Research Article

The Effects of Iconic Gestures and Babble Language on Word Intelligibility in Sentence Context

Veerle Wilms,[a] Linda Drijvers,[b] and Susanne Brouwer[a]

[a] Centre for Language Studies, Radboud University, Nijmegen, the Netherlands [b] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

ARTICLE INFO

Article History:
Received July 16, 2021
Revision received December 16, 2021
Accepted January 31, 2022

Editor-in-Chief: Bharath Chandrasekaran
Editor: Stephanie Borrie

https://doi.org/10.1044/2022_JSLHR-21-00387

ABSTRACT

Purpose: This study investigated to what extent iconic co-speech gestures help word intelligibility in sentence context in two different linguistic maskers (native vs. foreign). It was hypothesized that sentence recognition improves with the presence of iconic co-speech gestures and with foreign compared to native babble.

Method: Thirty-two native Dutch participants performed a Dutch word recognition task in context in which they were presented with videos in which an actress uttered short Dutch sentences (e.g., *Ze begint te openen,* "She starts to open"). Participants were presented with a total of six audiovisual conditions: no background noise (i.e., clear condition) without gesture, no background noise with gesture, French babble without gesture, French babble with gesture, Dutch babble without gesture, and Dutch babble with gesture; and they were asked to type down what was said by the Dutch actress. The accurate identification of the action verbs at the end of the target sentences was measured.

Results: The results demonstrated that performance on the task was better in the gesture compared to the nongesture conditions (i.e., gesture enhancement effect). In addition, performance was better in French babble than in Dutch babble.

Conclusions: Listeners benefit from iconic co-speech gestures during communication and from foreign background speech compared to native. These insights into multimodal communication may be valuable to everyone who engages in multimodal communication and especially to a public who often works in public places where competing speech is present in the background.

Everyday listening situations frequently present us with speech embedded in background noise such as the sound of traffic, music, or others conversing. Typically, people are rather successful at filtering the right information out of a message in such adverse listening conditions (i.e., cocktail party effect; Cherry, 1953). When a noisy environment complicates communication, meaningful hand gestures can help a listener understand what is being said (e.g., Drijvers & Özyürek, 2017). For example, when at a busy cafe, it can help to make a drinking gesture along with your order for the bartender to understand what you are asking. It is currently unknown how different linguistic background noises affect word intelligibility in sentences when meaningful gestures are present to aid a listener. The aim of this study is to investigate to what extent word intelligibility in sentence context is affected by meaningful gestures and background speech language. Insights into the effects of iconic gestures and background language on word intelligibility in sentences help develop existing theories on both word intelligibility in sentence context and multimodal communication. The societal implications of this study may be valuable not only to everyone who engages in multimodal communication but also, especially, to a public who often spends time (e.g., for leisure time and/or work) in public places. Specific knowledge on what makes communication easier or more complex and on how to deal with

Correspondence to Veerle Wilms: veerle.wilms@mpi.nl. Linda Drijvers and Susanne Brouwer share senior authorship. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

background noise can educate and help people experience less problems during face-to-face communication.

Enhancement Effects of Visual Speech and Iconic Gestures

During face-to-face communication, a listener is presented with both auditory and visual information. This visual information is, for example, visual speech and/or hand gestures. As these visual articulators are present in real-life communication, it is important to study to what extent they contribute to speech comprehension. Previous research on the effects of visual speech has established that visual speech enhances performance on speech recognition tasks. That is, participants perform better on speech recognition in noise (e.g., Ross et al., 2007; Sumby & Pollack, 1954) or in multitalker babble (e.g., Holle et al., 2010; Sommers et al., 2005; Stevenson et al., 2015; Tye-Murray et al., 2007, 2010) when the face of the speaker is visible, which allows for lipreading (also referred to as speechreading, e.g., Summerfield, 1992).

Apart from visual speech, it has also been demonstrated that iconic gestures, on top of visual speech, help word intelligibility by providing semantic cues that disambiguate degraded speech (e.g., Drijvers & Özyürek, 2017, 2020; Drijvers et al., 2018; for a review on how gestures play a role in the comprehension of speech, see Özyürek, 2014). Iconic gestures are hand movements that depict object attributes, actions, and space (e.g., Clark, 1996; Goldin-Meadow, 2005; McNeill, 1992) and are semantically related to speech because of the similarities to the objects, events, and special relations that they represent (such as a drinking gesture toward your mouth when you want to order a drink). For example, Drijvers and Özyürek (2017) performed a word comprehension study in which participants identified spoken action verbs in three different speech conditions (two-band noise vocoding, six-band noise vocoding, and clear), in three different multimodal conditions (speech and lips blurred; speech and visible speech; and speech, visible speech, and gesture), and finally in two visual-only conditions (visible speech, visible speech and gesture). Participants performed best on the word identification task when both visual speech and iconic gestures were presented in a joint context compared to one or none. Additionally, that benefit was larger at six-band than at two-band noise vocoding (for a follow-up study with nonnative listeners, see Drijvers & Özyürek, 2020). It is thus beneficial for listeners when both phonological cues from visible speech and semantic cues from iconic gestures are present to disambiguate degraded speech.

Apart from noise-vocoded speech, this enhancement effect of iconic gestures has also been found in the presence of a linguistic masker: Participants performed better on a word comprehension task in multitalker babble when they were able to see both visual speech and iconic gestures (Schubotz et al., 2020). Schubotz et al. performed a Dutch word recognition study and presented Dutch participants with the same task and stimuli as used in the study of Drijvers and Özyürek (2017) but with Dutch babble as a linguistic masker instead of noise-vocoded speech. Participants performed better on the task when both visual articulators (visual speech *and* iconic co-speech gestures) were present compared to one or none. The study by Schubotz et al. only used babble in one language (Dutch), but it is known from speech-in-speech research that different kinds of babble (i.e., different languages) affect speech intelligibility differently (e.g., Brouwer et al., 2012; Brungart, 2001; Garcia Lecumberri & Cooke, 2006; Van Engen, 2010; Van Engen & Bradlow, 2007). Speech intelligibility becomes poorer when the babble and the target language are linguistically similar and when a listener is familiar with the babble. Both linguistic maskers and iconic gestures are often part of real-life communication, and it is therefore important to study if and, if so, how much iconic gestures can aid listeners in such communicative situations. Taken together, previous studies suggest that iconic gestures enhance word intelligibility (Drijvers & Özyürek, 2017; Drijvers et al., 2018; Schubotz et al., 2020), but how great this enhancement effect is in sentence context and in different background languages has yet to be investigated.

Informational Masking Effects on Sentence Intelligibility

Not only gestures affect speech comprehension but background noise (i.e., a masker) does so as well. Within the field of speech-in-noise research, a distinction is often made between energetic and informational masking (e.g., Carhart et al., 1969; Darwin, 2008; Kidd et al., 2007; Pollack, 1975). Energetic masking is also referred to as peripheral masking, and it happens when noise physically interferes with target speech. For example, white noise, an energetic masker, has been found to negatively affect target sentence recognition when it is present as a masker in the background (e.g., Freyman et al., 1999). Background speech, for example, babble (i.e., more than one interfering talker), additionally causes informational masking, which happens when noise perceptually interferes with target speech. Informational masking is considered to be a higher level of masking and goes beyond the peripheral masking that energetic masking causes. Consequences of informational masking are incomplete linguistic separation between target and masker speech, captured attention by the masker speech, semantic interference, and associated cognitive load (Mattys et al., 2012).

Studies on the informational masking effects on speech-in-speech research have generally demonstrated that greater similarity between target and masker speech causes greater interference (e.g., Brouwer et al., 2012; Van Engen, 2010). This finding supports the Target–Masker Linguistic Similarity hypothesis (henceforth, TMLS hypothesis; Brouwer et al., 2012), which assumes that more (linguistic) similarity between target and masker speech and greater familiarity with the masker causes greater interference on speech intelligibility. There are several studies that have demonstrated evidence in favor of this hypothesis. For example, sentence intelligibility increases when the sex of the target and that of the masker speaker differ (Brungart et al., 2001; Williams & Viswanathan, 2020) or when the masker and the target are different languages compared to the same (e.g., Van Engen & Bradlow, 2007).

Similar effects have been found for target and masker speech language. Garcia Lecumberri and Cooke (2006) demonstrated that native English listeners (who were unfamiliar with Spanish) performed poorer on an English consonant identification task in an English masker than in a Spanish masker. Van Engen and Bradlow (2007) found a similar effect for sentence recognition instead of consonant recognition. Their English participants (to whom Mandarin was unfamiliar) were more negatively affected by English two-talker babble than by Mandarin two-talker babble. The authors explain that linguistic interference was the cause of the informational masking effect (see, e.g., Summers & Roberts, 2020, for more on linguistic contributors to informational masking, and Brouwer, 2017, for a similar linguistic similarity effect for dialect).

Besides target–masker similarity, the familiarity of the listener with the masker speech also affects sentence intelligibility. Van Engen (2010) performed an English sentence recognition study with native speakers of English and nonnative speakers of English whose first language was Mandarin. Performance on the recognition task was poorer for both participant groups in English than in the Mandarin babble, but the native speakers of Mandarin experienced a smaller release from masking in the Mandarin babble compared to the English babble. This study demonstrated that both the similarity between the target and masker speech (*similarity effect*) and the language familiarity of the listener (*familiarity effect*) influenced the intelligibility of target speech. Additionally, a study by Brouwer et al. (2012) demonstrated this familiarity effect with semantically meaningful and anomalous maskers. Brouwer et al. not only replicated the results by Van Engen but also demonstrated that semantically meaningful maskers decreased performance on a speech recognition task. Participants performed poorer on sentence recognition when the target speech was similar to the masker speech (i.e., two-talker babble), when they were familiar with the masker language, and when the masker was semantically meaningful babble compared to anomalous sentences. Participants thus performed better when they were unfamiliar with the masker speech. (Note that the studies by Calandruccio et al., 2018, Calandruccio & Zhou, 2014, and Tun et al., 2002, provided evidence against the hypothesis by Brouwer et al., 2012. The design of this study does not challenge these contradicting findings.)

In summary, previous studies have demonstrated that sentence intelligibility increases when target and masker speech are dissimilar (TMLS hypothesis; Brouwer et al., 2012). Note, however, that these studies were all unimodal: They focused on auditory stimuli only. This study fills this gap by studying speech recognition in different linguistic maskers, with gestures accompanying the target speech to aid listeners.

This Study

The aim of this study was to examine to what extent iconic co-speech gestures help Dutch participants with Dutch word recognition in context in both Dutch (native) and French (foreign) two-talker babble. As iconic hand gestures can aid listeners, their presence could help listeners overcome the obstacles that linguistic background noises pose during listening. Additionally, the previously discussed studies on gestural enhancement have focused on isolated target words (note that, e.g., Garcia & Dagenais, 1998, and Hustad & Garcia, 2005, studied whether sentence intelligibility of individuals with dysarthria improved with the use of iconic gestures, but background noise was not investigated in these studies), even though the production of gestures is often done (and thus comes more naturally) when uttering whole sentences. This study is the first to study the enhancing effects of iconic gestures and linguistic maskers in sentence context. In a word recognition task in sentence context, Dutch participants were presented with videos in which a Dutch actress uttered Dutch target sentences (e.g., *Ze begint te openen,* “She starts to open”). These target sentences were presented in six different conditions: no background noise (i.e., clear condition) without gesture, no background noise with gesture, French babble without gesture, French babble with gesture, Dutch babble without gesture, and Dutch babble with gesture. The iconic co-speech gestures made by the actress coincided with the pronunciation of the last word of the target sentence, which was always an action verb. The accurate identification of these action verbs was measured.

Several hypotheses can be proposed based on the literature previously discussed. First, it is predicted that participants perform better in the gesture conditions compared to the conditions without gesture (e.g., Drijvers & Özyürek, 2017; Schubotz et al., 2020). Second, it is expected that participants perform poorer in the Dutch than in the French babble because the Dutch targets and Dutch babble share great linguistic overlap and because participants are more familiar with Dutch than French (e.g., Brouwer et al., 2012; Van Engen, 2010). The current experiment therefore explores the TMLS hypothesis (Brouwer et al., 2012) in a context where the target speech is accompanied by iconic gestures. Finally, an interaction between gesture and babble on intelligibility might occur, and it can have two different outcomes. Iconic gestures could have the strongest positive effect in the most difficult condition (Dutch-in-Dutch) because participants will need the semantic information from iconic gestures most in this condition. In contrast, Drijvers and Özyürek (2017) showed that iconic gestures have a most optimal effect at an intermediate level of difficulty (Dutch-in-French in the current experiment). That is, iconic gestures cannot help when the target speech is too difficult to perceive. This study could show if the amount of help that an iconic co-speech gesture provides depends on the linguistic masker.

Method

Two pretests were conducted prior to the experiment. The pretests were necessary to create a suitable stimuli set and an appropriate signal-to-noise ratio (henceforth, SNR) level for the experiment.

The Method section is built up in the following way: first, a description of the materials used for both pretests and the experiment; second, full descriptions of the two pretests; and followed by participant, procedure, design, and coding information about the experiment. The materials, data, and analysis script can be retrieved from https://osf.io/29bmz/.

Materials

Target Sentences

The 180 Dutch target sentences used in the experiment were short, consisting of a minimum of four words and a maximum of six words, to avoid testing the participants' working memory instead of the intelligibility of the sentences (e.g., *Ze begint te openen,* "She starts to open"). Half of the target sentences started with a masculine pronoun (*hij,* "he"), and the other half started with a feminine pronoun (*zij,* "she"). All sentences followed this structure: subject–finite verb–(negation/preposition/adverb)–infinitive. Several finite verbs were used to avoid repetition (i.e., if all sentences had the same subject, had a finite verb, and only differed in the infinitive, the carrier phrase is more likely to be ignored during the experiment). No finite verb form was used more than 5 times. Every target sentence ended in a frequently occurring Dutch infinitive, which was always an action verb (e.g., *openen,* "to open"; *eten,* "to eat"; and *rollen,* "to roll"). The focus of this study was on these action verbs because many action verbs are easily accompanied by an iconic gesture. The choice was made to present all the action verbs in a sentence context instead of in isolation to increase ecological validity of the target items with regard to everyday face-to-face communication. All target sentences (and their English translations) can be found in the Appendix.

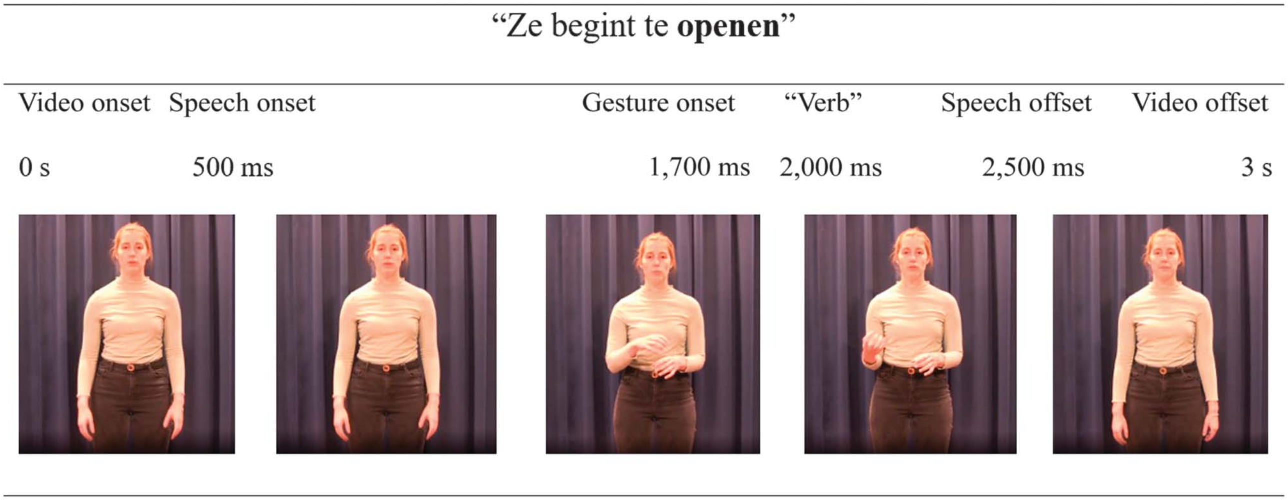
Video Stimuli

All videos were recorded with a Canon XF105 camera at the Centre for Language Studies lab at Radboud University Nijmegen. A schematic overview of a video stimulus can be found in Figure 1. The videos displayed a female 24-year-old native speaker of Dutch[1] who uttered short target sentences, with or without a gesture. The actress in the videos wore neutral light blue clothing in front of a neutral blue background, displayed from the head to just above the knee in the middle of the screen; her hair was tied up away from the face; and her starting position was the same for every video (standing straight, looking into the camera, and her arms loosely hanging on the sides of her body). The actress was instructed to utter the sentences at a comfortable but calm pace, to make a hand gesture at the infinitive in each sentence, not to make any facial expressions such as eyebrow raises or smiles, not to make any other bodily movements but the hand gestures, and to blink as little as possible. The actress was not instructed on how to make any of the hand gestures to prevent any of the gestures from looking unnatural. A pretest was conducted to determine the right selection of iconic gestures for the experiment (see Gesture Pretest section below).

The video footage was annotated in ELAN (Version 5.9; Wittenburg et al., 2006). For the nongesture videos, speech started 500 ms after video onset, and the videos ended 500 ms after the speech ended (following Brouwer & Bradlow, 2015). For the gesture videos, speech started 500 ms after video onset, and the videos ended after full execution of the gesture (when the hands were back at their starting position). The gestures made by the actress always started approximately 300 ms before pronunciation of the action verb with the stroke of the gesture (i.e., most meaningful component of the gesture) coinciding with the pronunciation of the verb. The duration of the videos was between 2 and 3 s, depending on the length of the verb and the time it took to place both hands back into starting position.

[1] The actress has given permission to use images of her in Figures 1 and 3.

Figure 1. Schematic overview of a video stimulus. English translation of the uttered sentence, "She starts to open."



Babble

The two-talker babble for this study was created by two 23-year-old female speakers of Dutch who are also highly proficient in French (one speaker was raised as a Dutch–French bilingual so French is her first language, and the second speaker had near-native proficiency in French due to working in France for several years). The beginning three pages of the novel *Du côte de chez Swann* by Marcel Proust were used for the babble, in both the French original and the official Dutch translation of the book.[2] This novel was chosen for two reasons. First, the Dutch text is a direct translation of the French text, which causes the differences between the two texts to be minimal. Second, this novel has rather long sentences and difficult terms in both the French and Dutch versions, which elicits monotonous reading behavior. This is favorable because the babble should not include any rapid changes in tempo, pitch, or volume (as was also the case for the target sentences).

Both the Dutch and French excerpts were about 6 min long to ensure a unique piece of babble for each target sentence in the experiment. The recordings by the speakers were made from their homes with their mobile phones (due to the COVID-19 pandemic). The speakers received both the Dutch and French texts, as well as written instructions. They were instructed to be in a quiet environment, record with their mobile phone, practice the stretches of text before recording them, use as little intonation as possible, start a sentence over when they made an error, and read at a pace appropriate for reading a book aloud. The recordings were checked to determine that they were of good quality and that there was no noise or distracting sounds.

Audacity® was used to delete large stretches of silence and reading mistakes in the recordings and to overlay the speakers to create a two-talker babble (following, e.g., Brouwer et al., 2012; Van Engen, 2010; Van Engen & Bradlow, 2007). The speakers sometimes made small mistakes or took some time to swallow or prepare for a next sentence. These instances of swallowing or long periods of silence were manually removed from the recordings in Audacity. This was done to make the recording sound like natural and comfortable reading aloud. Praat (Boersma, 2001) was used to normalize the long-term average speech spectra of the babble tracks and to set the appropriate dB levels. The target stimuli were normalized at a sentence level to have the same SNR (following, e.g., Brouwer et al., 2012; Calandruccio et al., 2018; Williams & Viswanathan, 2020). The intensity of the target sentences was set at 65 dB and piloting (see SNR and Gestural Enhancement Pretest section below) determined setting the intensity of the babble to 75 dB to create an SNR of −10 dB. To combine the babble and the target sentences, the audio of the videos was stripped off the videos (using VLC media player; VideoLan, 2006), to later match the videos with the corresponding audio again. The following two sections describe the two pretests that were conducted to create a suitable selection of gestures for the experiment (Gesture Pretest section) and to determine the appropriate SNR level for the experiment (SNR and Gestural Enhancement Pretest section).

Gesture Pretest

The aim of the gesture pretest was to determine (a) whether the action verbs in sentence context could be

[2] For the French edition of Marcel Proust's "Swann's Way," we used Du côte de chez Swann (Proust, 1954). For the Dutch edition we used the translation by Thérèse Cornips and Anneke Brassinga (Proust, 2019).

disambiguated by the recorded accompanying iconic gestures and (b) whether any of the recorded gestures could be considered pantomimes. It was examined whether the gestures made by the actress indeed matched the action verbs. Participants were presented with the gesture videos without audio and were asked how well the gestures matched the action verbs. The results of this pretest determined which action verbs were suitable for the experiment.

Participants. Twenty-one native speakers of Dutch participated voluntarily, but two participants were excluded from the data due to having hearing problems (n = 1) or later admitting to not having read the instructions of the task (n = 1), which resulted in a total of 19 participants (12 female; M_{age} = 26.1, SD = 9.52). All 19 participants were between the ages of 18 and 40 years and reported no visual or hearing impairments or any language-related or neurological disorders. None of these participants participated in the SNR and gestural enhancement pretest or in the experiment.

Materials. A total of 185 action verbs were recorded with and without gesture, with the exception of 10 verbs that did not have a suitable accompanying gesture (e.g., "to pole dance/pole dancing" is not easily captured in an iconic co-speech gesture produced with hands only). This pretest only focused on the iconic gestures and thus included 175 videos without audio.

Procedure. Participants were presented with 175 gesture video stimuli without any audio. The pretest was designed using Qualtrics software (https://www.qualtrics.com), which allowed participants to participate from their homes. Note that previous work has shown similar findings on an off-line compared to an online speech-in-noise task (e.g., Brouwer et al., 2021). Due to the COVID-19 pandemic, this pretest, as well as the SNR and gestural enhancement pretest and experiment, was fully conducted online. All participants were instructed to perform the pretest on a computer or laptop (no tablets or mobile phones were allowed due to the size of the video stimuli on the screen).

First, participants signed a form of consent and answered prescreening questions regarding age, gender, education, and visual and hearing impairments. After the prescreening questions, participants received written instructions on the screen. They were instructed to answer two questions after each video: first, what verb(s) they associated with the gesture in the video (minimum of one answer; maximum of two answers), and second, a 7-point Likert scale on which they had to indicate how well the verb that was matched with the gesture fit the gesture in the video (ranging from 1 = *does not fit the gesture very well* to 7 = *fits the gesture very well*). The first question followed the video immediately, but the second question was presented on the next screen to encourage the participants to answer instinctively. The 175 stimuli were pseudorandomized into four blocks of 44 or 43 videos each and were randomized within each block. Participants were given the opportunity to take a break after each block. To avoid dropouts as much as possible, the participants were shown how far along they were in the pretest at each break. The whole pretest was self-paced, and the length of the breaks was determined by the participants themselves. Every video could only be viewed once. The pretest took 45–60 min to complete.

Data coding and results. The typed answers to Question 1 ("Which verb(s) do you associate with the hand movement in the video?") were used to determine which of the action verbs were pantomimes rather than iconic gestures. Pantomimes can usually be understood without accompanying speech and are thus not optimal for the purpose of this study. Answers were coded "correct" when the correct verb was given between the maximum of two answers, when minor spelling mistakes were made that did not alter the meaning of the verb (e.g., *fotograveren* instead of *fotograferen* "to take a picture"), when the answer given was a perfect synonym of the correct verb (e.g., *hardlopen* instead of *joggen,* "to jog/run"), or when a preposition was added to the verb that did not alter the meaning of the verb (e.g., *uitwringen* instead of *wringen,* "to wring out"). The results showed a mean accuracy rate of 31% (SD = 0.46) over all gesture videos. There were four gestures in total with an accuracy rate of 100%: *vliegen* ("to fly"), *trekken* ("to pull"), *praten* ("to talk"), and *bellen* ("to call"). These gestures were removed from the data set for the experiment because the perfect accuracy rate suggests that these gestures behaved like pantomimes.

The answers to the second question ("How well do you think that our matched verb fits the hand movement in the video?") were used to find out if the gestures depicted the verbs that they were initially assigned to. The average Likert score for every item was calculated. The average score over all gestures was 5.19 (SD = 1.40). Only the gestures with a mean Likert score of 5 or higher fit the hand movement well enough to be used in the experiment (following Drijvers & Özyürek, 2017). This criterion, together with the criterion from Question 1 on accuracy, resulted in a total of 96 gestures (M_{Likert} = 6.04, SD = 1.18) suitable for the experiment.

SNR and Gestural Enhancement Pretest

The aim of the SNR and gestural enhancement pretest was to (a) establish that the iconic co-speech gestures caused gestural enhancement and (b) determine the optimal SNR level to be used in the experiment. It had to be made sure that the experiment would elicit neither floor or ceiling effects. Drijvers and Özyürek (2017) demonstrated that iconic co-speech gestures can lose their disambiguating abilities when a speech recognition task is too difficult (two-band noise vocoding in their study).

Participants. Ten native Dutch participants (eight female; $M_{\text{age}} = 25$, $SD = 4.59$) took part in the second pretest. All participants were between the ages of 18 and 40 years and reported no visual or hearing impairments or any language-related or neurological disorders. None of these participants participated in the gesture pretest or in the experiment. Participants self-rated their French listening skills to be 2.4, on average ($SD = 0.70$), on a scale from 1 (*no knowledge*) to 5 (*as my native language*). One participant indicated that he/she actively used French for roughly 1 hr per week in everyday life. Participants were recruited through the participant pool of the Max Planck Institute for Psycholinguistics and received €8 for participation.

Materials. The participants were presented with 144 audiovisual stimuli (12 items per condition). A 3 × 2 × 2 design presented the participants with 12 conditions: three different SNR levels; gesture vs. no gesture videos; and French vs. Dutch babble (within-subject design). The target sentences were set at 65 dB SPL (i.e., sound pressure level) and the two-talker babble at 72 dB, 75 dB, and 77 dB SPL, which created three different SNR levels of −7, −10, and −12 dB.

Procedure. The pretest was designed using Qualtrics software (https://www.qualtrics.com), and it was fully conducted online. As was the case for the gesture pretest, the participants were instructed to perform the pretest on a computer or laptop. Participants signed a form of consent and answered prescreening questions regarding age, gender, education, and visual and hearing impairments. Participants self-reported that they had ear pods/headphones at hand and that they were in a quiet environment. After the prescreening questions, participants received written instructions on the screen. They were instructed to answer two questions after each video: first, how comprehensible the signal was on a 4-point scale (ranging from 1 = *not comprehensible at all* to 4 = *perfectly comprehensible*; following the methodology by Drijvers & Özyürek, 2017), and second, what was said by the actress in the videos. Participants typed down their answer for the second question in an assigned box.

The 144 experimental items were preceded by eight (four gesture vs. four no gesture; four French babble vs. four Dutch babble) practice trials to familiarize participants with the task and the voice of the actress. The practice trials were presented at an SNR of +3 dB to clearly demonstrate that the focus of the questions was on the actress in the videos instead of on the babble in the background. Participants were instructed to use the practice trials to set their volume on a comfortable level. Volume level was not to be adjusted anymore during the experimental trials. The pretest was self-paced entirely so that participants could take short breaks whenever they desired to do so. To avoid dropouts as much as possible, the videos were numbered from 1 to 144. All experimental items were randomized. Every video could only be viewed once.

The experimental trials were followed by questions regarding Dutch and French proficiency in writing, speaking, listening, and reading. Participants subjectively rated their own proficiency levels on a 5-point scale (ranging from 1 = *no knowledge* to 5 = *as my native language*). Lastly, participants were asked to indicate approximately how many hours per week they were actively engaged with the French language. The pretest took approximately 45 min to complete.

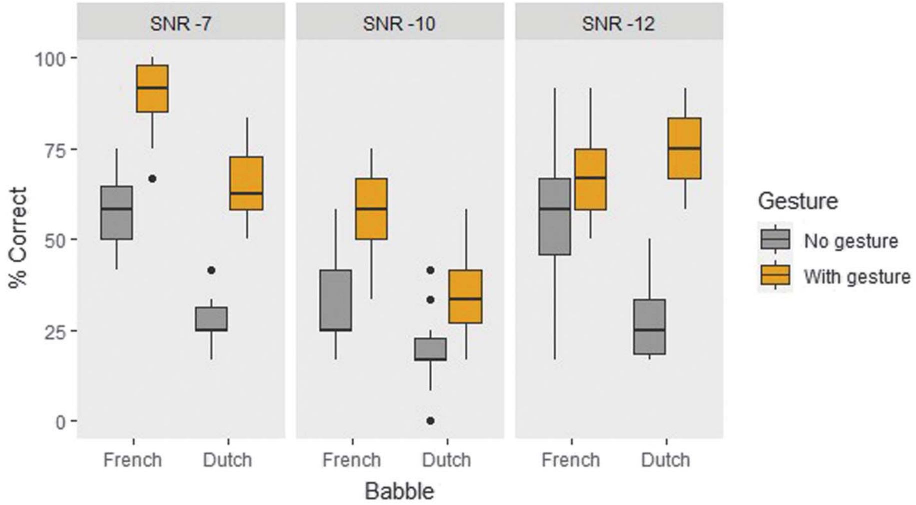
Data coding and results. The answers to the first question ("How comprehensible was the speech signal of the female speaker in the video?") were used to find out how comprehensible the actress was in the three SNR levels of babble. The average Likert score for every condition was calculated. Comprehensibility was, on average, rated 2.35 ($SD = 0.12$) for the easy condition (SNR −7 dB), 1.86 ($SD = 0.19$) for the medium condition (SNR −10 dB), and 2.38 ($SD = 0.13$) for the hard condition (SNR −12 dB).

The typed-down answers for the second question ("What was said by the female speaker in the video?") were coded "correct" when the correct verb was given, when minor spelling mistakes were made that did not alter the meaning of the verb (e.g., *dooden* instead of *doden,* "to kill"), when the answer given was a perfect synonym of the correct verb (e.g., *frutselen* instead of *friemelen,* "to fiddle"), or when a preposition was added to the verb that did not alter the meaning of the verb (e.g., *verbinden* instead of *binden,* "to join (together)").

Figure 2 below shows a boxplot of the mean accuracy percentages for all conditions. Accuracy was, on average, 59.58% ($SD = 24.79$) for the easy condition (SNR −7 dB), 36.25% ($SD = 18.25$) for the medium condition (SNR −10 dB), and 56.46% ($SD = 23.15$) for the hard condition (SNR −12 dB). Visual inspection of the boxplot suggests that iconic co-speech gestures enhance word intelligibility in sentence context at all SNR levels. Furthermore, Figure 2 demonstrates that the masking effects of French babble seem to be smaller than those of Dutch babble.

To establish whether the iconic co-speech gestures indeed caused gestural enhancement (Aim 1 of this pretest), the accuracy data of this pretest were analyzed in RStudio (Version 1.3.1073; RStudio Team, 2020) using the glmer function from the lme4 package (Bates et al., 2015). Mixed-effects logistic regression analyses were conducted, with keyword identification accuracy as the dichotomous dependent variable (1 = *correct*, 0 = *incorrect*). A logistic linking function was used to deal with the categorical nature of the dependent variable. Gesture, babble, and SNR were entered as categorical fixed effects, and gesture and babble were coded as numeric contrasts (gesture: without gesture as −0.5 and with gesture as +0.5;

Figure 2. Boxplot of the results of the signal-to-noise ratio (SNR) and gestural enhancement pretest. Boxplots show the interquartile ranges of accuracy scores (in %) on Dutch target word intelligibility in context for the two babble conditions (French babble and Dutch babble). The panel on the left presents the results for SNR level −7 dB, the panel in the middle for SNR −10 dB, and the right panel for SNR −12 dB. The gray boxes present the results for the conditions without gesture, and the yellow boxes present the gesture conditions. Whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range of the box.



babble: French as −0.5 and Dutch as +0.5). Participants and items were entered as random effects. The number of iterations was increased to 100,000 using a BOBYQA optimizer to solve the issue of nonconvergence (Powell, 2009).

The analysis demonstrated a main effect of gesture (β = 3.07, *SE* = 0.73, *z* value = 4.21, $p < .001$), which indicates that participants performed better in the gesture conditions than in the nongesture conditions. Furthermore, there was a main effect of babble (β = −2.31, *SE* = 0.72, *z* value = −3.20, $p < .01$), which suggests that performance decreased in Dutch babble compared to French babble. Finally, performance on SNR −10 dB was poorer than performance on SNR −7 dB (β = −1.61, *SE* = 0.49, *z* value = −3.28, $p < .01$) and SNR −12 dB (β = 1.32, *SE* = 0.48, *z* value = 2.76, $p < .01$). There were no significant interaction effects (all *p*s > .1).

These results show a gestural enhancement effect and, thus, positively confirm Aim 1 of this pretest. Aim 2 was to determine the optimal SNR level to be used for the experiment. All three SNR conditions elicited an adequate reduction in speech intelligibility. Therefore, the middle SNR level of −10 dB was selected for the experiment.

Participants

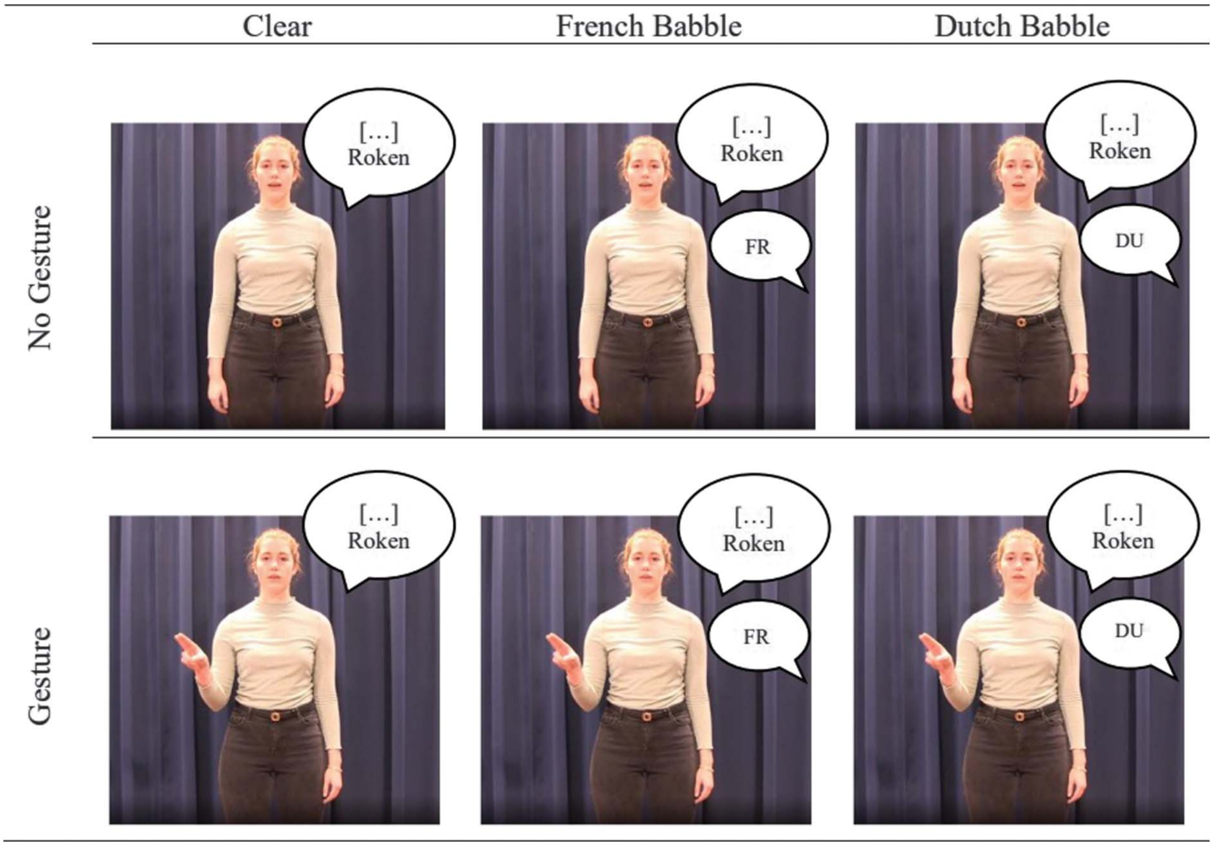
Thirty-seven native Dutch participants took part in the experiment, but five were excluded because their accuracy rate in the masker conditions was less than 10% correct (one participant did not fill out any answers, one participant did not understand the task and transcribed the multitalker babble, and the other three participants scored, on average, 5%, 3.33%, and 6.65% correct in the babble conditions). This resulted in a total of 32 participants (22 female, one other; M_{age} = 25.84, *SD* = 4.89). All participants filled out prescreening questions on their gender, age, first language, education level, and problems with their vision or hearing. Participants could not take part if their age was over 45 years, if their first language was not Dutch, or if they had vision or hearing problems. None of these participants took part in the gesture or SNR and gestural enhancement pretest. Participants self-rated their French listening skills to be 2.03, on average (*SD* = 0.59), on a scale from 1 (*no knowledge*) to 5 (*as my native language*). Three participants indicated that they actively used French in their daily life for 1–3 hr a week.

The study was approved by the Faculty Ethics Committee of the Radboud University Nijmegen (approval code: ECSW-2020-049). Participants were recruited through the participant pool of the Max Planck Institute for Psycholinguistics and received €10 for participation.

Procedure and Design

For the experiment, participants were presented with 180 short videos (30 items per condition; sentences were randomly selected for each condition; iconicity ratings of the gestures did not differ per condition, *F*(2, 87) = 0.746, *p* = .477). A 2 × 3 design presented the participants with six conditions: no background noise (i.e., clear condition) without gesture, no background noise with gesture, French

Figure 3. Overview of conditions of the experiment. The larger text boxes indicate what is said by the actress. The beginning of the target sentence is omitted and replaced by [. . .]. The uttered verb by the actress is *roken* ("to smoke"). The smaller text boxes indicate the babble in the background. French babble is abbreviated to "FR" and Dutch babble to "DU."



babble without gesture, French babble with gesture, Dutch babble without gesture, and Dutch babble with gesture. An overview of these six conditions can be found in Figure 3.

As was the case for the two pretests, the experiment was designed using Qualtrics software (https://www.qualtrics.com), and it was conducted online. Participants performed the experiment on a computer or laptop and signed a consent form before they answered the prescreening questions (same as for the pretests). Participants self-reported that they had ear pods/headphones at hand and that they were in a quiet environment. Written instructions were displayed on the screen. Participants were instructed to answer the following question after each video: "What was said by the actress in the video?" Answers could be typed down in an assigned box.

The 180 experimental trials were preceded by five practice trials (one clear background, two French babble, two Dutch babble, three with gesture, and two without gesture) to familiarize participants with the task and the voice of the actress. The practice trials were presented at an SNR of +3 dB to clearly demonstrate that the focus was on the voice of the actress in the videos. Participants were instructed to use the practice trials to set their volume on a comfortable level. Volume level was not to be adjusted anymore during the experimental trials. The experiment was self-paced entirely; hence, participants initiated every trial themselves and could take short breaks whenever they desired to do so. To avoid dropouts as much as possible, the videos were numbered from 1 to 180. All experimental items were randomized (i.e., randomization was done in such a way that every participant was automatically presented with a different order of trials). Note that the randomization causes the babble in the background to be randomly ordered as well. Every video could only be viewed once.

Questions regarding Dutch and French proficiency in writing, speaking, listening, and reading (on a 5-point scale ranging from 1 = *no knowledge* to 5 = *as my native language*) followed after the experiment. Lastly, participants were asked to indicate approximately how many hours per week they were actively engaged with the French language. The experiment took approximately 60 min to complete.

Data Coding

The typed-down answers (to "What was said by the actress in the video?") were coded "correct" when the correct verb was given or when minor spelling mistakes were made that did not alter the meaning of the verb (e.g., *gietem* instead of *gieten,* "to pour"). Participants who scored less than 10% correct (roughly 1.5 *SD*s below average) in the four masker conditions were excluded from the data set because they might not have understood the task correctly or might have not taken the experiment seriously.

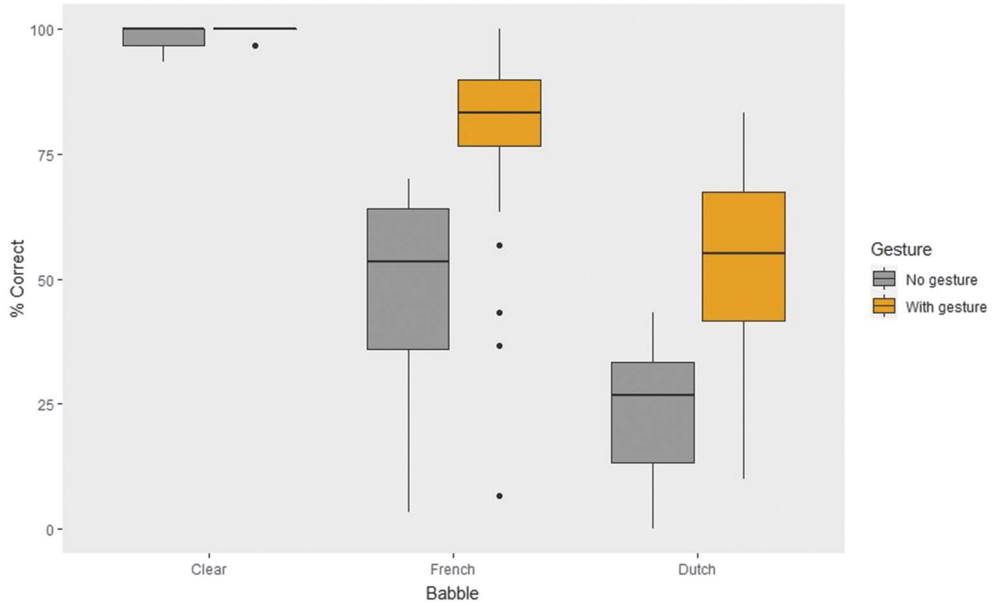
Results

Figure 4 below shows a boxplot of the mean accuracy percentages for all conditions. Accuracy for the clear condition without gesture was, on average, 98.54% (*SD* = 1.88) and 99.90% (*SD* = 0.59) with gesture. Accuracy for the French babble condition, on average, was 46.46% (*SD* = 20.88) without gesture and 79.27% (*SD* = 19.63) with gesture. Finally, accuracy for the Dutch babble condition, on average, was 24.48% (*SD* = 12.23) without gesture and 52.29% (*SD* = 18.63) with gesture.

The data were analyzed in RStudio (Version 1.3.1073; RStudio Team, 2020) using the glmer function from the lme4 package (Bates et al., 2015). Mixed-effects logistic regression analyses were conducted, with keyword identification accuracy as the dichotomous-dependent variable (1 = *correct*, 0 = *incorrect*). A logistic linking function was used to deal with the categorical nature of the dependent variable. Gesture and babble were entered as categorical fixed effects. Akaike information criterion model selection was used to determine that the most parsimonious model included participants and items as random effects and babble by participants as random slope. Helmert coding was used to set up two contrasts for the background conditions. Contrast 1 contrasted the two babble conditions (each coded as 1/3) with the clear condition (coded as −2/3). Contrast 2 contrasted the two babble conditions (French coded as −1/2 and Dutch coded as 1/2) without the clear condition (coded as 0). Contrast 2 was set up to directly compare the performance in Dutch and French babbles. The number of iterations was increased to 100,000 using a BOBYQA optimizer to solve the issue of nonconvergence (Powell, 2009). These models and the comparisons, as well as the data file, can be found in the analysis script on the Open Science Framework site (see https://osf.io/29bmz/).

The analysis revealed a main effect of gesture (β = 2.57, *SE* = 1.18, *z* value = 2.18, *p* = .03), which indicates that performance with an iconic gesture was better than without an iconic gesture. In addition, the analysis showed an effect of Contrast 1 ($\beta_{\mathrm{CLEARvsBABBLE}}$ = −5.94, *SE* =

Figure 4. Boxplot of the results of the experiment. Boxplot shows the interquartile ranges of accuracy scores (in %) on Dutch target word intelligibility in context for all three babble conditions (clear, French babble, and Dutch babble). The gray boxes present the results for the conditions without gesture, and the yellow boxes present the gesture conditions. Whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range of the box.



0.67, z value = −8.81, $p < .001$), indicating that word intelligibility in sentence context was better without babble compared to in babble. Finally, the analysis demonstrated an effect of Contrast 2 ($\beta_{\text{FRENCHvsDUTCH}}$ = −7.56, SE = 0.66, z value = −11.38, $p < .001$), which shows that word intelligibility in sentence context was better in French babble than in Dutch babble. The analysis did not reveal any interaction effects between gesture and babble (all ps > .1).

Discussion

The aim of this study was to answer the following research question: To what extent do iconic co-speech gestures help word intelligibility in sentence context in both native compared to foreign babble? This study was novel in two aspects: It presented target words in sentence context instead of isolated in a study about gestures and their effects on speech intelligibility, and it studied the effect of different linguistic maskers with the presence of iconic co-speech gestures. Dutch participants took part in a Dutch sentence recognition task and were presented with audiovisual stimuli with two gesture conditions (no gestures vs. gestures) and three masker conditions (clear vs. French babble vs. Dutch babble).

In line with previous research, it was hypothesized that performance would improve in the gesture conditions (Drijvers & Özyürek, 2017, 2020; Schubotz et al., 2020). Second, poorer performance was expected in Dutch babble compared to in French babble (following the TMLS hypothesis; Brouwer et al., 2012). This is the first study to investigate the effects of gesture in linguistic maskers; hence, a possible interaction between gesture and babble can have two outcomes. Iconic gestures could have the strongest positive effect in the most difficult condition because the semantic cues from iconic gestures are most needed in this condition. Contrastingly, iconic gestures could have the most optimal effect at an intermediate level of difficulty based on the results by Drijvers and Özyürek (2017).

This study showed two main findings. First, results demonstrated that iconic co-speech gestures help speech intelligibility in linguistic maskers. Previous research on the enhancement effect of iconic gestures has demonstrated that iconic gestures help word intelligibility (e.g., Drijvers & Özyürek, 2017; Schubotz et al., 2020), and this study has replicated that finding in sentence context. Participants' performance on the recognition task increased in the gesture conditions. This finding confirms the hypothesis that participants would perform better in the gesture conditions compared to the conditions without gesture. Schubotz et al. found this gestural enhancement effect with Dutch targets in the Dutch babble, but they had not considered other linguistic maskers. This is the first study to show the enhancement effect of iconic gestures in both a foreign (French) and a native (Dutch) masker, as well as in sentence context.

The second main finding of this study is that speech intelligibility with and without iconic co-speech gestures was better in the French babble than in the Dutch babble. This is in line with the second hypothesis that participants would perform poorer in the Dutch babble than in the French babble across both gesture conditions. The participants had little experience with French, and the French masker and Dutch targets did not share linguistic similarities. The Dutch babble, however, was known to the participants and shared great linguistic overlap with the Dutch target sentences. This difference between French and Dutch causes the French babble to be an energetic masker for the participants, whereas the Dutch babble caused additional informational masking. The informational masker (Dutch) caused a larger masking effect than the energetic masker (French). This result suggests that the predictions by the TMLS hypothesis (Brouwer et al., 2012) are also valid when the target speech is accompanied by iconic hand gestures.

The effect of iconic gestures in different linguistic maskers had not been studied before; hence, there were two possible outcomes for a possible interaction between babble and gesture. On the one hand, iconic gestures can have the strongest enhancement effect in the most difficult condition (Dutch-in-Dutch) because participants need the semantic information from iconic gestures most in this condition. The results of this study did not reveal an interaction effect between gesture and babble, which suggests that the gestural enhancement effect did not increase or decrease in French or Dutch babble. The iconic co-speech gestures did not aid listeners more in Dutch babble, which is evidence against the first interaction speculation. This suggests that listeners do not rely on iconic gestures more in a more difficult linguistic masker than in an easier linguistic masker.

On the other hand, Drijvers and Özyürek (2017) found a greater enhancement effect in moderately degraded speech (six-band noise vocoding) compared to in severely degraded speech (two-band noise vocoding). They propose that there exists an optimal range for multimodal enhancement: Gestures have their maximum effect when "auditory cues are moderately reliable" (p. 219). Following this finding, it could be expected that iconic co-speech gestures in this study aid listeners more in French babble (i.e., the moderate level of difficulty). The absence of an interaction effect suggests that such an optimal range for multimodal enhancement was not created in this experiment. A possible explanation for the absence of this effect is the use of only one SNR level of −10 dB. It is unknown what level of speech degradation best optimizes the effect of gestures in these stimuli. Modifying the level of degradation across the

French and Dutch babble conditions might further optimize the gestural enhancement effect.

An innovative aspect of this study was the use of target sentences instead of isolated words. Previous research has established that iconic gestures can help target word recognition (e.g., Drijvers & Özyürek, 2017; Schubotz et al., 2020). However, face-to-face communication also involves conversations with sentences. The results of this study showed that word intelligibility in sentence context improved when iconic gestures were present. This suggests that gestures do not only help listeners recognize single words but that they can also help listeners recognize words in longer stretches of speech.

There are some limitations to this study that could be addressed in future research. First, it has to be noted that the target speech and babble were recorded with different equipment. The videos of the actress were made in a video laboratory, but the speakers for the babble recorded the texts with their own mobile devices (due to the COVID-19 pandemic). It is therefore difficult to guarantee that the sound quality of the babble was similar and as good as that of the stimuli sentences. Second, this study tested the TMLS hypothesis (Brouwer et al., 2012) with gestures accompanying the target speech, but, as suggested by an anonymous reviewer, to paint a more complete picture of the hypothesis in a multimodal setting, iconic gestures also need to be added to the masker(s). An evaluation of masker languages that have linguistically similar gestures would be needed to investigate whether similarity of the target and masker with regard to gestures also negatively affects speech intelligibility.

Furthermore, it remains unclear if the gestural enhancement effect found in this study would be the same for Dutch–French bilinguals. The participants that took part in the experiment knew little to no French, and most of them indicated to never actively use French. The results of this study support the TMLS hypothesis, but no claims can be made regarding the familiarity effect (Brouwer et al., 2012; Van Engen, 2010) because the participants were unfamiliar with French. A future multimodal experiment with Dutch–French bilinguals could show if a familiarity effect arises when the participants are familiar with both masker languages. Additionally, Drijvers and Özyürek (2020) studied gestural and visual enhancement in nonnatives and found that nonnative listeners benefited less from the combined gestural and visual enhancement effect than native speakers. Future research with nonnative speakers could reveal if they benefit less from the iconic gestures than native speakers of Dutch in sentence context.

Lastly, future research could further manipulate the languages used for the linguistic maskers. This study focused on Dutch and French because they are highly dissimilar and were thus expected to elicit a familiarity effect. It remains unclear if the same gestural enhancement effect can be found in different language pairs than Dutch and French. Calandruccio et al. (2010), for example, compared the masking effects of English and Mandarin because these languages vary in linguistic content, as well as phonetically and acoustically. Future experiments with different language pairs such as Dutch–Mandarin (highly spectrally, and highly linguistically distinct) and Dutch–English (less spectrally, but more linguistically distinct because these are more closely related Germanic languages) could reveal if their masking effects in a multimodal context differ or not due to spectral differences between the languages.

The societal implications of this study are invaluable to anyone who engages in multimodal communication. Many situations occur in which people are talking in the background. For example, at gatherings, shops, amusement parks, restaurants, or universities, there are always others in conversation. This competing background speech can complicate conversation, but this study provides evidence that iconic gestures along with spoken sentences can help a listener. Additionally, this study has demonstrated that the kind of masker language influences communication. Competing speech in a foreign language (French) provides less of a masking effect than a native language (Dutch) in a setting where target speech is supported by gestures. Furthermore, these results may be valuable to everyone who engages in online communication. Many people around the globe have adjusted to working in an online environment during the past 1.5 years. There is no doubt that communication has become more difficult through online platforms, and there are many factors that can cause this difficulty. When communicating through a video call, body language cannot be read as well as it can be in real life because often only faces are visible; facial movements, such as lipreading, cannot be spotted as easily; the online environment can be noisy due to bad connection or echoes; and turn taking is more difficult online, so people often talk over each other. This study strongly suggests that using iconic gestures helps communication, perhaps also during online video calls.

To conclude, this study investigated the effects of iconic co-speech gestures and masker language on word intelligibility in sentence context. The results indicate that word intelligibility in context improves in the presence of iconic gestures and that linguistic similarity between the target and masker language complicates intelligibility. Word intelligibility in sentence context becomes better with iconic co-speech gestures and when target–masker linguistic similarity is small. The results are valuable to everyone who engages in multimodal communication, online or off-line, and especially to a public who often communicates in public places.

Acknowledgments

We would like to thank Lieke Hendriks for being the actress in the videos, and Daniëlla van Uden and Nadine Courbois for help with the babble. We would also like to thank the Technical Group from the Max Plank Institute for Psycholinguistics for help with the stimuli. This work has been supported by a Minerva Fast Track Fellowship from the Max Planck Gesellschaft, awarded to Linda Drijvers.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boersma, P. (2001). PRAAT, a system for doing phonetics by computer. *Glot International, 5*(9–10), 341–345.

Brouwer, S. (2017). Masking release effects of a standard and a regional linguistic variety. *The Journal of the Acoustical Society of America, 142*(2), EL237–EL243. https://doi.org/10.1121/1.4998607

Brouwer, S., Akkermans, N., Hendriks, L., van Uden, H., & Wilms V. (2021). "Lass frooby noo!" The interference of song lyrics and meaning on speech intelligibility. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000368

Brouwer, S., & Bradlow, A. R. (2015). *The effect of target-background synchronicity on speech-in-speech recognition*. Proceedings of the 18th International Congress of Phonetic Sciences 2015 (ICPhS XVIII), Glasgow.

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America, 131*(2), 1449–1464. https://doi.org/10.1121/1.3675943

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America, 109*(3), 1101–1109. https://doi.org/10.1121/1.1345696

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America, 110*(5), 2527–2538. https://doi.org/10.1121/1.1408946

Calandruccio, L., Buss, E., Bencheck, P., & Jett, B. (2018). Does the semantic content or syntactic regularity of masker speech affect speech-on-speech recognition? *The Journal of the Acoustical Society of America, 144*(6), 3289–3302. https://doi.org/10.1121/1.5081679

Calandruccio, L., Dhar, S., & Bradlow, A. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America, 128*(2), 860–869. https://doi.org/10.1121/1.3458857

Calandruccio, L., & Zhou, H. (2014). Increase in speech recognition due to linguistic mismatch between target and masker speech: Monolingual and simultaneous bilingual performance. *Journal of Speech, Language, and Hearing Research, 57*(3), 1089–1097. https://doi.org/10.1044/2013_JSLHR-H-12-0378

Carhart, R., Tillman, T. W., & Greetis, E. S. (1969). Perceptual masking in multiple sound backgrounds. *The Journal of the Acoustical Society of America, 45*(3), 694–703. https://doi.org/10.1121/1.1911445

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975–979. https://doi.org/10.1121/1.1907229

Clark, H. H. (1996). *Using language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539

Darwin, C. J. (2008). Listening to speech in the presence of other sounds. *Transactions of the Royal Society B, 363*(1493), 1011–1021. https://doi.org/10.1098/rstb.2007.2156

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

Drijvers, L., & Özyürek, A. (2020). Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language and Speech, 63*(2), 209–220. https://doi.org/10.1177/0023830919831311

Drijvers, L., Özyürek, A., & Jensen, O. (2018). Hearing and seeing meaning in noise: Alpha, beta, and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human Brain Mapping, 39*(5), 2075–2087. https://doi.org/10.1002/hbm.23987

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America, 106*(6), 3578–3588. https://doi.org/10.1121/1.428211

Garcia, J., & Dagenais, P. (1998). Dysarthric sentence intelligibility. *Journal of Speech, Language, and Hearing Research, 41*(6), 1282–1293. https://doi.org/10.1044/jslhr.4106.1282

Garcia Lecumberri, M. L., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America, 119*(4), 2445–2454. https://doi.org/10.1121/1.2180210

Goldin-Meadow, S. (2005). *Hearing gesture: How our hands help us think*. Harvard University Press. https://doi.org/10.2307/j.ctv1w9m9ds

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

Hustad, K., & Garcia, J. (2005). Aided and unaided speech supplementation: Effect of alphabet cues and iconic hand gestures on dysarthric speech. *Journal of Speech, Language, and Hearing Research, 48*(5), 996–1012. https://doi.org/10.1044/1092-4388(2005/068)

Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2007). Informational masking. In W. Yost (Ed.), *Springer handbook of auditory research 29: Auditory perception of sound sources* (pp. 143–190). Springer. https://doi.org/10.1007/978-0-387-71305-2_6

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

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

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1651), 20130296. https://doi.org/10.1098/rstb.2013.0296

Pollack, I. (1975). Auditory informational masking. *The Journal of the Acoustical Society of America, 57*(S5). https://doi.org/10.1121/1.1995329

Powell, M. J. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives* [Technical report], Department of Applied Mathematics and Theoretical Physics, University of Cambridge.

Proust, M. (1954). *Du côte de chez Swann* [Swann's way]. Éditions Gallimard.

Proust, M. (2019). *De kant van Swann* [Swann's way]. De bezige bij.

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

RStudio Team. (2020). *RStudio: Integrated development for R*. RStudio, PBC. http://www.rstudio.com/

Schubotz, L., Holler, J., Drijvers, L., & Özyürek, A. (2020). Aging and working memory modulate the ability to benefit from visible speech and iconic gestures during speech-in-noise comprehension. *Psychological Research, 85*(5), 1997–2011. https://doi.org/10.1007/s00426-020-01363-8

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory–visual speech perception and auditory–visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26*(3), 263–275. https://doi.org/10.1097/00003446-20050 6000-00003

Stevenson, R. A., Nelms, C. E., Baum, S. H., Zurkovsky, L., Barense, M. D., Newhouse, P. A., & Wallace, T. (2015). Deficits in audiovisual speech perception in normal aging emerge at the level of whole-word recognition. *Neurobiology of Aging, 36*(1), 283–291. https://doi.org/10.1016/j.neurobiola ging.2014.08.003

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

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions: Biological Sciences, 335*(1273), 71–78. https://doi.org/10.1098/rstb.1992.0009

Summers, R. J., & Roberts, B. (2020). Informational masking of speech by acoustically similar intelligible and unintelligible interferers. *Journal of the Acoustical Society of America, 147*(2), 1113–1125. https://doi.org/10.1121/10.0000688

Tun, P., O'Kane, G., & Wingfield, A. (2002). Distraction by competing speech in young and older adult listeners. *Psychology and Aging, 17*(3), 453–467. https://doi.org/10.1037//0882-7974.17.3.453

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing, 28*(5), 656–668. https://doi.org/10.1097/AUD.0b013e31812f7185

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audiovisual integration, and the principle of inverse effectiveness. *Ear and Hearing, 31*(5), 636–644. https://doi.org/10.1097/AUD.0b013e3181 ddf7ff

Van Engen, K. J. (2010). Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. *Speech Communication, 52*(11–12), 943–953. https://doi.org/10.1016/j.specom.2010.05.002

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America, 121*(1), 519–526. https://doi.org/10.1121/1.2400666

VideoLan. (2006). *VLC media player*. https://www.videolan.org/vlc/index.html

Williams, B., & Viswanathan, N. (2020). The effects of target-masker sex mismatch on linguistic release from masking. *The Journal of the Acoustical Society of America, 148*(4), 2006–2014. https://doi.org/10.1121/10.0002165

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 1556–1559.

Appendix (p. 1 of 3)

Target Stimuli Used in the Experiment

Gesture condition	Target sentence	English translation
Gesture	Ze begint te ritsen	She begins to zip
Gesture	Ze begint te fluiten	Ze begins to play flute
Gesture	Hij begint te vertellen	He begins to talk
Gesture	Ze begint te openen	She begins to open
Gesture	Ze begint te kruipen	She begins to crawl
Gesture	Hij durft niet te kloppen	He does not dare to knock
Gesture	Ze durft niet te steken	She does not dare to stab
Gesture	Ze durft niet te doden	She does not dare to kill
Gesture	Hij durft niet te rollen	He does not dare to roll
Gesture	Ze gaat vaak vissen	She often goes fishing
Gesture	Hij gaat vaak boksen	He often goes boxing
Gesture	Ze gaat vaak fietsen	She often goes cycling
Gesture	Ze gaat vaak wandelen	She often goes walking
Gesture	Ze heeft zin om te proosten	She is excited to toast
Gesture	Hij heeft zin om te eten	He is excited to eat
Gesture	Hij heeft zin om te stempelen	He is excited to stamp
Gesture	Ze heeft zin om te joggen	She is excited to jog
Gesture	Hij heeft zin om te mixen	He is excited to mix
Gesture	Hij houdt erg van drummen	He really likes to drum
Gesture	Hij houdt erg van roeien	He really likes to row
Gesture	Ze houdt erg van knuffelen	She really likes to hug
Gesture	Ze houdt erg van tekenen	She really likes to draw
Gesture	Hij houdt ervan om te krassen	He really likes to scratch
Gesture	Hij houdt ervan om te lezen	He really likes to read
Gesture	Ze houdt ervan om te computeren	She really likes to work on the computer
Gesture	Ze houdt van timmeren	She likes to do carpentry
Gesture	Ze houdt van signeren	She likes to sign
Gesture	Ze houdt van vegen	She likes to sweep
Gesture	Hij houdt van geven	He likes to give
Gesture	Hij is bang om te aaien	He is afraid to pet
Gesture	Ze is bang om te vliegen	She is afraid to fly
Gesture	Ze is bang om te klimmen	She is afraid to climb
Gesture	Hij is klaar om te bidden	He is ready to pray
Gesture	Ze is klaar om te drinken	She is ready to drink
Gesture	Ze is klaar om te buigen	She is ready to bow
Gesture	Ze is klaar om te stoppen	She is ready to stop
Gesture	Hij is klaar om te werpen	He is ready to throw
Gesture	Hij kan erg goed darten	He is very good at darts
Gesture	Ze kan erg goed zwemmen	She is very good at swimming
Gesture	Hij kan erg goed tennissen	He is very good at tennis
Gesture	Hij kan erg hard slaan	He can hit very hard
Gesture	Ze kan erg hard knijpen	She can squeeze very hard
Gesture	Ze kan erg hard gooien	She can throw very hard
Gesture	Hij kan erg hard schieten	He can shoot very hard
Gesture	Hij kan erg hard trekken	He can pull very hard
Gesture	Hij kan erg mooi schrijven	He can write very prettily
Gesture	Ze kan erg mooi scheuren	She can tear very prettily
Gesture	Hij kan erg mooi filmen	He can film very prettily
Gesture	Ze kan erg mooi zingen	She can sing very prettily
Gesture	Hij kan erg snel beitelen	He can chisel very quickly
Gesture	Hij kan erg snel typen	He can type very quickly
Gesture	Ze kan erg snel vouwen	She can fold very quickly
Gesture	Hij kan erg snel marcheren	He can march very quickly
Gesture	Ze kan goed touwtjespringen	She is good at jumping rope
Gesture	Ze kan goed fotograferen	She is good at taking pictures
Gesture	Ze kan goed stapelen	She is good at stacking
Gesture	Hij kan goed vangen	He is good at catching
Gesture	Ze kan niet goed skieen	She is not good at skiing
Gesture	Hij kan niet goed wassen	He is not good at washing
Gesture	Ze kan niet goed balanceren	She is not good at keeping balance
Gesture	Hij kan niet snel lopen	He cannot walk fast
Gesture	Hij kan niet snel schilderen	He cannot paint fast
Gesture	Hij kan niet snel groeien	He cannot grow fast
Gesture	Hij kan niet snel markeren	He cannot mark fast

Appendix (p. 2 of 3)

Target Stimuli Used in the Experiment

Gesture condition	Target sentence	English translation
Gesture	Ze kan snel smeren	She can smear quickly
Gesture	Hij kan snel graven	He can dug quickly
Gesture	Hij kan snel schudden	He can shake quickly
Gesture	Ze kan snel tellen	She can count quickly
Gesture	Ze kan snel knopen	He can tie quickly
Gesture	Hij ontspant van scrollen	Scrolling relaxes him
Gesture	Ze staat altijd te praten	She is always talking
Gesture	Hij staat altijd te appen	He is always texting
Gesture	Ze staat altijd te zwaaien	She is always waving
Gesture	Hij staat altijd te frutselen	He is always fiddling
Gesture	Ze staat altijd te bedienen	She is always waitressing
Gesture	Ze staat daar te wijzen	She is pointing there
Gesture	Hij staat daar te gieten	He is watering there
Gesture	Ze staat daar te duiken	She is diving there
Gesture	Hij staat op om te klappen	He stands up to clap
Gesture	Ze staat op om te strijken	She stands up to iron
Gesture	Ze staat op om te verplaatsen	She stands up to move
Gesture	Hij staat op om te bezorgen	He stands up to deliver
Gesture	Hij staat te waaieren	He is waving
Gesture	Hij staat te beven	He is shivering
Gesture	Ze staat te roeren	She is stirring
Gesture	Hij staat te wiegen	He is cradling
Gesture	Hij staat te zeven	He is sieving
Gesture	Ze staat te kruiden	She is seasoning
Gesture	Hij zit altijd te gamen	He is always gaming
Gesture	Ze zit altijd te klikken	She is always clicking
Gesture	Ze zit altijd te duwen	She is always pushing
Gesture	Ze zit altijd te bladeren	She is always browsing
Gesture	Hij zit altijd te ontstoppen	He is always unclogging
Gesture	Hij zit vaak te duimen	He always has his fingers crossed
Gesture	Hij zit vaak te bellen	He is always calling
Gesture	Hij zit vaak te glijden	He is always sliding
Gesture	Hij houdt niet van wringen	He does not like to wring
Gesture	Ze houdt niet van vijlen	She does not like to file
Gesture	Ze houdt niet van snijden	She does not like to cut
Gesture	Ze houdt niet van scheiden	She does not like to separate
No gesture	Hij begint te knippen	He begins to cut
No gesture	Hij begint te roken	He begins to smoke
No gesture	Hij begint te kogelstoten	He begins to play shot put
No gesture	Ze durft niet te wurgen	She does not dare to strangle
No gesture	Hij durft niet te blussen	He does not dare to extinguish
No gesture	Hij durft niet te begraven	He does not dare to burry
No gesture	Hij gaat vaak golven	He often goes to play golf
No gesture	Ze gaat vaak kopieren	She often goes to make a copy
No gesture	Hij gaat vaak plakken	He often goes to stick
No gesture	Hij gaat vaak persen	He often goes to push
No gesture	Ze heeft zin om te bowlen	He is excited to bowl
No gesture	Hij heeft zin om te videobellen	He is excited to video call
No gesture	Ze heeft zin om te kleien	She is excited to clay
No gesture	Ze houdt erg van bakken	She really likes to bake
No gesture	Hij houdt erg van sporten	He really likes to play sports
No gesture	Ze houdt erg van bewateren	She really likes to water
No gesture	Ze houdt ervan om te dobbelen	She likes to dice
No gesture	Ze houdt ervan om te dippen	She likes to dip
No gesture	Hij houdt ervan om te skateboarden	He likes to skateboard
No gesture	Hij houdt ervan om te brabbelen	He likes to babble
No gesture	Hij houdt van jojoen	He likes to play yoyo
No gesture	Hij houdt van kaarten	He likes to play cards
No gesture	Ze houdt van tuinieren	She likes to garden
No gesture	Ze houdt van ventileren	She likes to ventilate
No gesture	Hij houdt van giechelen	He likes to giggle
No gesture	Ze is bang om te vliegen	She is afraid to fly
No gesture	Hij is bang om te ontkurken	He is afraid to uncork
No gesture	Hij is bang om te slijpen	He is afraid to sharpen
No gesture	Ze is bang om te paaldansen	She is afraid to pole dance
No gesture	Hij is bang om te handballen	He is afraid to play handball

Appendix (p. 3 of 3)

Target Stimuli Used in the Experiment

Gesture condition	Target sentence	English translation
No gesture	Hij is klaar om te melken	He is ready to milk
No gesture	Ze is klaar om te hakken	She is ready to chop
No gesture	Hij kan erg goed sjoelen	He is very good at playing shuffleboard
No gesture	Hij kan erg goed poolen	He is very good at playing pool
No gesture	Ze kan erg goed speerwerpen	She is very good at javelin throwing
No gesture	Ze kan erg goed luisteren	She is very good at listening
No gesture	Hij kan erg goed springen	He is very good at jumping
No gesture	Ze kan erg goed fluisteren	She is very good at listening
No gesture	Hij kan erg hard kietelen	He can tickle very hard
No gesture	Hij kan erg hard schreeuwen	He can shout very hard
No gesture	Ze kan erg hard schoppen	She can shop very hard
No gesture	Ze kan erg hard stampen	She can kick very hard
No gesture	Ze kan erg mooi naaien	She can sew very prettily
No gesture	Ze kan erg mooi slingeren	She can swing very prettily
No gesture	Hij kan erg mooi fileren	She can fillet a fish very prettily
No gesture	Hij kan erg mooi borduren	He can stitch very prettily
No gesture	Ze kan erg snel schillen	She can peel very quickly
No gesture	Ze kan erg snel plukken	She can pick very quickly
No gesture	Ze kan erg snel verkorten	She can shorten very quickly
No gesture	Hij kan erg snel hoepelen	He can hula hoop very quickly
No gesture	Hij kan goed koken	He is good at cooking
No gesture	Hij kan goed dirigeren	He is good at being a conductor
No gesture	Hij kan goed voetballen	He is good at playing football
No gesture	Ze kan goed monteren	She is good at
No gesture	Ze kan goed ondersteunen	She is good at supporting
No gesture	Hij kan niet goed vlechten	He cannot braid well
No gesture	Hij kan niet goed rijden	He cannot drive well
No gesture	Ze kan niet goed zagen	She cannot saw well
No gesture	Hij kan niet goed sluipen	He cannot sneak well
No gesture	Ze kan niet goed drijven	She cannot float well
No gesture	Ze kan niet snel raspen	She cannot grate quickly
No gesture	Ze kan niet snel proppen	She cannot cram quickly
No gesture	Ze kan snel verbinden	She can connect quickly
No gesture	Hij kan snel dribbelen	He can dribble quickly
No gesture	Hij ontspant van knikkeren	Playing marbles relaxes him
No gesture	Ze ontspant van borstcrawlen	Breast crawl relaxes her
No gesture	Ze ontspant van breien	Knitting relaxes her
No gesture	Hij staat altijd te nieten	He is always stapling
No gesture	Ze staat altijd te frankeren	She is always franking
No gesture	Hij staat altijd te dansen	He is always dancing
No gesture	Hij staat daar te hijsen	He is hoisting there
No gesture	Ze staat daar te slepen	She is dragging there
No gesture	Hij staat daar te verdelen	He is dividing there
No gesture	Ze staat daar te planten	She is planting there
No gesture	Hij staat daar te spuiten	He is spouting there
No gesture	Ze staat op om te wenken	She stands up to wave
No gesture	Hij staat op om te stofzuigen	He stands up to vacuum clean
No gesture	Ze staat op om te vragen	She stands up to ask
No gesture	Hij staat op om te faxen	He stands up to fax
No gesture	Ze staat op om te schroeven	She stands up to screw
No gesture	Ze staat te knikken	She is nodding
No gesture	Hij staat te trappen	He is kicking
No gesture	Hij zit altijd te toeteren	He is always honking
No gesture	Hij zit altijd te parfumeren	He is always spraying perfume
No gesture	Ze zit altijd te spelen	She is always playing
No gesture	Ze zit vaak te poetsen	She is often cleaning
No gesture	Ze zit vaak te boren	She is often drilling
No gesture	Ze zit vaak te schommelen	She is often swinging
No gesture	Ze zit vaak te huilen	She is often crying
No gesture	Hij zit vaak te haken	He is often crocheting
No gesture	Hij houdt niet van schrobben	He does not like to scrub
No gesture	Hij houdt niet van stikken	He does not like to suffocate
No gesture	Hij houdt niet van verminderen	He does not like to decrease
No gesture	Ze houdt niet van masseren	She does not like to massage