who have found that speakers prolong the strokes and post-stroke holds of their gestures so that these can be more communicative, more so when the gesture carried most of the communicative load. Campisi & Özyürek (2013) found that, when participants were asked to demonstrate an action to a child, a novice adult and an expert adult, the gestures aimed at the child were more informative and bigger. Plus, the participants gestured more often when gesturing to the child. The authors suggest that the speakers adjust the way in which they convey the message according to the presumed state of knowledge of the listener.

Holler & Beattie (2003) aimed to study if speakers use gestures in order to resolve verbal ambiguity. They were asked to read sentences containing ambiguous words and then to explain these to the experimenter. They found that all the participants used representational gestures despite providing disambiguating information in the speech too, and that 90% of the speakers used iconic gestures. In 46.4% of the sentences they solved the ambiguity by using gesture, and if the word was disambiguated with gesture alone (so not together with speech), then the gesture was more elaborate than when it was accompanied by speech. This suggests that "*the form of the gestures employed in association with the resolution of verbal ambiguity depends on how suitable the speaker perceives speech to fulfil the communicational task at hand, and thus that gesture must be directly linked to the speaker's communicative intent*" (p. 143). When participants told a story with homonyms in it, they used significantly more

gestures with these homonyms than with control items. These results imply that speakers modulate gesture and speech according to how effective the speakers think they are in the communicative context. The 'context' thus does not only hold the current narrative, but also the communicative needs as seen by the speaker.

So, Kita & Goldin-Meadow (2009) have investigated how speakers do coordinate speech and gesture to disambiguate important information. Focusing on investigating whether speakers produce gestures in referent identification when speech fails to do so, So et al. let participants watch short stimulus videos, and then let them describe what happened in these videos. In the videos the lexical specificity was manipulated in the genders of the protagonists: it was either a Man-Man story (M-M) or Man-Woman story (M-W). The authors assumed that speakers are less likely to uniquely specify the referents in the story with speech in the former condition than in the latter. The results surprisingly showed that the participants used gestures to specify referents less often when speech failed to be specific as well. In other words, participants used gestures to refer to the protagonists, but only gestured to specify a referent when that referent was also referred to in speech. The gestures thus did not compensate for an under-specification in speech, but paralleled with it.

These studies show that speakers tend to adjust their gesture productions to the needs of the communicative partner and the communicative task.

2.4 Speech production and perception

In communication in a noisy environment, speech is also affected by the background noise. Research has shown that speakers increase their vocal amplitude when their surroundings contain noise. This is known as the Lombard effect (Lombard, 1911), and was originally thought of as an automatic regulation of the intensity of the voice as a result of auditory feedback. The Lombard effect not only holds that noise causes the vocal amplitude to rise, but also that the vocal amplitude changes, which includes "*a rise in fundamental frequency, a flattening of the spectral slope (or "tilt"), and an elongation of signal duration*" (Zollinger & Brumm, 2010, p. 1). Studies that have focused on speech that is produced in a noisy environment have shown that this speech has not only an increase in intensity, the perceived loudness of the sound, and amplitude, the size of oscillations of the vocal folds, but that it is also defined by a decrease in speech rate, phoneme modifications, a shift in spectrum that goes more towards the medium frequencies, and change in pitch (Castellanos, Benedi & Casacuberta, 1996; Davis, Kim, Grauwinkel & Mixdorff, 2006; Elman, 1981; Garber, Siegel

& Pick, 1981; Garnier, 2008; Junqua, 1993; Kim, 2005; Stanton, Jamieson & Allen, 1988; Van Summers, Pisoni, Bernacki, Pedlow & Stokes, 1988).

Studies have also shown that Lombard speech is different and more intelligible for the listeners than speech in a clear environment (Dreher & O'Neill, 1958; Pittman & Wiley, 2001; Tufts & Frank, 2003;Van Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988). Pittman & Wiley examined the speech produced in a clear environment, in wide band noise and in multi-talker babble noise. The results showed that the speakers' vocal levels had increased by 14.5 dB on average in both the wide band noise and the multi-talker babble noise conditions as compared to the quiet condition. Furthermore, on average the speakers' words lasted 77 ms longer in both of the noise conditions in relation to the no-noise condition. The productions in the noise levels were also characterised by an increase in F0, and a decrease in spectral tilt. Furthermore, in their second experiment, they focused on the recognition of speech that is produced in a clear environment and in noise, creating two conditions. In one condition, the differences in vocal levels were preserved, in the other one the signal-to-noise ratios were equated. In the equated condition, the speech produced in both the wide band noise and the multi-talker babble noise was recognised 15% more often than the speech produced in the no-noise environment. In the preserved condition, the recognition of the speech was on average 69% higher in both noisy environments as compared to the quiet environment. The results suggest that the recognition of speech utterances was better for speech that was produced in noise than for speech produced in clear a environment.

These results are in line with those found in the study of Van Summers, Pisoni, Bernacki, & Stokes (1988). In that study, participants were asked to read aloud words shown on a screen. The participants performed the task either in a silent environment or were exposed to several degrees of noise. The results showed that an increase in noise level led to an increase in amplitude, an increase in word duration and fundamental frequency, and a decrease in spectral tilt. When all stimuli were equated and presented at equal SNR ratios, the authors found that digits that were produced in a noisy environment had a higher identification rate than those produced in silence. There are seemingly characteristics of speech produced in noise that make it more intelligible. Van Summers et al. state the following concerning these results:

"*In trying to articulate speech more precisely under these adverse conditions, the talker introduces certain changes in the acoustic–phonetic correlates of speech that are similar to those distinguishing stressed utterances from unstressed utterances. The changes in the*

prosodic properties of speech which occur in noise are also similar to changes that occur when subjects are explicitly instructed to "speak clearly". However, the F1 and F2 data suggest that the changes in productions that subjects automatically make when speaking in noise are not identical to the changes that occur when subjects are given clear speech instructions or when subjects put stress or emphasis on particular utterances" (p. 15).

Garnier, Henrich & Dubois (2010) compared the modification of speech perception and production with self-monitoring feedback with different noise types, and also compared the moderation of acoustic and lip articulatory parameters in interaction. They argued that the speech adaption made by speakers did not only consist of acoustical and articulatory moderations, but also of prosodic moderations that can serve to maintain intelligibility for the speech partner. This suggests that the Lombard effect is not only an automatic regulation of the voice, but also a communicative adaption.

Adaption can also be found in the speech rate. It has been shown that speakers use more speech, produce more content in their speech and also that they include more details in their speech if there is a gap in common ground between speaker and listener (Campisi & Özyürek, 2013; Isaac & Clark, 1987). This result thus opposes the findings of studies concerning speech in noisy environments: it was found that speaker decrease their speech rate when communicating in noise.

2.5 Communicative strategies

2.5.1 Communicative intent

Trujillo, Simanova, Bekkering & Özyürek (2018) have studied the communicative actions and gestures in the context of production and comprehension. They state that, for communication in general, there are two requirements: the speaker must make the communicative intention recognisable for the listener, and they must represent the semantic information that they want the listener to observe (p. 38-39). In their first experiment, they asked participants to perform sets of everyday actions using objects (for example *pour the water into the glass*), in either a more communicative or less communicative context. In the more communicative context, the participants were told a confederate would watch them through a camera placed directly in front of them to study their gestures. In the less communicative context, they were told the confederate would watch them through the camera to learn about the set-up of the experiment. Furthermore, they were split into an action group and a gesture group. The former was asked

to perform the action using the presented objects, the latter to gesture the action, i.e. to perform the action as if using the objects but without touching them. The results showed that both of the modalities were regulated in size, number of submovements and maximum amplitude: in a more communicative context, gestures were made larger, had greater vertical amplitude and had a more complex movement in comparison to the less communicative context. On top of that, in the more communicative context, both modalities contained more addressee-directed eye-gaze. In the second experiment, participants were shown videos containing the same stimuli as in experiment 1, and were asked to judge whether an action was performed for the speaker self or for the listener, thus being communicative or non-communicative. It was found that not so much the kinematics but the addressee-directed eye-gaze were considered cues for communicative intent. In a third experiment, which focused on the kinematics alone without the addressee-directed eye-gaze, the faces of the actors in the videos were blocked. This resulted in a marginal increase in recognition in a more-communicative context than in a less-communicative. In the gesture modality, a strong relation was observed between the increased maximum amplitude and a higher recognition rate, suggesting that the participants interpreted kinematics more easily as more communicative. The authors propose that eye-gaze serves to initiate interaction, while kinematics enhance the legibility of the movement.

In a follow-up study, Trujillo, Simanova, Bekkering & Özyürek (2019) aimed to investigate if and how the kinematic modulation influences gestural comprehension. The stimuli were the same as the previous study, but with the actor's face blurred in half of the videos. The participants were asked to watch the video and indicate which action they thought was depicted, with two answers they could choose from. The authors found a higher recognition rate for pantomime gestures and initial fragments in the more communicative compared to the less communicative context. The visibility of the actor's face did not significantly influence the results, which causes the authors to suggest that "*the improved comprehension may come from fine-grained kinematic cues, such as hand-shape and finger kinematics*" (p. 7). To eliminate the influence of face and finger kinematics, in the second experiment, the stimuli were reduced to a visually simplified stick-figures. In this experiment, too, there was a higher recognition rate in the more communicative than the less communicative context overall, as well as for medium fragments. Actions produced in more communicative contexts were thus more easily understood early on, and kinematic modulation causes better recognition even if the visuals are reduced.

2.5.2 Communicative failures

In a communicative context, the communication is not always successful. It is known that, when a new referent is successfully introduced in the description, afterwards reduced references can be applied; this signals an increase in common ground between speaker and listener (Clark & Wilkes-Gibbs, 1986; Holler & Stevens, 2007; Hoetjes Krahmer & Swerts, 2015). Holler & Wilkin (2011) found that, when the speaker receives negative feedback, they use slightly more gestures after the feedback, though this was not significant. Hoetjes, Krahmer & Swerts (2015) have followed up on this and studied the gesture rate and form in unsuccessful communicative situations. In an experiment, participants had to refer to complicated figures that were hard to describe. They communicated with a confederate, who gave either positive or negative feedback. The results show that the negative feedback caused the linguistic references to be shorter and to contain fewer words. The speech rate was also found to be lower. After each production following the negative feedback, the gesture rate had increased, and the number of repeated gestures also increased slightly, which kept increasing for every production after the feedback. These results suggest that speakers tend to rely more on gesture when communication turns out to be unsuccessful, and that the produced gestures after negative feedback was more effortful. In the current paper, we will call every communicative production an attempt.

2.6 The tradeoff vs. hand-in-hand hypotheses

An influential theory regarding the production side of the relationship between gesture and speech is the tradeoff hypothesis (Bangerter, 2004; De Ruiter, 2006; Melinger & Levelt, 2004; Van der Sluis & Krahmer, 2007). This theory holds that there exists a tradeoff relation between gesture and speech when it comes to communicative load. In other words, according to the tradeoff hypothesis, if it becomes more difficult to convey a message through speech (when the speech requires more effort), it becomes more likely that gestures occur, which instead of the speech convey the message. Also, when it becomes harder to make gestures, then speakers will rely more on speech. Several studies have been conducted that support this hypothesis. For example, Graham & Heywood (1975) studied the effect of gesture prohibition on the speech production by asking participants to describe two-dimensional figures and either allowing or prohibiting them to produce gestures. When participants were not allowed to gesture, they produced a higher amount of words that were used to describe spatial relations. They also used less deictic expressions than when they were allowed to gesture.

Graham & Heywood's findings suggest that speech does take over the information often conveyed by gesture (i.e. spatial relations) when gesture is prohibited.

Melinger & Levelt (2004) asked participants to describe the space and colour of several circles to a listener in a picture description task. It was found that, when participants used iconic gestures to represent the spatial relations of the circles, they omitted more spatial relations from speech than participants who did not produce gestures.

Bangerter (2004) used a matching task procedure in which the speaker (or director, as called here) and listener (or matcher) were sitting next to each other, and the director had to describe pictures of people to the matcher at varying distances ranging from 0 cm (arm length) to 100 cm. He found that not only deictic gestures decreased when the distance to the target object increased, but also that pairs that were visible to one another used fewer words when targets got closer. Pointing thus reduced verbal effort.

So, Kita & Goldin-Meadow (2009) however, have studied whether speakers use gesture and speech in order to help them specify referents when they cannot do so in speech, and how speakers semantically coordinate gesture and speech in order to disambiguate information that is needed for discourse processing. They suggested that the gestures that speakers make tend to follow the speech, rather than compensate it; in their study they found that 35% of the produced gestures were linked to locations associated with a character, thus used to specify the identity of a referent. The speakers did not produce gestures when the referent was not referred to in the previous speech. The authors suggest that specificity in speech concerning referents goes hand in hand with specification of those referents in the gesture. De Ruiter, Bangerter & Dings (2012) have named this the hand-in-hand hypothesis: gestures follow, or go hand in hand with the speech.

De Ruiter et al. have aimed to investigate these two opposing theories. They used the matching task procedure of Bangerter (2004) to study the relationship between speech and gesture in collaborative referring to something in the shared visual environment. They asked the producers, or directors, to identify targets (tangram figures; little figures consisting of several wooden shapes) to listeners, or matchers, from a set of targets that were visible to both of them. The authors manipulated the codability of the tangrams (i.e. simple tangrams, humanoid tangrams -for example *ice dancer* - and complex abstract tangrams) and the repetition of reference, i.e. whether the target is old or new. Of the results, there has been only one that supports the tradeoff hypothesis, which is that the deictic gesture rate decreased when the directors repeated an expression with referents. The authors argue that this result

emphasises the role of conceptual pacts in order to facilitate conversational referring. However, the authors also found that the iconic gesture rate was not systematically affected at all, and the manipulations which made it harder to speak were found to have a strong effect on speech, but not on any of the gesture types. These results do not support the tradeoff hypothesis. The found results that show that the rate of deictic gestures was positively correlated with the amount of locative expressions in speech support the hand-in-hand hypothesis.

It should be noted that, in De Ruiter et al.'s study, the manipulation consisted of the difficulty of the tangrams (the codability) and the repetition of the figures. These are arguably not the most impacting variables to manipulate in order to study the tradeoff hypothesis. That is to say, with these manipulations the production of speech utterances or gestures is not necessarily complicated; in both codability and repetition speech and gesture can still contribute to the communicative message. For example, simple and humanoid or abstract tangrams will naturally cause gestures that are different by nature (i.e. gestures describing a circle vs. an ice dancer), but both of these conditions do not make gesturing itself harder; there is no factor that prevents the directors from producing gestures. The same is to be said for speech. The repetition manipulation can cause directors to produce linguistic descriptions that differ content-wise, but there is no factor present that prohibits them from speaking.

It is because of this that the tradeoff hypothesis and the hand-in-hand hypothesis should be studied to a further extent, in an experiment that does make communication harder with the use of background noise. Background noise was chosen because a noisy environment will likely hinder the production of speech, which allows us to study the speaker's modulations. Because the variable manipulation of the De Ruiter et al.'s paper did not necessarily complicate the production of gesture and speech and therefore the tradeoff hypothesis was not directly tested, we assume that speakers do follow this principle: when speaking is made harder, they will rely more on gesturing.

3. Present study

From what we have seen in the previous chapter, there seems to be more evidence for the tradeoff than the hand-in-hand hypothesis: Graham and Heywood (1975) found that speakers produce more speech to describe spatial relations when they were not allowed to gesture; Melinger & Levelt (2004) argued that speakers who used spatial relations in their gestures omitted more spatial relations from speech compared to speakers that didn't gesture; according to Bangerter (2004), the speakers produced fewer gestures when the distance to the target object increased, and fewer words when the target objects got closer. Yet, De Ruiter et al. (2012) found that the use of deictic gestures was positively correlated with the amount of locative expressions produced in speech, which supports the hand-in-hand hypothesis. We however propose that the manipulations applied by De Ruiter et al. concerned the codability of objects (i.e. the difficulty of the target objects), but not the difficulty to speak or gesture in itself. Therefore, we deem the results that argue for the tradeoff hypothesis more convincing than the ones for the hand-in-hand hypothesis. In the current study we will thus go by the tradeoff hypothesis.

Numerous researches have been conducted concerning communication in noise, though these studies have often presented participants with video stimuli of an actor or actress who uttered speech that was either clear or degraded. For this reason, questions remain how multimodal productions are *created in* a noisy environment. More specifically, more research is needed to study the multimodal productions and communicative strategy speakers apply when communicating though different levels of noise. This is what will be studied in this paper.

The aim is to focus on the production of both gesture and speech in a noisy environment. In an experimental set-up, participants will be exposed to three different levels of noise: there is a no-noise condition, a moderately noisy 4-talker babble condition, and a highly noisy 8-talker babble condition, in which they are asked to convey action verbs. The gestures the speech utterances are coded and analysed, as well as the gesture feature changes that the participants produced. The goal is to study the influence of the different noise levels on the production of both speech utterances and gestures. Furthermore, the communicative attempts are a point of focus. With communicative attempts we mean the communicative production that a speaker creates to get the message across to the listener. Going by the results found in Hoetjes et al. (2015), who studied communicative failure after negative feedback, we want to know whether a communicative failure (i.e. when conveying these action verbs is not

successful), while being surrounded by noise will influence the communicative productions of the speaker. The aim is to find out if directors make adjustments in their multimodal strategy in their second communicative attempt if their first one has failed to get the message across to the matcher.

This paper will aim to answer the following questions:

1) Which differences can be found in the production of gesture strokes in a moderately and highly noisy environment, as compared to a no-noise environment?
2) Which differences can be found in the production of speech utterances in a moderately and highly noisy environment, as compared to a no-noise environment?
3) Which differences can be found between the second communicative attempt and the first attempt with reference to gestures, speech utterances and change in gesture features?

The influence of noise is thus studied on speech utterances and gestures. This paper takes the gesture strokes as a dependant variable, and not the entire gesture phrase, as we wanted to focus on the part of the gesture most meaningful to the communication, the part that contains the communicative message, and to see how this part interacts with speech in noise.

Given the research that has been carried out thus far in this domain and following the tradeoff hypothesis, it is expected that the gesture and speech will compensate rather than parallel each other. More specifically, it is expected that, of the three noise conditions, the least amount of gestures and the biggest amount of speech utterances will be produced in the no-noise condition. This is expected as speaker will experience no communication problems in this clear environment. The contradicting theory, the hand-in-hand hypothesis, would predict that the speech and gesture follow each other, which would here mean that both modalities would increase or decrease together as the noise level would get higher. Following Bangerter's (2004), Graham & Heywood's (1975), and Melinger & Levelt's (2004) studies, we go by the predictions of the tradeoff hypothesis. In sum, we think the no-noise condition will cause the most speech utterances, and the least gesture strokes.

In the moderate noise level with 4-talker babble noise, we predict that more gestures will be produced than in the no-noise condition, as well as less speech. It has also been shown that a double enhancement of gestures and visible speech positively influence speech comprehension mostly in a moderately noisy environment (Drijvers & Özyürek, 2017). We however don't know if the *productions* are significantly different in these conditions. At this

moderate noise level, the babble noise will not be interfering enough to completely mask the producer's speech utterances, but as it is likely that this noise hinders speech production, it is expected to influence the speech rate in that less utterances will be produced. The producer is expected to try to get the message across by using both the modalities of speech and gesture.

In the highly noisy condition, we will again assume the expectations of the tradeoff hypothesis: when the speech becomes difficult, then gesture will take over the communicative load. It is to be expected that the speakers will produce the least amount of speech utterances in this condition, since this level contains the highest level of noise interference which will cause speech transfer to become difficult, and the producer will rely more on the gestures to convey the message. We therefore predict that the gesture rate is highest in this condition, and speech rate the lowest.

As for our third research question, concerning the different attempts, we expect that the second communicative attempt will hold more gestures and speech utterances than the first attempt, as well as more gesture feature changes. We make this assumption as we expect that a speaker, once noticing the failed first attempt, will change the used communicative strategy to try to be more effective. We follow the results of Hoetjes, Krahmer & Swerts (2015), who showed that a failed communicative attempt leads to a lower speech rate and a higher gesture rate. Therefore, we assume that speakers in our study, after a failed first attempt, will adjust the communicative strategy and try to be more effortful: so they will use both gesture and speech to be as informative as possible, resulting in more gestures, more speech utterances and a wider variety of gestures, i.e. more gesture feature changes. We expect more changes in gesture feature also by taking into account the results Goldin-Meadow et al.'s (1996) study, where speakers produced a series of different gestures to describe an event when they were not allowed to speak. To be as informative as possible, we expect subjects to produce gestures that describe different parts or features, or a combination of different gestures.

4. Methods

4.1 Participants

The participants were recruited at Lowlands, a yearly music festival that takes place in the Netherlands. Participants volunteered for the experiment at the festival itself, and people could volunteer until all available places were filled.

A total of 182 participants, 91 dyads were recruited (97 females), most of whom knew each other (n = 86 dyads). Age of the participants ranged from 17 to 62 (M_{age} = 28,55 years). Participants gave written consent before the start of the experiment; if the participant did not sign the consent form, the participant was excluded (n = 7).

Of all the participants, 175 had Dutch as their native language. Of the seven remaining participants, of five participants the data were missing, and two reported a different native language (Russian and Armenian). In the case of these participants, Dutch was their second language. All of them reported their alcohol and drug use: either 0 drinks (n = 74), between 1 and 3 drinks (n = 70), between 4 and 6 drinks (n = 17) or more than 6 drinks (n = 17).

Of the 91 pairs that participated in the experiment, twenty were excluded due to audio-visual failures (n = 13) or problems with the consent forms (i.e. when the forms were not signed) (n = 7). Subsequently, the participants who first took on the role of matcher and then of director were also excluded, as they were primed. So, of every dyad, only one person was taken into account for the analysis, resulting in 71 participants. For the current study, the productions of a total of 56 participants were coded and included in the analyses. The 15 individuals that were not taken into account had been excluded due to the scope of the paper. Of the participant group that was taken into analysis, 24 were male, with M_{age} = 28,52 (min = 17, max = 62). Fifty-five participants had Dutch as a native language; one had another (Armenian). Fifty-two directors knew their communicative partner. Most had 0 drinks (n = 27) or between 1 and 3 (n = 21). Eight participants reported either between 4 and 6, or more than 6 drinks (n = 5 and n = 3 respectively).

4.2 Stimulus materials

In the experiment, one of the two participants was assigned the role of director, the other one of matcher. The director was presented with twenty Dutch action verbs, written on a piece of paper. These verb served as the stimulus materials of several studies (see Drijvers, 2019). To

make sure that the Dutch action verbs could easily be expressed with iconic gestures, the verbs were all highly frequent and were all pre-tested.

4.2.1 Pre-test

The pre-test served to test whether the verbs could be disambiguated by iconic gestures. In order to do this, stimulus materials of an actress uttering the verbs, and making gestures that either matched or mismatched the uttered verb. In the pre-test, 170 video stimuli were presented on a computer screen to twenty native Dutch speakers (10 female, M_{age} = 22,2, SD = 3,3) who had no motor, neurological, visual, hearing or language impairments. The video stimuli contained a gesture but were presented without the auditory information that contained the verb. The stimuli were presented in a randomised order. The participants first saw a fixation cross for 1000 ms, which was followed by the video. Afterwards, the participants indicated the verbs that they associated the movement they saw with. They were then shown the verb that the actress had originally uttered, after which participants were asked to indicate if they thought the movement fit the verb presented on the screen, going from "does not fit the movement at all" to "fits the movement really well" on a 7-point scale. The participants could take two breaks during the course of the experiment (after items 55 and 110), and they completed the experiment in approximately 35 minutes.

The answers that were given concerning the verbs that the movement was associated with, were used to determine if verbs had to be renamed, or if there were verbs in the stimuli that were unrecognisable and had to be deleted. Correct verbs or synonyms were coded as correct, unrelated verbs as incorrect. The mean recognition rate was 59% for all gesture videos. This is an acceptable rate, given that the gestures used in the videos were co-speech gestures, and therefore produced together with speech, but the speech itself was left out. Very high recognition rates would indicate that the gesture was more of a pantomime than a co-speech gesture. There were four items that had a very high (> 95%) or very low (< 15%) recognition rate which were removed from the data set.

The results of the answers participants gave that concerned how well the movement fit the verb showed that six videos did not get a rating above 5 on the 7-point scale. Over the other videos, the mean score of iconicity was 6.1 (SD = 0.64). Ten items were removed from the data set, which resulted in 160 verbs with a gesture, of which 80 contained a gesture-speech match and 80 a gesture-speech mismatch. Of this dataset the stimuli for the main experiment were created (see also Drijvers, 2019).

4.2.2 Main experiment

A total of four sets of each twenty verbs were made, to prevent a priming effect for the following participants. A couple of verbs (four in total – *beklimmen* (to climb), *bidden* (to pray), *drijven* (to float) and *filmen* (to film)) were used in more than one verb set. In total, 76 different verbs were used. An overview of all four verb sets is presented below in Table 2.

Table 2. Overview of the stimuli. Four verb sets were used, to avoid priming of the previous and following pair. In every experiment conducted, the director was given one set, and the matcher (who later took on the role of director) the other. On top of that, the succeeding pair were given the two other sets, in order to prevent a priming effect.

	Verb set			
	Set 1	*Set 2*	*Set 3*	*Set 4*
1.	beklimmen	rollen	roeren	ophangen
2.	bidden	dobbelen	hameren	uitrekken
3.	drijven	strijken	afgieten	boren
4.	filmen	afdrogen	filmen	klappen
5.	stempelen	slaan	beklimmen	schudden
6.	raspen	drummen	bidden	zagen
7.	skiën	verplaatsen	drijven	zouten
8.	boksen	opzoeken	vangen	aankruisen
9.	bladeren	tekenen	wijzen	hakken
10.	slingeren	wassen	dippen	plukken
11.	verbinden	toetsen	groeien	stapelen
12.	wenken	smeren	poolen	pellen
13.	krassen	snijden	vlechten	vissen
14.	weggooien	zwemmen	breien	toeteren
15.	kietelen	plakken	verschuiven	ritsen
16.	schrijven	darten	jojoën	fietsen
17.	steken	ontkurken	indrukken	omdraaien
18.	schrobben	verdelen	opkloppen	knippen
19.	kloppen	schroeven	hijsen	schuiven
20.	timmeren	golven	rijden	mixen

Throughout the experiment, both participants were wearing headphones. Through these headphones, noise could be channelled, which consisted of babble noise; unintelligible

speech of people talking at the same time. This babble either consisted of four people, or eight people talking. The difference between these two conditions was not the amplitude of the speech signal, but merely the interference: the 8-people babble noise has a higher interference than the 4-people babble noise. These noise levels were high enough to shut the participants off from their environment, but were below the pain threshold. Other than the babble-noise, one condition consisted of no noise, giving three used conditions in total: no noise, 4-people babble noise, and 8-people babble noise. The noise level was changed for the director every round. The matcher heard the same noise level throughout the experiment: the 4-talker babble noise.

In order to avoid priming as much as possible, four different verb sets were used, two per pair. For the majority of the participants, these verb sets were applied in ascending order (so verb set 1 for participant 1 of pair 1, verb set 2 for participant 2 of pair 1, verb set 3 for participant 1 of pair 2, verb set 4 for participant 2 of pair 2, etc.). To further avoid the influence of priming, the participants who first took on the role of matcher and then of director were excluded from analysis. These participants in the matcher position usually (but not always) were attributed the verb set 2 or 4 (1 and 3 often being given to the participants who first took on the director position). For this reason, verb set 4 was not at all included in the analysis, and verb set 2 only several times (due to a mix in verb sets). A total of 56 verbs remained (i.e. verb sets 1, 2 and 3 with 4 double verbs).

4.3 Set-up

In the experiment, two participants were standing in a face-to-face position in an indoor space on the grounds of Lowlands festival. During the course of the experiment, one of the participants took on the role of director, the other of matcher. The role of the director was to convey the twenty Dutch action verbs, the role of the matcher was to guess which verb the director tried to convey. In between the participants was a one-way mirror, allowing the matcher to see the director, but not vice versa. With this one-way mirror, the director knows that he/she is visible to the matcher, and thus is not likely to reduce his/her gestures. At the same time, the (gestural) feedback of the matcher will not influence the director.

At the side of the participants is the experiment leader seated, who shows the director the cards with the Dutch action verbs, one at a time. Furthermore, the experiment leader communicates to the director whether or not the matcher has correctly guessed the word, or

signals the director to move on to the next word. The experiment leader is seated facing the mirror, so that he/she can see both the director and the matcher.

Two cameras were positioned at the side of the setting, which are aimed at each of the participants, one camera for each. Other than the cameras, the director is tracked by two kinects, and eye-tracking was also done for the matcher. Apart from the experiment leader, two experiment assistants were seated also at the side of the participants, further removed than the experiment leader. The assistants controlled the noise level of the participants: whenever the matcher had guessed a word and the round ended, the assistants changed the noise level that the director heard.

4.4 Procedure

Before the experiment was started, the experiment leader gave a verbal instruction on what the participants could expect. They were told that the director had to convey the action verb to the matcher, and that they both could hear noise through the headphones. The director was also told to use whichever strategy he/she felt most comfortable using: gesture, speech or both.

During the experiment, the director tried to convey the twenty Dutch action verb that was shown by the experiment leader on the card. Every verb indicated an experiment round. With twenty verbs, the experiment thus consisted of twenty rounds. Whenever the matcher correctly guessed the word, the experiment leader raised a thumb, to indicate to the director that this round has successfully ended. In case the round would take too long, and the matcher failed to guess the word, the experiment leader put a thumb down, to indicate to both parties that the current round has been unsuccessful, and the next one will start. In case the director feels stuck in conveying the verb and thinks the matcher will not guess it, he/she can also lower a thumb. In this case, the round will also end, and the experiment leader will show the next verb.

4.5 Coding

The coding of the video data is done in the multimodal analysis programme ELAN (Max Planck Institute for Psycholinguistics, Nijmegen, https://tla.mpi.nl/tools/tla-tools/elan, see

also Geerts, 2018). For every video, the speech and the gestures are annotated, as well as the gesture features (see 4.5.2).

4.5.1 Gesture phrases and strokes

The gesture phrase can consist of a preparation phase, a pre-stroke hold, a stroke phase, a post-stroke hold and a retraction phase (Kita, Van Gijn & Van der Hulst, 1997; McNeill, 2005). Two consecutive gesture phrases can mean that the exact same gesture is produced twice, or can mean a gesture with slight changes, or a complete new gesture. Sometimes, the gestures are succeeded one after the other without gap. In this case, a clear change in gesture feature decides the beginning of a new gesture. This can mean a kinematic change, but also a semantic one. It is explained in more detail below.

The gesture strokes are also coded separately, not only as part of the gesture phrase. This has been done because coding just the main part of the gesture can give more insight on how the stroke is produced to co-communicate and interact with the speech utterances. The stroke is characterised by more force being exerted than its neighbouring phases, and acceleration (or deceleration) are indicators of the exerted force (Kita et al., 1997, p. 8). As the stroke of the gesture contains the communicative message, we used the gesture strokes for analyses.

4.5.2 Change in gesture features

We have seen that changes in referent can occur when every gesture described a different referent or action than the previous one, like in Goldin-Meadow, McNeill & Singleton (1996), or that gestures may depict different viewpoints (Parrill, 2009).

We propose that there are four more different gesture features that can be combined to create descriptions. These are direction, location, hand and arm. Our own data have provided examples of participants producing the same gesture, but merely changing the direction in which the gesture is performed. It could be argued that this is done either to provide an expansion of the ways in which the action verb that is being embodied can move (for example ‘to iron’ which is gestured in different directions by one of our participants), or simply to provide the matcher with a better view of the gesture.

Other gesture combinations can hold different locations in order to provide a more

thorough description of the conveyed action or event. For example, when gesturing ‘to tickle’, participants in our data produced a tickling gesture on the belly, chest and the arm pit, to fully describe the action of being tickled.

Furthermore, changes can occur in the hands and arms of the participants. Hands can change in shape or position, as shown in our experiment. It was stated by Trujillo et al. (2019) that “*the improved comprehension* [in a more communicative compared to a less communicative context] *may come from fine-grained kinematic cues, such as hand-shape and finger kinematics*” (p. 7). Arguably, change in hand can be done in order to put more emphasis on the fingers performing the action: in the data, a director changed her handshape when gesturing ‘to write’. In the first gesture, the thumb and index finger were pressed on top of each other, as to hold a pen, while the other three fingers were curled in. In the following gesture, these three fingers were extended, thus arguably giving the conversational partner a better view on the two fingers performing the action. Gestures can also change in arm, i.e. a gesture that is performed with the right arm can be produced with the left one in the succeeding gesture. It is also possible that the subject produces the gesture with one arm first, and then produces the same gestures while adding the other arm or vice versa. Our data has shown subjects gesturing ‘to hammer’, and first producing this gesture with one arm, immediately followed by a gesture with an added arm. It is not yet clear why this occurs.

These six changes in gesture features (referent, viewpoint, direction, location, hand and arm) will be part of this current study. In the rest of this paper, we will refer to these gestures as gesture feature changes.

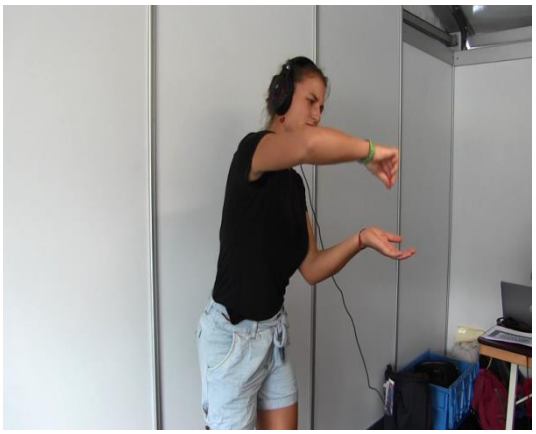
The gestures are annotated based on a (partial) retraction of the limbs, a gap between them or a clear change in gesture features. The gestures in one round (thus which all refer to one verb) are compared with each other. If there is a change in the gesture features that is immediately prior to the gesture in question, then this is indicated in ELAN.

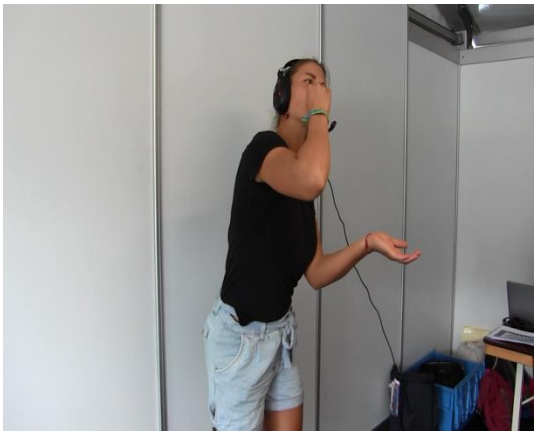
In sum, we coded seven different gesture feature changes, as proposed in the literature chapter: change of *referent*, change of *viewpoint*, change in *direction*, change in *location*, change in *hand*, change in *arm*, and change *other*. These will all be explained with examples below.

4.5.2.1 Change of referent

A change of referent indicates a new gesture that refers to a new item, part or action that is associated with the verb that is to be conveyed, usually with the intention of giving a better understanding of the context of the verb. An example of this is given below. Here, the director tries to convey the verb *dippen* (to dip) towards the matcher. Instead of just producing the action verb of dipping, the director first makes a circle gesture with her right hand on top of her left, to indicate a bowl (first image), followed by the dipping gesture made with the right hand in the second image, the left hand still indicating the bowl. In the third image, it is shown that the director elaborates on the verb even more by producing an eating gesture: the right hand is moved from the bowl towards the mouth. These three gestures a coded as belonging to one single attempt with related gestures when the gestures in isolation refer to another verb than the verb to be conveyed. In other words, while the second gesture in the example refers to the action of dipping, the other two refer to an item and action linked with it. Seen in isolation, the depictions of a bowl and an eating action would not convey the verb *dipping*. These are therefore seen as multiple related gestures, providing different features or actions of one verb.

Figure 1: Change of referent while conveying the verb dippen *(to dip): the first image the director makes a circle gesture with her right hand on top of her left to depict a bowl. In the second, she makes a dipping gesture with her right hand (in the bowl that was referred to), after which she brings her right hand to her mouth to depict the action of eating.*

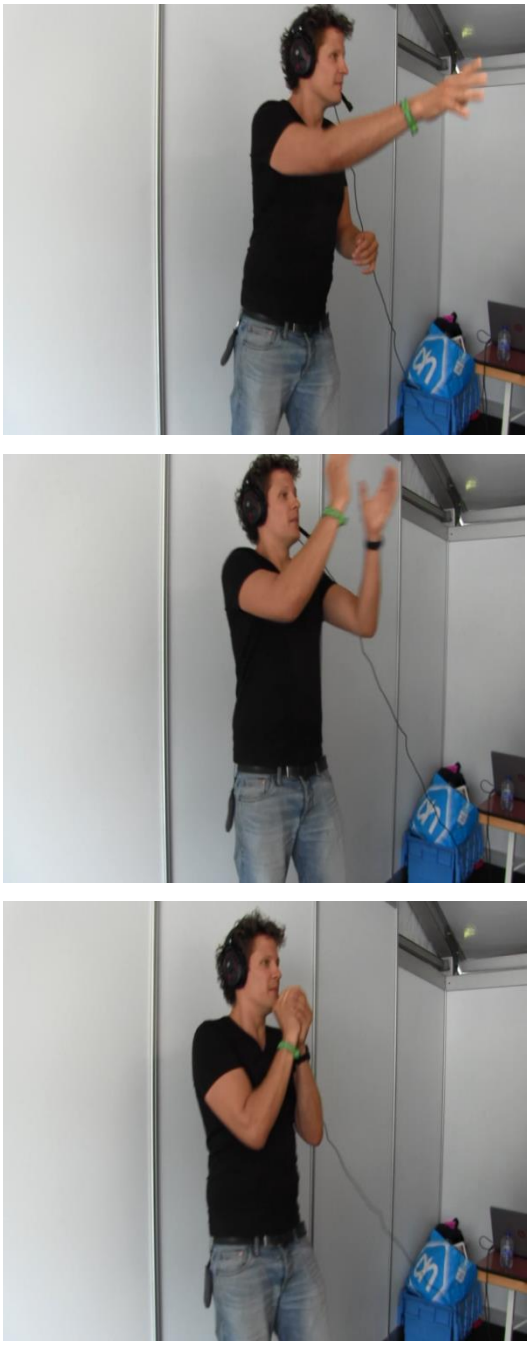




4.5.2.2 Change of viewpoint

When two succeeding gestures indicate a change of viewpoint, then the second gesture is seen from a different point of view than the first one. In other words, in one gesture the action is performed on the director, the director being the receiver of the action, and in the other, the director performs the action himself, and is the producer of the action. In the example below, the director tries to convey the verb *vangen* (to catch). Here, the director first makes a throwing gesture, after which he makes a catching gesture. He thus tries to convey the verb by embodying both sides of the action.

Figure 2: Change of viewpoint while conveying the verb vangen *(to catch). In the first image, the director extends his right arm forward in a throwing gesture. In the second and third images, he brings his hands together in the air and then lowers his arms, as to depict to catch something.*



4.5.2.3 Change in direction

A gesture phrase is coded as showing a change in direction when nothing in the movement of the gesture changes, apart from the direction in which the action is performed. Only moving action verbs are coded as showing a change in direction. Figure 3 shows a participant gesturing the verb *strijken* (to iron). The director continues the gesture in the same speed, size, and movement (moving back and forth of the arm), only the direction has changed: the first

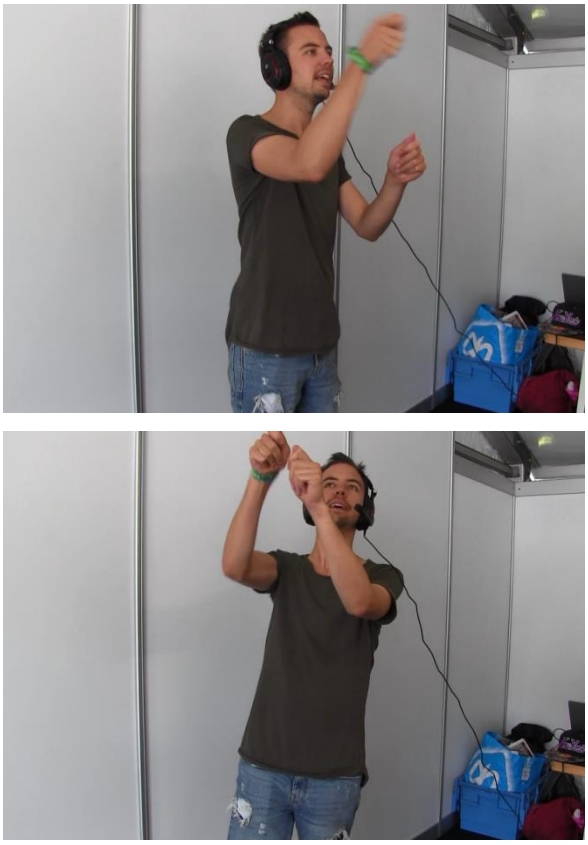
image shows the movement going to the left side and back, the second image goes to the right side and back.

Figure 3: Change in direction while conveying the verb strijken *(to iron). The director holds his right hand still as to hold something (i.e. clothes), and his right hand makes ongoing movements that go to the left and then back to the middle. In the second image, the director is holding his left hand in the same position, only now the right hand makes ongoing movements going to the right and then back to the middle.*

4.5.2.4 Change in location

This gesture change refers to gestures of which there is no or not much movement in space. Instead, the gesture is performed in a specific space, or on a specific place on the body, after which the hands move and the exact same gesture is performed in another place. An example of this is a participant conveying the verb *timmeren* (to hammer), which is performed in two or more different places in space. In the figure below, the director first makes a hammering gesture in the gesture space right in front of him, and then moves to his right to repeat the gesture.

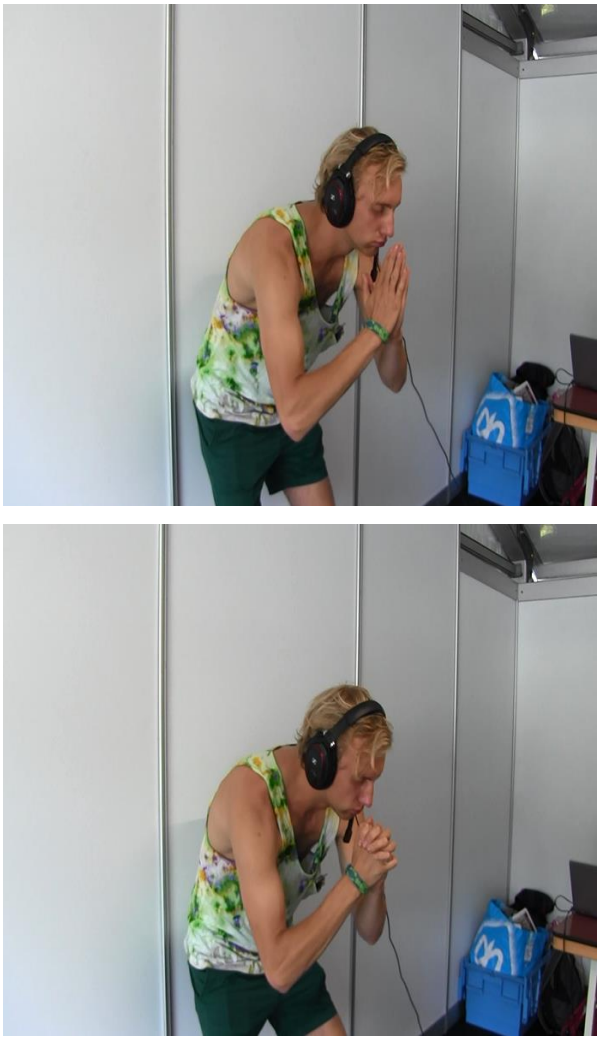
Figure 4: Change in location while conveying the verb timmeren *(to hammer). The director holds his left hand still as if to hold a nail, and with the right hand he makes quick movements to depict a hammering gesture. In the second image, the movements are the same, only the director has moved the gesture space from the one in front of him to the high right gesture space.*



4.5.2.5 Change in hand

A change in hand indicates a change that happens in the shape of the hand, or the position of the hand. Figure 5 depicts a change in hand, when the director is conveying the verb *bidden* (to pray). The director keeps the same (arm) position, but merely folds the hands. These differences tend to be very small. Therefore, two gestures are considered as having a change in hand only if the matcher can notice it. In order to decide this, the gestures were played in real time. Were these changes visible to the naked eye, then it was coded as change in hand. In this paper, hand or finger movements that differed 90 degrees or more from the previous gesture depicted a change in hand. This real time coding has led to change in handshape mostly occurring in gestures that happen immediately one after the other, with no considerable gap in between.

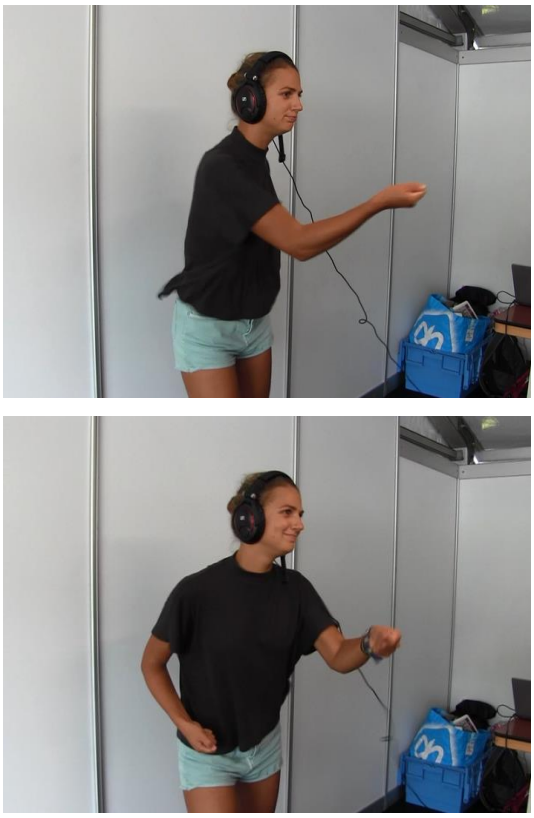
Figure 5: Change in handshape while conveying the verb bidden *(to pray). The director stays in the same position in both images, but only folds his hands in the second gesture.*



4.5.2.6 Change in arm

When a participant depicts the same (one armed) gesture, but changes his arm the second time, then this is coded as a change in arm. Also coded as change in arm are cases where the director adds an arm (when a one handed gesture becomes two handed) or omits one arm (when a two handed gesture becomes one handed). A change of the position of the arm also falls in this category. Figure 6 depicts the change in arm: when the director tries to convey the verb *steken* (to stab), she first makes a stabbing gesture using the right hand, and then goes on to repeat the gesture with the left hand.

Figure 6: Change in arm while conveying the verb steken *(to stab). The director lifts her right arm from her hip and extends it in a quick, flicking way, as to stab someone, and then immediately retracts it. After that, the left hand leaves the resting position and makes the same quick, flicking gesture, and also quickly retracts the arm.*



4.5.2.7 Change other

The final category of gesture changes contains the remaining gestures that did not fit any other category. Most of these cases contained a change in speed, a change in size, and sometimes even a whole new interpretation of the verb. The figure below shows an example of a change in size. The same gesture is being produced (a hammering gesture), with the only difference being that the first gesture uses a small gesture space in front of the director, and the second using a large one, with the working hand being lifted all the way above the director's head. In order to decide if two gestures differ enough to be coded as having different sizes, the gesture phrases are played in real time. As with the hand gestures, the reason for this is that for two gestures to be coded as having different sizes, the matcher should notice the difference between them in real time. The gestures are thus not compared pixel by pixel: if the naked eye can see the difference (and hence the matcher can spot the difference), change in size is considered to occur.

Figure 7: Change other (size) while conveying the verb hameren *(to hammer). The director holds her left hand still in front of her to depict holding something, and the right hand makes an up and down flicking gesture that stays in front of her chest. In the second image, the left hand stays in the same position, but the right and is raised to the level of the director's face, and is then flicked up and down in the same manner as in the first gesture.*

The change in interpretation concerns verbs that can be explained in more than one way. In case of a whole new interpretation, the new gesture always marks the start of a new attempt. Below is an example of a whole new interpretation of the verb *hijsen* (to lift). In the first attempt, the director lifts his hands one after the other as to pull something up with the use of ropes. In the second interpretation, he lifts his trousers.

Figure 8: Change other (interpretation) while conveying the verb hijsen (to lift). First, the director raises his arms one at a time to make a grabbing gesture in the gesture space above him and then lower the arm as to lift (or haul) something. In the second gesture, the director grabs the band of his trousers with both hands and lifts it upwards, as to hold up ('lift') his trousers.

Gesture changes that have been coded as *other* are always accompanied with a comment on a separate tier, explaining the kind of change.

Change in gesture feature was indicated for every gesture in the round of the experiment, apart from the first one. Each following gesture would then be compared to the one directly preceding it.

4.5.3 Speech coding

Speech utterances that were relevant for the experiment were coded. This could mean that the participant said the verb itself, or described it in one way or another. These speech utterances can thus consist of one single word, but also of a sentence. Below, some examples of speech utterances are shown that were produced by the participants:

3) Zwemmen | in het water
   Swimming | in the water
   verb: to swim
4) Hihihi | kietelen
   Hihihi | tickling
   verb: to tickle
5) Je doet het met komkommers
   You do it with cucumbers
   verb: to grate
6) Papier | pen | schrijven
   Paper | pen | writing
   verb: to write
7) Je bent klein, je bent groot, wat doe je dan
   You are small, you are big, what do you do then
   verb: to grow

In example 3, first the verb is mentioned, and when that was not informative enough, the director added more information as to where the action happened (*in the water*). Example 4 is characterised by a sound closely linked to the verb tickle: a laughing sound. This laughing sound was accompanied with a tickling gesture. The verb that followed was merely shouted at the matcher (without a gesture). Examples 5 and 7 do not contain the verb itself, but are descriptions of them. The third sentence indicates with what the verb (to grate) can be used. In the fifth, the director describes the state of the subject before and after the action (being small and being big). Finally, in example 6, first two different objects are described that are linked to the gesture: paper and pen. These objects were accompanied by gestures: the director gestured a sheet of paper and a pen. It is only after that, that the verb itself is uttered. Here, the director gestured the action of writing.

The speech was coded by looking at the gaps between speech utterances. On average, a pause is thought to be 200 to 400 ms, and the minimal response time is about 200 ms (Bögels & Torreira, 2015; Fry, 1975; Stivers et al., 2009). Therefore, when coding the speech utterances, a gap of 200 ms marked the start of a new attempt.

It is however possible that the director has trouble finding words and pauses mid-sentence. This can cause a gap of at least 200 ms, even though the utterances before and after are clearly linked to each other. Because of this, the Turn Constructional Units were also

looked at. These TCU's are described by Clayman (2013) as "*a coherent and self-contained utterance, recognizable in context as 'possibly complete*' (p. 151). The ending of a TCU is characterised by whether an utterance is intonationally complete (a falling tone), grammatically complete (the syntax signals the end of the utterance), and pragmatically complete (the utterance has accomplished its purpose). The ending of a Turn Constructional Unit marks the possibility of the listener to respond or give feedback (Clayman, 2013; Ford & Thompson, 1996). A speech segment was thus coded as one annotation if they indicated one 'unit'. Following these results, either a gap of 200 ms or more, or a clear ending of a TCU was considered a new attempt. These two ways of coding were not found to contradict each other.

4.5.4 Attempts

The produced gestures and speech utterances are all part of communicative attempts. A single communicative attempt holds a single time the director tries to get the verb across to the matcher, after which the director provides the matcher with an opportunity to respond or give feedback (for example with eye-gaze) (based on Hoetjes et al., 2015; Trujillo et al., 2018, 2019).

Producing an attempt can be done with gestures alone, speech alone, or a combination of the two. If the attempt consists of gesture(s) alone, the attempt starts with the preparation phase of the first gesture, when the hands leave the resting position. The attempt ends after the retraction phase of the last gesture, when the hands return to a rest position or move on to a new gesture. If several gestures are succeeding one after the other, then the gap between them is looked at. The minimal response time is taken as an indicator for the start of a new gesture. So, if a gap of 200 ms fell between the gestures, then the two gestures were annotated as two different attempts; a smaller gap would result in one attempt. If however, a gesture is immediately followed by another one, without a gap in between, but there is a clear retraction of the arms after the first gesture before going on to the second, then this is seen as a clear indication of the ending of an attempt, and the second gesture would be coded as a new attempt. Only the first and (if present) second communicative attempt were coded, due to the scope of this paper.

In coding the attempt, we paid attention to the communicative intent of the speaker. The communicative intent of the speaker was decided by his or her initiated addressee-directed eye-gaze, as found by Trujillo et al. (2018).

As has been shown in the example 1 to 5 above, directors often used gesture and speech at the same time. The produced gesture and speech did however often not start nor end at the same time. As proposed by Habets et al. (2011), speech and gesture indeed don't always start at the same time: the producer often inserts speech after the gesture, to make the message as informative as possible for the receiver. Speech and its accompanying gesture often partly overlap, but the gesture usually starts before the speech (and not after). Habets et al. investigated when exactly the asynchrony of speech and gesture onsets are most optimal for the semantic integration of the gesture and speech. In their experiment they delayed the speech onset in comparison to gesture onset with 160 and 360 ms, and measured the semantic integration of speech-gesture matches and mismatches. They found that speech and gesture are integrated most efficiently at a speech delay of 160 ms, but are not at a speech delay of 360 ms, implying that there is a time span (somewhere between the SOA of 160 and 360 ms) up until which gesture and speech are most efficiently integrated.

Both the results of the paper of Habets et al., and the minimal response time were considered during the coding of the results of this paper. Whenever the stroke of a gesture phrase overlapped with a speech utterance, then this was coded as one attempt. In case the speech utterance overlapped with the gesture preparation or retraction phase, an overlap of 200 ms was considered one attempt. If the overlap was less than this time window, the speech utterance and the gesture phrase were coded as two separate attempts.

4.6. Data analysis

After the coding was done in ELAN, the data was exported to the statistical programme R. The data was checked by means of residual plots, density plots, histograms and Q-Q plots (Winter, 2013), all for both the gesture strokes and the speech utterances. A violin plot was produced for the gesture strokes. For both the gestures and the speech, the residual plots showed linearity. However, the histograms and Q-Q plots showed that the data were not normally distributed: the histograms were not bell-shaped and the Q-Q plots did not show a straight line. To check the data distribution, several normality tests were performed. Skewness was 2.719 for the gesture strokes and 0.945 for the speech. Both are thus positively skewed; for both variables the probability mass concentrates on the right tail. Kurtosis for both variables is 9.392 for gesture strokes and -1.108 for speech. This, too, indicates distributions that are not normal. To verify that the data are not normal, several distribution tests were

performed. The Anderson-Darling test, Cramer-von Mises test and the Lilliefors (Kolmogorov-Smirnov) test (Gross & Ligges, 2015) all suggest that the data are not normally distributed ($p < .001$ for all tests for both variables). Data are thus not normal and positively skewed. In order to solve this non-normal distribution, log-transformations were applied to both variables: a common (base) log-transformation with the gesture strokes, and a log2 (a binary, base 2) transformation with the speech. The log2-transformation for the speech was chosen because the response is binary (the speech utterances were coded as either being present or absent).

After the log-transformations, generalised linear mixed models (GLMM) were applied (Bates et al., 2019), with Poisson family (Jabeen, 2019). This Poisson Regression Model is "*a Generalized Linear Model (GLM) that is used to model count data and contingency tables*" (*ibid*), and it is used when the results are counts (i.e. the number of times that an event occurs: the number of times the directors produce gesture strokes and speech utterances) (*ibid*).

Several models with different variables were made to see which would fit the data best; the fixed and random variables taken in to account were noise level, verb, participant, age and gender. By looking at the AIC (Sakamoto & Ishiguro, 1986), it was determined which model fit the data best: a lower AIC score fits the data better than a high one.

The generalised linear mixed effects model that fit the data best looks like this:

gesture strokes / speech utterances ~ noise level + (1 + noise level | participant) + (1 | verb)

These analyses were executed on both the gesture strokes and the speech utterances with the *noise level* containing factors 0, 4 and 8 as a fixed effect, and the variables *participant* and *verb* as random effects.

In other words, the independent variable is the noise level, which is manipulated and subdivided into the three noise levels. The dependant variables are the number of gesture strokes and the number of speech utterances. The variable *participant* was included as a random effect, as the factor noise level will be taken into consideration more by some participants than others. This results in a random slope in the model. The variable *verb*, too, is included in the generalised linear mixed model, since some verbs might cause more gestures (or speech utterances) to be produced than other verbs, causing a random intercept to be included in the model. For these analyses, all rounds and attempts were included: thus the speech-only, gesture-only and the multimodal ones, as well as attempt 1 and attempt 2.

In order to answer the third research question concerning the differences between attempt 1 and 2, an adjusted data set was made which only contained the rounds that have two attempts. Rounds consisting of only one attempt were thus excluded. With this adjusted data set, two subsets were made of attempts 1 and 2 (in each subset respectively). In other words, all the first attempts were assembled in subset 1, and the second attempts in subset 2. Subsequently, Wilcoxon rank sum tests were carried out for the variables *speech*, *strokes*, and the gesture feature changes. The Wilcoxon test was used because both samples are from repeated observations of the same subject group, and this analyses can be executed with data that are not normally distributed.

5. Results

As explained in earlier chapters, this paper follows the tradeoff hypothesis. To operationalise this hypothesis, the amount of gestures per attempt as well as the number of speech utterances was computed. The focus of the speech was thus merely on the *frequency* with which the producer made an utterance related to the verb, and not on *what* was said.

For the gesture production, a violin plot was created. The plot is shown in figure 9. It shows the gestures for every noise level, as well as the mean and the median. The median of produced strokes is 1. In other words, most often, only one gesture was produced by the participants. The mean for all three noise levels is at around 1,3.

Figure 9. Violin plot of number of strokes per noise level. Included are median (black line) and mean (red point).

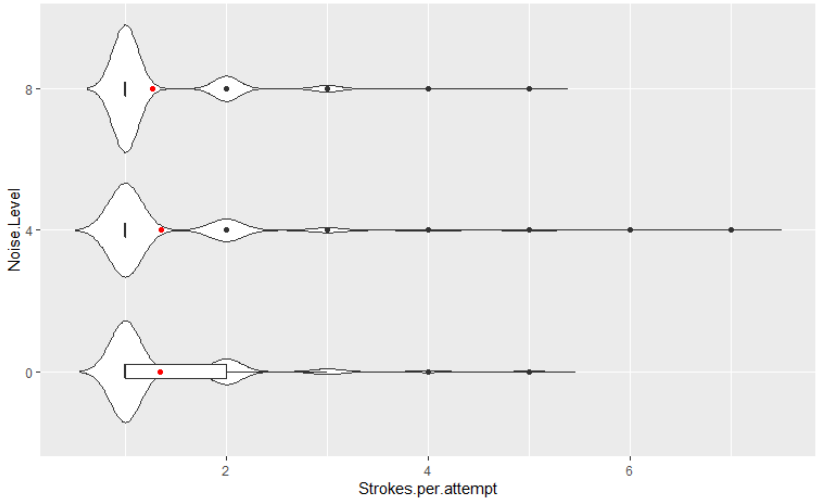


Figure 9 shows that, in most cases, directors gestured only once. We can also see that they barely gestured three or more times. Importantly, the figure shows that most gestures were produced in the 4-talker babble condition: up to seven gestures were produced. This is in contrast with both the no-noise condition and the 8-talker babble condition.

After the distribution of the data was checked with distribution plots, normality tests and a violin plot for the gesture strokes, a Generalised Linear Mixed Effect Regression model was produced to predict the dependant variable with *noise level* as a fixed effect and *participant* and *verb* as random effects.

5.1 Gesture strokes

A generalised linear mixed model showed that there were on average 1.34 gestures in the no-noise condition (SD = 0.685). A similar result was found in the 4-talker babble condition (1.35, SD = 0.743). In the 8-talker babble condition a mean of 1.26 gestures were produced (SD = 0.573). A Wilcoxon rank sum test (Bauer, 1972; Hollander & Wolfe, 1973) performed between the three noise conditions showed that there was a significant difference between the no-noise and the 8-talker babble condition ($p = 0.01$), as well as between the 4-talker and the 8-talker condition ($p = 0.02$). No significant difference was found between the no-noise and the 4-talker babble condition.

However, the maximum likelihood ratio test (Laplace Approximation) (Bates et al., 2019) showed that, overall, noise level did not affect the production of gestures significantly ($X^2(2) = 2.520$, $p = 0.285$. This suggests that speakers do not significantly adjust their gesture production when the noise level changes.

5.2 Speech utterances

A total of 2358 speech utterances were made. On average, a mean of 0.30[1] speech utterances were produced in the no-noise condition (SD = 0.459). In the 4-talker babble condition, this amount is slightly lowered to 0.28 (SD = 0.450). In the 8-talker babble condition, the mean was 0.27 (SD = 0.445). The Wilcoxon rank sum test showed that there is a significant difference in the speech utterances of the no-noise condition and the 8-talker babble condition. The other two comparisons showed no significance.

The likelihood ratio test showed that here too, noise level did not significantly affect the speech production ($X^2(2) = 1.335$, $p = 0.513$). Manipulating the noise levels was thus not of influence on the speech production.

5.3 Attempts 1 and 2

In order to compare gesture and speech productions between attempt 1 and 2, an adjusted data frame was made from the original data, which consisted of only the rounds in which two attempts were produced. The rounds in which one attempt was sufficient for the

[1] In the case of speech, utterances were converted to 1, as the amount of utterances were counted. Coding for speech was thus done with either 0 or 1: 0 indicating absence of speech, 1 indicating presence of speech. Therefore, the means of speech are between 0 and 1.

communicative task to be successful were excluded. This new data frame was then split into two subsets: one containing the gesture strokes and speech utterances made in the first attempt, the other containing those made in the second. Subsequently, a Wilcoxon rank sum test was performed on several variables. These showed interesting outcomes. A summary of the Wilcoxon test results is given below in table 3. Also included is the number of cases the variable is coded to occur (n), the mean and the interquartile range. This is a measure of the range of the middle 50% of the distribution, the difference between the upper quartile and lower quartile (Jassim, 2013)[2]. The *n* is obtained with the count function in R, the mean and the IQR with a summary of the two subsets.

Table 3: Wilcoxon test results with W-value and p-value for variables strokes per attempt, speech, and change in referent, viewpoint, direction, location, hand, arm and other. Included also are number of coded occurrences (n), mean value and interquartile range for attempt 1 and 2.

	Attempt 1			Attempt 2			Wilcoxon rank sum test	
Variable	**N**	**Mean**	**IQR**	**N**	**Mean**	**IQR**	**W-value**	**p-value**
Strokes	265	1.288	0	369	1.404	1	240375	8.126 e-05
Speech	67	0.293[3]	1	134	0.280	1	1751306	0.404
Change of referent	116	0.436	1	127	0.346	1	53646	0.021
Change of viewpoint	7	0.026	0	3	0.008	0	50106	0.069
Change in direction	35	0.135	0	51	0.138	0	49087	0.928
Change in location	30	0.113	0	27	0.073	1	51169	0.083
Change in hand	30	0.113	0	25	0.068	1	51435	0.046
Change in arm	13	0.049	0	26	0.070	1	48157	0.268
Change other	25	0.094	0	41	0.111	0	48237	0.486

The results show that three variables show a significant difference: strokes (mean 1.29 for attempt 1, 1.40 for attempt 2, p = 8.126 e-05), change of referent (0.436 for attempt 1, 0.346 in attempt 2, p = 0.021) and change in hand (0.113 in attempt 1, 0.067 in attempt 2, p = 0.046). Speech thus did not change significantly, despite the amount of speech utterances

[2] Here, the interquartile range is computed instead of the standard deviation as the variables are not numerics.
[3] It should be noted that, apart from the gesture strokes, all variables in this table are coded with 0 or 1. For the change in gesture features, the gesture was compared to the one directly preceding it (given that they depicted the same verb), and were coded with either 0 (this gesture feature change did not occur) or 1 (this gesture feature change did occur).

increasing in attempt 2. The table also shows that, when a different iconic gesture was produced, this mostly resulted in a change in referent.

6. Discussion

6.1 Interpretations of the results

In this study, the aim was to investigate the influence of background noise on the production of gestures and speech and the differences in communicative attempts. More specifically, this study asked what differences are to be found in the production of both gesture strokes and speech utterances in a moderately and highly noisy environment as compared to a clear environment, as well as the differences in regards to gesture strokes, speech utterances and change in gesture features that are found between attempt 1 and attempt 2. There were three research questions that were central to this paper:

1) Which differences can be found in the production of gesture strokes in a moderately and highly noisy environment, as compared to a no-noise environment?
2) Which differences can be found in the production of speech utterances in a moderately and highly noisy environment, as compared to a no-noise environment?
3) Which differences can be found between the second communicative attempt and the first attempt with reference to gestures, speech utterances and change in gesture features?

Central to investigating the first part is the tradeoff hypothesis (Bangerter, 2004; De Ruiter et al., 2012; Melinger & Levelt, 2004) and the contrasting view, the hand-in-hand hypothesis (So et al., 2009). The former holds that gestures and speech compensate each other: when speaking gets harder, the speaker will rely more on gestures to take over the communicative load, and vice versa. The latter hypothesis presumes that gesture and speech parallel each other: if the speech rate decreases, then the gesture rate will decrease as well. We have studied which approach directors take in a communicative context when trying to solve the knowledge gap between them and the matcher. It is known that during an explanation or communicative context, the director tends to produce gestures that are larger, more complex and precise and are higher up in the gesture space (Campisi & Özyürek, 2013; Galati & Brennan, 2014; Hilliard & Cook, 2016; Trujillo, Simanova, Bekkering & Özyürek, 2018, 2019).

We have seen in the result section that for all three of the noise levels, in the majority of rounds, only one gesture stroke was produced, with the mean of produced strokes being 1,3. Plus, according to the violin plot shown in figure 9, there are not many rounds in which three or more gesture strokes were used to describe the verb. The maximum likelihood ratio test has

proven that there is however no significant difference in gesture strokes over the different noise levels (even though the Wilcoxon rank sum test showed a significant difference between the no-noise and 8-talker babble condition and the 4- and 8-talker babble conditions). This suggests that the director does not alter his or her gestures significantly when the noise level changes. As for the speech utterances, the likelihood ratio test has shown that the directors' speech utterances, too, were not significantly affected by the change in noise level (though here, too, the Wilcoxon rank sum test showed a significant difference between no-noise and 8-talker babble backgrounds).

These results are interesting, as they do not confirm nor deny the tradeoff hypothesis or the hand-in-hand hypothesis. The outcomes could imply that the directors choose a certain communicative approach, and do not make adjustments in it throughout the experiment. A possible explanation for this could lie in the design of the experiment and the duration of the noise levels. In the present study, the director and matcher wore headphones through which noise was channelled. The noise level of the matcher was the same throughout the experiment with a 4-talker babble, but the noise level the director heard changed per round. The round lasted until either the matcher had correctly guessed the word, or the director or experiment leader indicted that the round took too long and started the next round. Rounds could last for several minutes, but there have also been rounds that lasted some seconds before the matcher guessed the word. It is possible the noise condition did not stay on for long enough for the director to adjust his or her modality use to it. It anyhow seems that the directors were not so sensitive to the change in noise level that they immediately carried through changes in their gesture and speech rate. Should they have done so, then the director would have produced more gestures and less speech when the sound channelled through the headphones became more noisy as compared to the previous noise level, and similarly would have decreased the gesture rate and increased the speech rate when the succeeding noise level was less noisy than the previous one.

It should also be noted however that, of the studies that have been discussed in this paper that have studied the verbal, gestural or multimodal communication through noise in a face-to-face situation, most have not dealt with a change in noise level. Not much is clear or known when it comes to a speaker's adaption to changing noise levels in fluent communicative situations. In the studies discussed in earlier chapters of this paper, participants were often shown a video of a narrator producing speech utterances that were either created in noise or in a clear environment which the participants were asked to judge or recognise (for example Drijvers & Özyürek, 2017; Pittman & Wiley, part II, 2001; Van

Summers et al., 1988), or participants were asked to read sentences that were spoken in clear or noisy environments (for example Pittman & Wiley, part I, 2001; Tufts & Frank, 2003).

We do know that speakers tend to adjust their linguistic and gestural production depending on who their listener is (Campisi & Özyürek, 2013), and whether the context is communicative or not (Trujillo et al., 2018), but in both studies the speaker had to adjust their communicative strategy once upon learning about the listener or task. Plus, in our research, the directors were asked to convey a message in fast changing levels of noise. In other words, it might be that participants do not adjust to their environmental conditions, whether this is their listener or the surrounding noise, when the situation changes at a fast pace, which resulted in a non-effect of noise level on gesture and speech productions in the current study. Another possibility is that directors choose a strategy at the start of the experiment, and stick to it without letting the change in the surroundings affect it.

The second part of the experiment focused on the differences between attempt 1 and attempt 2. The Wilcoxon rank sum test, performed with the variables gesture strokes, speech utterances and all seven possibilities of gesture feature changes, has shown that only three of the variables were significantly different across the two attempts: these are the gesture strokes, the change in referent and the change in hand. The gesture strokes appear more often in the second attempt than in the first, the change in referent and hand less often in the second than in the first. In other words, the director produced significantly more gesture sequences that are characterised by a change in referent or hand in the first attempt than in the second attempt, but more gestures overall in the second. This first result is in line with our predictions: we expected that the gesture rate would increase in attempt 2, after a failed communicative attempt (as seen in Hoetjes et al., 2015).

The second result however is surprising, as we expected that more gesture feature changes would occur in attempt 2, meaning that the director would expand on the gestures by providing different referents of a gesture and create a more thorough description (which thus would result in a change of referent) the moment he or she would notice that the message had not yet come across to the matcher successfully. This result is in contrast with the results found in Campisi & Özyürek (2013), who found that subjects make their gestural productions more informative when talking to a child instead of an adult (and hence experienced a knowledge gap between them). The authors claimed that subjects used this communicative strategy to teach new knowledge and to create common ground between them and their listener. In the current study, directors produced more gestures that were characterised by

change of referent in the first attempt and thus embodied different angles and dimensions of one verb, hence creating a broader picture of the verb in question in the first and not the second attempt. It seems that the communicative strategy of the directors was to create a broader dimension of the verb immediately from the start. Furthermore it seems that, once the director has noticed the communicative attempt has failed, he or she tries a different strategy that contains significantly more gestures in the second attempt, but with gesture sequences that are not characterised by changes of referent. These gestures then, either contain a gesture feature change of one of the other categories, or the gestures are exactly the same to one another and are therefore not coded as having one of the gesture feature changes at all.

The fact that the gesture strokes significantly increased in the second attempt can be explained with a change in communicative strategy once noticing the failed first attempt. In short, we assume that once the first attempt has failed, the director seems to adjust their strategy and improve the effectiveness of their message. The increased gesture rate in attempt 2 is in line with the predictions proposed earlier in this paper: when the director has failed to communicate with the matcher we expected that he or she will produce more gestures to try to be more informative.

Other predictions that were made earlier in this paper were that the gestures produced in the second attempt would be characterised by a bigger variety in gesture features as compared to the first attempt. The results have shown the opposite: the only two gesture feature changes that were shown to be significant, change of referent and hand, were produced significantly more often in the first attempt. Producers thus did not produce a significant wider variety of gestures in the second attempt.

Deepening our knowledge concerning the production side of communication in suboptimal environments gives us more knowledge on which communicative approach is applied most often, and also which one is most effective. This knowledge, in turn, can be expanded to communication to people that are hard of hearing.

6.2 Limitations and suggestions for future research

The current paper has aimed to find some answers in the process of communicating in noisy environments. By doing this, we have ran an experiment of participants communicating a range of Dutch action verbs through noise. We have segmented and coded the gesture phrases, strokes and attempts, but also the seven gesture feature changes and the speech utterances, the latter which were also transcribed. With this experiment we have thus

assembled a broad data collection, which allows for more research to be done that goes beyond the scope of this paper. Some recommendations are discussed below.

As for the speech utterances, this experiment has not so much focused on the Lombard effect in terms of vocal intensity or spectral tilt, but more in terms of the amount of speech utterances made by the director. It would be interesting for future research to focus on these aspects of speech in combination with the change in noise levels: to investigate whether the vocal intensity increases when the noise level changes. Furthermore, this paper has analysed the *frequency* with which directors have uttered speech. The focus was thus not so much on the content of their speech. Follow-up researches could aim to investigate if the content of speech utterances does change, by either the noise level or over the attempts. We have seen in the method chapter (4.5.3) that there are several ways with which a verb can be explained: the producer can simply utter the verb, or can use noises or descriptions that are often linked to the verb in question (as is the case with "*hihihi"* in example 4 for the verb *to tickle*, or "*in the water"* in example 3 with the verb *to swim*). Moreover, directors can provide a description of the verb: in example 7) the director describes the verb *to grow* by uttering "*you are small, you are big, what do you do then*". We have seen in earlier chapters that directors make their speech more informative when they notice a knowledge gap between them and their listener. It would be interesting to see if they would adjust the informative content of their utterances if they are exposed to a certain noise level, or if they are forced to produce more than one communicative attempt.

Another interesting elaboration on this paper with the use of the same data collection is to focus on the director-specific communicative strategy. Our results have shown that the change in noise level did not significantly affect the speech utterance production nor the gesture production. It would be interesting to see if, rather than following the surrounding noise level, participants choose a certain communicative strategy (i.e. relying more on gesture, more on speech, or apply a balanced multimodal strategy). If this is indeed the case, and directors tend to stick to the same noise level throughout the experiment, then it would also be interesting to see which strategy directors applied most often, i.e. if the directors had a preference for the speech-only, gesture-only, or equally balanced attempts, and if the noise level they heard initially might have influenced their strategy choice. It could be possible that noise level does play a role, only in a different way than was anticipated for this paper. It thus could be the case that the directors attend to the noise level that they hear, but choose a strategy depending on the noise level they hear in the beginning of the experiment and stick to

it for the duration of the test. If we follow the tradeoff hypothesis, and the first noise level heard by the director does indeed play a role, then it can be expected that the director would choose a speech-based strategy when hearing no noise through the headphones. Upon first hearing the highly noisy 8-talker babble however, the director possibly chooses a gesture-based approach, i.e. producing many gestures and relatively little speech, as the tradeoff hypothesis would predict.

During the coding of the data, only attempt 1 and 2 were segmented and included in the analyses. This was done due to the scope of the paper, and because the initial research questions could be answered with the first two attempts. In future research, it would be interesting to include all the attempts produced by the director. This research could compare the differences between attempt 1 and 2 with those later on in the round. It might be possible, for example, that directors significantly change their gesture and speech rate, as well as their production of gesture feature changes over the course of several attempts, but that we have not found this difference to be significant between the first and second attempt. If the director needs to produce several communicative attempts in order to convey the message successfully to the matcher, he or she might realise the lack of clarity in his current communicative strategy and change it in order to be more informative.

Related to this is the fact that the current paper has not delved into the different attempts divided over the three noise levels. If the highly noisy environment does indeed make it more difficult to communicate, then it could be expected that the director will need to produce more attempts for the matcher to understand which verb is being described. The no-noise condition, on the other hand, would then be expected to cause the least amount of attempts, as the communication between director and matcher is not hindered by (extra) babble noise.

Another suggestion for future research concerns the change in noise levels and its durations. In this experiment, the noise that the director heard though the headphones changed per round. As these rounds did not last long (varying from several minutes to down to only several seconds), it might be the case that the director did not adjust the modality use simply because the noise level did not stay the same for long enough. In a follow-up study, this can be solved by dividing participants in a permanent noise level so that they can perform the experiment in the same noise condition. This would mean that a third of the participants would complete the experiment in the no-noise condition, a third in the moderately noisy environment with the 4-talker babble condition, and a third in the highly noisy 8-talker babble

condition. That way, it can give clearer insight in the gesture and speech productions in different noise levels. The disadvantage to this approach is that it does not test how the director might adjust his communicative strategy (should he or she do so) as a result of the change in noise level. To solve this, another way is to divide the experiment in three parts, in which every part would contain one of the noise levels. Following this method, one noise condition would be attributed to six or seven consecutive rounds: round 1 until 7 would consist of one noise level, then it gets changed in round 8 to the next noise level, and again in round 14, so that round 14 to 20 would contain the last noise level. This set up would allow to not only to study the multimodal communicative strategies that the directors choose, but also to investigate the changes they make when the noise level changes. Contrary to the set-up mentioned above, which is a between-subjects set-up, this one is a within-subject set-up. In sum, we have a large data base at our disposal which allows for more research to be done, which in turn can provide us more insight in modality modulations of producers in noise.

7. Conclusion

In this study we have aimed to investigate the influence of noise level on the production of gesture strokes and speech utterances, as well as the differences in gestures, speech and the gesture feature changes. Based on the results that were gained from experimental data using Dutch action verbs, it can be concluded that neither the gesture strokes nor the speech utterances were influenced by the changing noise levels. These outcomes cause that neither the tradeoff hypothesis not the hand-in-hand hypothesis can be confirmed nor denied. More research is needed to gain more insight in how, and if, directors adjust to their changing surrounding noise level. We have proposed several suggestions for further research, among others a more steady noise level change instead of a different one after every round. This could possibly result in a change in communicative strategy in the directors, as they would have more time to accustom to the new noise level and make changes in their production accordingly.

The third research question revolved around the different attempts that were produced by the directors. It was found that, in comparison to the first attempt, the second attempt consisted of a significant different amount of gesture strokes, gesture feature changes of referent and changes of hand; the former is produced more in the second attempt, the latter two less. The fact that directors produced more gestures in attempt 2 but less of these gestures are characterised by change of referent, implies that the directors either produce other gesture feature changes or produce the exact same gestures. However, none of the other gesture feature changes are produced significantly more often in the second attempt than in the first. Future research could focus on all the attempts produced by the director, instead of just the first two that were the focus of this paper. Coding all the attempts might give more results in the director's adjusted communicative behaviour over time.

Conducting more research on the production side of communication in noisy environments will gain us more insight in the initial and natural strategy the directors apply in order to communicate in a suboptimal environment. Investigating what the best strategy is to communicate in noise can provide answers that can be extended to, for example, communicating with hard of hearing people.

More face-to-face communicative data is needed to give answers to the remaining questions and give us further insight in how production is realised in suboptimal environments, and can provide answers as to how gesture and speech interact with each other in communicative productions.

References

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15, 415-419.

Bates, D., Maechler, M., Bolker, B., Walker, S., Bojesen Christensen, R. H., Singmann, H., … Fox, J. (2019). *Lme4: Linear Mixed-Effects Models using 'Eigen' and S4*. Obtained through the internet: https://cran.r-project.org/web/packages/lme4/.

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to semantic information conveyed by speech? An experimental investigation. *Semiotica*, 123 (1/2), 1-30.

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52, 46-57.

Bolker, B. M., Brooks, M. E., Clarke, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24 (3), 127-135.

Brugman, H., & Russel, A. (2004). *Annotating multimedia / multi-modal resources with ELAN*. Proceedings of the 4th International Conference of Language Resources and Language Evaluation (LREC 2004), 2065-2068.

Campisi, E., & Özyürek, A. (2013). Iconicity as a communicative strategy: Recipient design in multimodal demonstrations for adults and children. *Journal of Pragmatics*, 47 (1), 14-27.

Cassell, J., McNeill, D., & McCullough, K. (1998). Speech – gesture mismatches: Evidence for one underlying representation of linguistic & nonlinguistic information. *Pragmatics & Cognition*, 6 (2), 1-24.

Castellanos, A., Benedi, J. M., & Casacuberta, F. (1996). An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. *Speech Communication*, 20, 23-35.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1-39.

Clayman, S. E. (2013). Turn-constructional units and the transition-relevance place. In J. Sidnell & T. Stivers (Eds.), *The Handbook of Conversation Analysis* (pp. 151-166). London: Blackwell.

Cohen, A. A. (1977). The communicative functions of hand illustrators. *Journal of Communication*, 27, 54-63.

Davis, C., Kim, J., Grauwinkel, K., & Mixdorff, H. (2006). Lombard speech: Auditory (A), visual (V) and AV effects. In R. Hoffmann & H. Mixdorff (Eds.), *Proceedings of the Third International Conference on Speech Prosody* (pp. 248-252). Dresden, Germany: TUD Press.

De Ruiter, J. P. (2006). Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech Language Pathology*, 8, 124-127.

De Ruiter, J. P., Bangerer, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the Tradeoff Hypothesis. *Topics in Cognitive Science*, 4, 232-248.

Dreher, J. J., & O'Neill, J. (1958). Effects of ambient on speaker intelligibility for words and phrases. *The Laryngoscope*, 68, 539-548.

Drijvers, L. (2019). *On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions*. Nijmegen: Ipskamp.

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, 212-222.

Drijvers, L., & Özyürek, A. (2018). Native language status of the listener modulates the neural integration of speech iconic gestures in clear and adverse listening conditions. *Brain and Language*, 177/178, 7-17.

Drijvers, L., Özyürek, A., & Jensen, O. (2018a). Hearing and seeing meaning in noise: Alpha, beta and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human Brain Mapping*, 39 (5), 2075-2087.

Drijvers, L., Özyürek, A., & Jensen, O. (2018b). Alpha and beta oscillations index semantic congruency between speech and gestures in clear and degraded speech. *Journal of Cognitive Neuroscience*, 30 (8), 1086-1097.

Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *The Journal of the Acoustical Society of America*, 70, 45-50.

Feyereisen P. (2006). Further investigation on the mnemonic effect of gestures: Their meaning matters. *European Journal of Cognitive Psychology*, 18 (2), 185-205.

Ford, C. E., & Thompson, S. A. (1996). Interactional units in conversation: Syntactic, intonational and pragmatic resources for the management of turns. In E. A. Schegloff & S. A. Thompson (Eds.), *Interaction and Grammar* (pp. 135-184). Cambridge: Cambridge University Press.

Fry, D. B. (1975). Simple reaction-times to speech and non-speech stimuli. *Cortex*, 11, 355-

360.

Galati, A., & Brennan, S. E. (2014). Speakers adapt gestures to addressee's knowledge: Implications for models of co-speech gesture. Language, *Cognition and Neuroscience*, 29 (4), 435-451.

Garber, S. R., Siegel, G. M., & Pick, H. L. Jr. (1981). Regulation of vocal intensity in the presence of feedback filtering and amplification. *Journal of Speech and Hearing Research*, 24, 104-108.

Garnier, M. (2008). May speech modifications in noise contribute to enhance audio-visible cues to segment perception? In R. Göcke, P. Lucey, & S. Lucey (Eds.), *Proceedings of AVSP '08, the International Conference on Audio-Visual Speech Processing* (pp. 95-100). ISCA Archive.

Garnier, M., Henrich Bernardoni, N., & Dubois, D. (2010). Influence of sound immersion of communicative interaction on the Lombard effect. *Journal of Speech, Language, and Hearing Research*, 53 (3), 588-608.

Geerts, J. (2018). *ELAN – Linguistic annotator*. The Language Archive, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

Goldin-Meadow, S., McNeill, D., & Singleton, J. (1996). Silence is liberating: Removing the handcuffs on grammatical expression in the manual modality. *Psychological Review*, 103 (1), 34-55.

Graham, J. A., & Heywood, S. (1975). The effects of elimination of hand gestures and of verbal codability on speech performance. *European Journal of Social Psychology*, 5, 189-195.

Gross, J., & Ligges, U. (2015). *Nortest: Tests for Normality*. R package version 1.0-4. Retrieved from: https://cran.r-project.org/web/packages/nortest/.

Habets, B., Kita, S., Shao, Z., Özyürek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech - gesture integration during comprehension. *Journal of Cognitive Neuroscience*, 23 (8), 1845-1854.

Hagoort, P. (2013). *MUC (Memory, Unification, Control) and beyond. Frontiers in Psychology* (Vol. 4). Elsevier Inc. https://doi.org/10.3389/fpsyg.2013.00416.

Hilliard, C., & Cook, S. W. (2016). Bridging gaps in common ground: Speakers design their gestures for their listeners. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 42 (1), 91-103.

Hoetjes, M., Krahmer, E., & Swerts, M. (2015). On what happens in gesture when communication is unsuccessful. *Speech Communication*, 72, 160-175.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, 19 (7), 1175-1192.

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, 875-884.

Holler, J., & Beattie, G. (2003). Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener? *Gesture*, 3 (2), 127-154.

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin & Review*, 25 (5), 1900-1908.

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, 692-697.

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.

Holler, H., & Stevens, R. (2007). The effect of common ground on how speakers use gesture and speech to represent size information. *Journal of Language and Social Psychology*, 26 (1), 4-27.

Isaac, E. A., & Clark, H. (1987). Reference in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116 (1), 26-37.

Jabeen, H. (2019). *Tutorial: Poisson Regression in R*. Data Science Tutorials. Retrieved from https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/.

Jassim, F. A. (2013). Image denoising using Interquartile Range filter with local averaging. *International Journal of Soft Computing and Engineering*, 2 (6), 424-428.

Junqua, J. (1993). The Lombard reflex and its role on human listener and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93, 510-524.

Kelly, S., Healey, M., Özyürek, A., & Holler, J. (2015). The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review*, 22 (2), 517-523.

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.

Kendon, A. (1972). Some relationships between body motion and speech. In A. W. Sigman &

B. Pope (Eds.), *Studies in dyadic communication* (pp. 177-216). New York: Pergamon Press.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In: M. R. Key (ed.), *The relation between verbal and nonverbal communication* (pp. 207-227). Mouton, The Hague,.

Kendon, A. (1983). Gesture and speech: How they interact. In J. M. Weimann & R. P. Harrison (Eds.), *Nonverbal interaction* (pp. 13-45). Beverly Hills, CA: Sage.

Kim, S. (2005). Durational characteristics of Korean Lombard speech. In *Proceedings of Eurospeech '05, the 9th European Conference on Speech Communication and Technology* (pp. 2901-2904). ISCA Archive.

Kita, S. (1990). *The temporal relationship between gesture and speech: A study of Japanese – English bilinguals*. Master's thesis, University of Chicago.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Eds.), *Language and gesture* (pp. 261-283). Cambridge: Cambridge University Press.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for and interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48, 16-32.

Kita, S., Van Gijn, I., & Van der Hulst, H. (1997). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and sign language in human – computer interaction* (pp. 23-35). Berlin, Germany: Springer.

Lombard, E. (1911). Le signe de l'élévation de la voix. *Annales des Maladies de L'Oreille et du Larynx*, 37, 101-119.

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, 23, 45-66.

McNeill, D. (1989). A straight path – To where? Reply to Butterwort and Hadar. *Psychological Review*, 96, 175-179.

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

McNeill, D. (2005). *Gesture and thought*. Chicago: Chicago University Press.

McNeill, D., Cassell, J., & McCullough, K. (1994). Communicative effects of speech-mismatched gestures. *Research and Language and Social Interaction*, 27 (3), 223-237.

Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture*, 4, 119-141.

Nobe, S. (2000). Where do most representational gestures actually occur with respect to speech? In D. McNeill (Eds.), *Language and Gesture* (pp. 186-198). Cambridge: Cambridge University Press.

Obermeier, C., Dolk, T., & Gunter, T. C. (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex*, 48, 847-870.

Parrill, F. (2009). Dual viewpoint gestures. *Gesture*, 9 (3), 271-289.

Peeters, D., Chu, M., Holler, J., Hagoort, P., & Özyürek, A. (2015). Electrophysiological and kinematic correlates of communicative intent in the planning and production of pointing gestures and speech. *Journal of Cognitive Neuroscience*, 27 (12), 2352-2368.

Pittman, A. L., & Wiley, T. L. (2001). Recognition of speech produced in noise. *Journal of Speech, Language and Hearing Research*, 44, 487-496.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and SPLUS*. New York: Springer.

Rogers, W. T. (1978). The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances. *Human Communication Research*, 5 (1), 54-62.

Sakamoto, Y., Ishiguro, M., and Kitagawa G. (1986). *Akaike Information Criterion Statistics*. D. Reidel Publishing Company.

So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science*, 33, 115-125.

Stanton, B. J., Jamieson, L. H., & Allen, G. D. (1988). Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions. In *Proceedings of ICASSP '88, the International Conference on Acoustics, Speech and Signal Processing* (pp. 331-333).

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, 100 (26), 10587-10592.

Student Learning Development. (2009). *Measures of variability: The range, inter-quartile range and standard deviationar charts*. University of Leicester. Retrieved from https://www2.le.ac.uk/offices/ld/resources/study-guides-pdfs/numeracy-skills-pdfs/measures-ariability-v0.1.pdf

Trujillo, J. P., Simanova, I., Bekkering, H., & Özyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A Kinect study. *Cognition*, 180, 38-51.

Trujillo, J. P., Simanova, I., Bekkering, H., & Özyürek, A. (2019). The communicative advantage: how kinematic signaling supports semantic comprehension. *Psychological Research*. https://doi.org/10.1007/s00426-019-01198-y.

Tufts, J. B., & Frank, T. (2003). Speech production in noise with and without hearing protection. *Acoustical Society of America*, 114 (2), 1069-1080.

Valbonesi, L., Ansari, R., McNeill, D., Quek, F., Duncan, S., McCullough, K. E., & Bryll, R. (2002). Multimodal signal analysis of prosody and hand motion: Temporal correlation and speech and gestures. In *Proc. Eur. Signal Process. Conf.* (EUSPICO ’02) (pp. 75-78).

Van der Sluis, I., & Kramer, E. (2007). Generating multimodal references. *Discourse Processes*, 44, 145-174.

Van Summers, W., Pisoni, B., Bernacki, H., Pedlow, R., & Stokes, M. (1988). Effect of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84, 917-928.

Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications*. arXiv:1308.5499. [http://arxiv.org/pdf/1308.5499.pdf].

Zappavigna, M., Cléirigh, C., Dwyer, P., & Martin, J. R. (2010). The coupling of gesture and phonology. In M. Bednarek & J. R. Martin (Eds.), *New discourse on language: Functional perspectives on modality, identity, and affiliation* (pp. 219-236). London and New York: Continuun.

Zollinger, S. A., & Brumm, H. (2010). The Lombard effect. *Current Biology*, 21 (16), 614-615.