

<Martin's pithy title about gesture and speech>.

JK, MC, HR

Psychology, PPLS, University of Edinburgh

Josiah King

Philosophy, Psychology and Language Sciences

University of Edinburgh

7 George Square

Edinburgh EH8 9JZ, UK

J.P.J.King@sms.ed.ac.uk

Abstract

When a task becomes more conceptually demanding, speakers tend to produce more gestures. This might be because gestures help the speaker package more complex information into appropriate units for speech, or because some of the communicative load is being traded off from speech to gesture. The present study is a dialogue study designed to tease apart these views.

Previous research has found evidence for both views. Speech and gesture have been found to specify referents in parallel, but a general increase in gesture duration over a trial has been found when referents are less describable.

Studies such as these have tended to look at the relationship between gesture and speech where there is no addressee, or where the addressee is not explicitly contributing to a dialogue. Moreover, gesture has tended to be measured as the number of discrete gestures per word or trial. Because gestures and vocalisations can vary in durations, these metrics fail to capture the relative contributions of each. In the present study we use naturalistically-occurring dialogues, and measure the relative durations of gestures and speech directly, in order to establish whether they co-vary (as would be predicted by the packaging account) or correlate inversely (as would be predicted by a trade-off).

Twenty-two pairs of participants took part in a shape-matching game, in which a director described two target shapes, and a matcher attempted to identify the correct referents among an array of competitors. Roles alternated on each trial, and the shapes were either easy or difficult to verbally encode. Directors saw two shapes (one easy, one difficult) for two seconds. They then had 10 seconds to describe these shapes to their partner. Points were scored for each shape which the matcher correctly identified in an array of six shapes.

In line with an information packaging account, speech duration and gesture duration increased in parallel for easy-to-name objects $\beta = 0.57$, $SE = 0.06$, $t > 2$. Importantly, when participants were referring to objects which were more difficult to verbally encode, gesture duration increased at a higher rate than speech duration $\beta = 0.27$, $SE = 0.06$, $t > 2$.

This suggests that, when naming difficult objects, gesture does more than simply help the speaker to package verbal information. Instead, gesture serves an additional communicative purpose, adding to, but not trading off against, the verbal descriptions which are uttered.

<Martin's pithy title about gesture and speech>.

During conversation, speakers often move their hands and arms in meaningful ways which co-express the content of their speech (McNeill, 1992): These movements might add emphasis to speech; indicate location; or depict properties of objects, movements, actions or space. Not all speech is accompanied by gesturing, and recent research has aimed at better understanding the relationship between the two modalities. Few studies, however, have investigated the speech-gesture relationship in a truly dialogical setting.

One prominent theory of how speech and gesture interact (Bangerter, 2004; de Ruiter, 2006) is that there is trade-off in communicative load between the two modalities: When speaking becomes more difficult, gesturing takes over more of the informational content (and vice versa). This is supported by studies which have found increased gesturing when speakers are describing referents with lower verbal codability (Morsella & Krauss, 2004); describing novel (as opposed to repeated) referents and narratives (Hoetjes, Koolen, Goudbeek, Krahmer, & Swerts, 2015; Jacobs & Garnham, 2007); and referring to targets which are further away (Bangerter, 2004). These findings suggest that as the effort required to verbally encode meaning increases, so does the likelihood of a gesture occurring. However, this increase alone does not necessitate a trade-off in communicative load between the two modalities. This evidence comes from Melinger and Levelt (2004), who found that participants' speech was less explicit when it was accompanied by gesturing than when it was not, suggesting that this was due to gesturing assuming some of the communicative load from speaking.

Melinger and Levelt's findings contrast with that of (So, Kita, & Goldin-Meadow, 2009), who suggests an alternative account: As conceptual load increases, speech and gesture do not trade-off against one another, but increase in parallel. In So et al.'s study, speakers' use of gesture did not compensate for, but instead paralleled underspecifications in speech. Participants were asked to describe scenes from videotaped vignettes (e.g., a man giving woman a basket), and their use of both speech and gesture to indicate characters in the scene was measured. So et al. found that speakers more often used a gesture to identify a referent if it was also specified in speech. The idea that speech and

gesture increase hand-in-hand with one another is often accompanied by the theory that gesturing is motivated by the cognitive benefits it has for the speaker. Many studies suggest this might be the case: Gesturing has been found to aid speech production and planning processes (Kita, 2000; Krauss & Hadar, 1999; Morsella & Krauss, 2004; Rauscher, Krauss, & Chen, 1996; Rose & Douglas, 2001), spatial working memory (Morsella & Krauss, 2004; Wesp, Hesse, Keutmann, & Wheaton, 2001) and conceptual planning (Melinger & Kita, 2007).

However, whilst much gesturing might be motivated by the benefits it has for the speaker, some gestures are clearly explicitly intended to communicate (e.g. "A shape like this [traces a shape in the air]"). It may be that the trade-off hypothesis holds for certain types of gesturing, or when the information being communicated lends itself to being more easily specified in gesture than in speech (i.e. novel shapes which can be easily and clearly represented in space). Evidence for different types of gesture potentially serving different purposes can be seen in findings in which mutual visibility affects speakers' use of gestures. Alibali, Heath, and Myers (2001) found mutual visibility to influence *representational* gestures, but not *beat* gestures. Similarly, de Ruiter, Bangerter, and Dings (2012) found that speakers' who could see their addressee used pointing and *obligatory iconic* gestures more (gestures containing information not represented in speech but required for disambiguation), but not *non-obligatory iconic* gestures (gestures which are not essential to the understanding of co-occurring speech).

Additionally, Mol, Krahmer, Maes, and Swerts (2011) found that over computer-mediated conversation, more gestures were produced when participants knew they were visible to the addressee than when the addressee was visible to them, suggesting that these gestures were produced to be seen. Mol et al.'s finding reflects that of Bavelas, Kenwood, Johnson, and Phillips (2002), in which participants described stimuli to a supposed future addressee who would receive either audio or audiovisual recordings. Bavelas et al.'s results showed participants gesturing more when they believed their descriptions would be seen by the future addressee. It is not just the visible presence of an addressee which influences a speaker's gesturing: Several studies suggest

that speakers tailor their use of gesture to specific contexts and addressees: As well as mutual visibility (Alibali et al., 2001; Cohen & Harrison, 1973; Hoetjes et al., 2015), an increase in the quantity or precision of speakers' gestures has been associated with less common ground (Gerwing & Bavelas, 2005; Holler & Stevens, 2007), and information which is novel for the addressee (Jacobs & Garnham, 2007).

Much of the evidence above suggests that the addressee plays an important role in a speaker's use of speech and gesture. Additionally, gesturing is often useful for the addressee (for an overview, see Hostetter (2011)). Despite this, there has been little research into how speech and gesture interact in a truly dialogical setting. To date, much of the research into gestures has involved studies in which a single speakers' gestures are evaluated under various conditions. Experimental paradigms have tended towards those in which participants produce speech and gesture either to an imagined future addressee (e.g. Morsella and Krauss (2004); Wesp et al. (2001)) or to an addressee who is present but in a comparatively passive role (e.g. a "matcher" to a "director", as in Bangerter (2004); de Ruiter et al. (2012); Hoetjes et al. (2015); Holler and Stevens (2007)). These sorts of designs fail to capture the dynamic process of conversation - that is, the fact that dialogue is a joint activity (Clark, 1996).

A study by Bavelas, Gerwing, Sutton, and Prevost showed independent effects of visibility and dialogue on gesturing: Speakers produced more gestures when speaking to an occluded but dialogically involved listener than when speaking to a tape recorder. Furthermore, recent research suggests that speakers will mimic their interlocutors gestures (Holler & Wilkin, 2011; Kimbara, 2008).

JK got this far

trying to work out where this next bit goes..

gesture rate

One reason for the mixed results in studies surrounding how speech and gesture interact may be the differences in how the relationship between the two is measured: Tending towards *rate* measures (i.e. gesture as a function of speech), studies have differed in their choice of measures for both modalities. Much research has involved measuring the number of discrete gestures produced when speaking (e.g. de Ruiter et al. (2012); Gerwing and Allison (2011); Hoetjes et al. (2015); Hostetter, Alibali, and Kita (2007)). However, coding a speaker's arm movements into distinct gestures is not straightforward. Unlike speech — where distinct phonemes and words offer comparatively clear means of measuring utterance length and duration — multiple pieces of information (and multiple types of gesture) may be produced in the time between the raising and lowering of hands. In the literature, defining individual gestures has varied with the thing being described, from “illustrating a feature of the target (for instance its shape).” when describing tangram figures (de Ruiter et al., 2012) to change in any one of “shape and placement of the hand, trajectory of the motion” when identifying referents in a narrative (So et al., 2009). Furthering to this confusion, rates of discrete gestures have been calculated per minute (Mol et al., 2011); per 100 words (Gerwing & Allison, 2011; Hoetjes et al., 2015; Hostetter et al., 2007; Masson-Carro, Goudbeek, & Krahmer, 2015); per *feature description* (de Ruiter et al., 2012) or per *semantic attributes* (Hoetjes et al., 2015).

For representational/iconic gesturing, we propose that a duration-based account offers a simpler (but underused) measure. This measure involves simply comparing the durations for which a speaker conveys (or attempts to convey) information via different channels. Gesturing is often considered to be split into three main phases: Preparation; Stroke or Hold; and Retraction (McNeill, 1992). Excluding the preparatory and retraction phases, comparing the duration for which a speakers hands are in the air with the duration for which they are speaking offers a more objective measure of how information is distributed in multimodal communication.

Experiment

?? ?? ??

Setup

Pairs of participants engaged in a collaborative matching game, in which they were tasked with matching two target shapes seen by one participant from a set of six shapes seen by the other participant. Participants took turns in the roles of *director* and *matcher*. Target shapes varied in verbal codability: they were either easy- or hard-to-name. We also manipulated participants' familiarity with the target shapes in both roles (director and matcher).

Materials

Pairs of shapes in critical trials were selected from a set of 20 critical shapes (10 easy-to-name, 10 hard-to-name). Each of the 10 easy-to-name shapes was altered (sections of the shape were rotated and/or flipped) to create the 10 hard-to-name variants. For filler trials, shapes were selected from a further set of 40 shapes (20 easy-to-name shapes and 20 hard-to-name variants).

A set of 20 *distractor* shapes (10 easy-, 10 hard-to-name) were visually and descriptively similar to the 20 critical target shapes. These shapes were never described, only being presented as part of the matchers' arrays, with the aim of encouraging specificity in future descriptions.

In critical trials, the two shapes seen by the director differed in codability. In filler trials, target shapes were both the same codability (both easy-to-name or both hard-to-name).

Matchers' arrays were composed of six shapes. Along with the two target shapes for that trial, this was comprised of two randomly selected filler shapes and two randomly selected distractor shapes. Codability of the shapes in the array matched that of the target shapes (i.e. for critical trials, half were easy-to-name, half were hard-to-name). Positions of shapes were randomly selected for both the director (Left vs. Right) and for

the matcher (in a 2x3 grid).

Experimental blocks

The experiment consisted of two blocks, each containing 40 trials (20 critical and 20 fillers). In each block, every critical shape was seen twice.

In the first block, critical trials were sampled without replacement from a list of 20 trials, with the constraint that target shapes were never repeated in consecutive critical trials. Whilst the majority of the shapes were described once by each participant, the probability that a given shape was described twice by the same participant was 25%¹.

For the 20 filler trials, two target shapes of the same codability were randomly selected from the set of filler shapes.

In the second block, pairs of consecutive critical trials alternative with pairs of filler trials. For the consecutive critical trials, the difficult-to-name shape was repeated. This meant that for critical trials where participant B was the director, they were tasked with describing the difficult shape which had just been described to them in the previous trial by participant A. Filler trials remained the same as in block 1.

Procedure

Participants sat facing each other on chairs (without any arms), with an unobstructed view of each other. Each participant had a monitor and a mouse on a table to their left, positioned such that they could not see what was on their partner's monitor. The setup was designed to encourage face-to-face dialogue, and to discourage participants from leaving their hand resting on the mouse whilst speaking (the position being uncomfortable for a right-handed mouse user), thus leaving both arms free to gesture. Audio and video was recorded by two cameras positioned to the right of each participant, facing their partner.

Taking turns in the roles of director and matcher, participants were tasked with successfully matching what was seen on the director's monitor from a set of possibilities

¹Whilst the intention was to make it so that each participant described each critical shape once, a typing error in the experiment script resulted in these distributions

on the matcher's monitor.

As a pair, participants were awarded points for successful matching. As an incentive, the highest scoring pair received 40, and participants were informed of this beforehand.

A high-score table was shown prior to the experiment, with participants adding their score to the table after they had played. To encourage face-to-face dialogue, participants were only allowed to communicate within a restricted time-window of 10 seconds during which no images were present on either screen. Participants were told that during this period, they were "both allowed to talk, gesture, ask questions, and so on". Thus gesturing was permitted but not explicitly encouraged. The end of this window was signified by the sound of a bell (see Figure 1).

Feedback was given at the end of each trial (The number of correctly matched shapes was signified by an equivalent number of bell rings, with zero correct resulting in a buzzer), and the participants' cumulative score was displayed.

Coding

Audiovisual data for each pair of participants was coded using a three stage process: Audio-only and Video-only stages were used to code for speech and gesture respectively, with the third stage (both Audio and Video) used to confirm the variables resulting from the previous stages. As each trial consisted of describing two shapes, special care was taken in the third stage to ensure that utterances and gestures were assigned to the correct referents².

Speech Coding

Utterance duration, utterance length and disfluencies were coded in the Audio-only stage. Only the first mention of each shape was used. Utterance duration (ms) was coded from the onset of the noun-phrase up until either a) speech-offset, or b) a valid interruption from the listener in either modality. Listeners' use of the collateral channel (for instance: "yep", "mmhm", [nods head]) were not considered valid interruptions.

²There was potential for descriptions in both modalities to merge into one another

Utterance length (number of words) was coded analogously, with disfluencies within the utterance period being identified and excluded from this measure.

Disfluencies were coded as falling into one of six categories: Filled pauses; Insertions; Substitutions; Articulation Errors; Deletions; Repetitions. (ref Shriberg). Any speech prior to the onset of the noun-phrase was coded as either fluent or disfluent.

Conceptual pacts were coded during the third (both Audio and Video) stage.

Gesture Coding

Gestures were identified in the Video-only stage of the coding process. Any movement from the fingers up to the shoulder were considered. Only gestures which partially overlapped an identified utterance period were included, and were assigned to the utterance which they primarily overlapped. This pairing was then confirmed in the third stage of the coding process. In any cases of a gesture being ambiguous as to which utterance it accompanied (i.e. adaptor gestures which overlap both utterance periods), the gesture was assigned to both referents.

Gestures were categorised into five types: Iconics; Beats; Points; Adaptors; and Others. Any gesture which was considered to be an attempt to represent any feature of the target shape was coded as an Iconic gesture. Beat gestures were identified as any movements which rhythmically matched prosody in speech but which *did not* represent any feature of the target shape. Point gestures were extensions of the index finger used to refer deictically to either present objects or people, or to previous parts of the discourse. Other movements were categorised as either adaptor gestures (scratching, stroking, manipulating clothing, etc.), or other miscellaneous gesticulations.

Individual gestures were identified by onset of movement, and continued either until the start of the retraction phase, or until transformation into a) a different category of gesture or b) iconic gesturing referring to a different shape³.

The third stage of the coding process (audio and video) was used to confirm the coding of the first and second stages, specifically gesture categorisation and pairing of

³Because each trial involved a participant describing two shapes...

gestures with referents. Additionally, this stage was used to code whether or not the utterance referred explicitly to the gesture being made (e.g. "like this", "like that", "a bit here", etc.) Several gestures remained ambiguous between iconic and beat even after this third stage (n referencing easy shapes, and n referencing difficult shapes). To err on the side of caution, these were considered to be imprecise/lax attempts at representing the shapes in space, and thus coded as iconic gestures.

Once identified, iconic gestures were coded for gesture duration analogously to the measure of utterance duration: Gesture duration for iconic gestures was measured from the onset of the first stroke or hold phase up until the retraction phase, or until interruption. End-of-gesture hangs (uninformative hangs immediately prior to a retraction phase) were not included⁴.

This measure of gesture duration included any hangs, false starts, or preparation which occurred within-gesture, just as utterance duration included within-utterance pauses and disfluencies. Any suspected false starts and repetitions were counted (a finger trace which is subsequently reversed counts as repetition, as does a static hold with a distinct beat gesture incorporated).

Additionally, all gestures were coded for the hands used (Left, Right, or Both), and whether the representational part of the gesture was conveyed dynamically, statically, or as a combination of both.

Analysis

Results

⁴We discerned here between end-of-gesture *hangs* and end-of-gesture *holds* which continued to convey some representational content, and were thus included as part of gesture duration

References

- Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of Visibility between Speaker and Listener on Gesture Production: Some Gestures Are Meant to Be Seen,. *Journal of Memory and Language*, *44*, 169–188. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0749596X00927529> doi: 10.1006/jmla.2000.2752
- Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, *15*(6), 415–419. doi: 10.1111/j.0956-7976.2004.00694.x
- Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, *58*(2), 495–520. doi: 10.1016/j.jml.2007.02.004
- Bavelas, J., Kenwood, C., Johnson, T., & Phillips, B. (2002). An experimental study of when and how speakers use gestures to communicate. *Gesture*, *2*(1), 1–17.
- Clark, H. H. (1996). *Using language*. Cambridge university press.
- Cohen, A. A., & Harrison, R. P. (1973). Intentionality in the use of hand illustrators in face-to-face communication situations. *Journal of Personality and Social Psychology*, *28*(2), 276.
- de Ruiter, J. P. (2006). Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech Language Pathology*, *8*(2), 124–127. doi: 10.1080/14417040600667285
- de Ruiter, J. P., Bangerter, 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*(2), 232–248. doi: 10.1111/j.1756-8765.2012.01183.x
- Gerwing, J., & Allison, M. (2011). The flexible semantic integration of gestures and words: Comparing face-to-face and telephone dialogues. *Gesture*, *11*(3), 308–329. Retrieved from <http://www.jbe-platform.com/content/journals/10.1075/gest.11.3.03ger>

doi: 10.1075/gest.11.3.03ger

Gerwing, J., & Bavelas, J. (2005). Linguistic influences on gesture's form. *Gesture*, 4(2), 157–195. Retrieved from

<http://www.jbe-platform.com/content/journals/10.1075/gest.4.2.04ger>

doi: 10.1075/gest.4.2.04ger

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2015). Reduction in gesture during the production of repeated references. *Journal of Memory and Language*, 79-80, 1–17. Retrieved from

<http://dx.doi.org/10.1016/j.jml.2014.10.004> doi:

10.1016/j.jml.2014.10.004

Holler, J., & 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. Retrieved from [http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=24195444&site=ehost](http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=24195444&site=ehost-live)

[-live&url=http://jls.sagepub.com/cgi/doi/10.1177/0261927X06296428](http://jls.sagepub.com/cgi/doi/10.1177/0261927X06296428)

doi: 10.1177/0261927X06296428

Holler, J., & Wilkin, K. (2011). Co-Speech Gesture Mimicry in the Process of Collaborative Referring During Face-to-Face Dialogue. *Journal of Nonverbal Behavior*, 35(2), 133–153. doi: 10.1007/s10919-011-0105-6

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

Hostetter, A. B., Alibali, M. W., & Kita, S. (2007). I see it in my hands' eye:

Representational gestures reflect conceptual demands. *Language and Cognitive Processes*, 22(3), 313–336. doi: 10.1080/01690960600632812

Jacobs, N., & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language*, 56(2), 291–303.

Kimbara, I. (2008). Gesture form convergence in joint description. *Journal of Nonverbal Behavior*, 32(2), 123–131. doi: 10.1007/s10919-007-0044-4

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.),

Language and gesture (Vol. 1, pp. 162–185). Cambridge University Press.

- Krauss, R. M., & Hadar, U. (1999). The Role of Speech-Related Arm/Hand Gestures in Word Retrieval. *Gesture, speech, and sign*, 93–116. Retrieved from <http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CFkQFjAA&url=http://www.columbia.edu/~rmk7/PDF/K%26H.pdf&ei=157HT-fMM4TPhAfU5qC0Cw&usg=AFQjCNHZzh3m5L3{ }4EQ1Na5ce2FZ9gWkwQ{ }sig2=5-DHpjgqUxXJvfY5Bo2F2w>
doi: 10.1080/14417040600667293
- Masson-Carro, I., Goudbeek, M., & Krahmer, E. (2015). Can you handle this? The impact of object affordances on how co-speech gestures are produced. *Language, Cognition and Neuroscience*, 3798(March), 1–11. Retrieved from <http://www.tandfonline.com/doi/full/10.1080/23273798.2015.1108448>
doi: 10.1080/23273798.2015.1108448
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.
- Melinger, A., & Kita, S. (2007). Conceptualisation load triggers gesture production. *Language and Cognitive Processes*, 22(4), 473–500. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/01690960600696916> doi: 10.1080/01690960600696916
- Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture*, 4(2)(2004), 119–141. doi: 10.1075/gest.4.2.02mel
- Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2011). Seeing and being seen: The effects on gesture production. *Journal of Computer-Mediated Communication*, 17(1), 77–100. doi: 10.1111/j.1083-6101.2011.01558.x
- Morsella, E., & Krauss, R. M. (2004). The Role of Gestures in Spatial Working Memory and Speech. *American Journal of Psychology*, 117(3), 411–424. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15457809> doi: 10.2307/4149008
- Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7(4),

226. Retrieved from <http://pss.sagepub.com/content/7/4/226.short> doi: 10.1111/j.1467-9280.1996.tb00364.x
- Rose, M., & Douglas, J. (2001). The differential facilitatory effects of gesture and visualisation processes on object naming in aphasia. *Aphasiology*, 15(10), 977–990. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/02687040143000339> doi: 10.1080/02687040143000339
- 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(1), 115–125. doi: 10.1111/j.1551-6709.2008.01006.x
- Wesp, R., Hesse, J., Keutmann, D., & Wheaton, K. (2001). Gestures maintain spatial imagery. *American Journal of Psychology*, 114(4), 591–600. doi: 10.2307/1423612

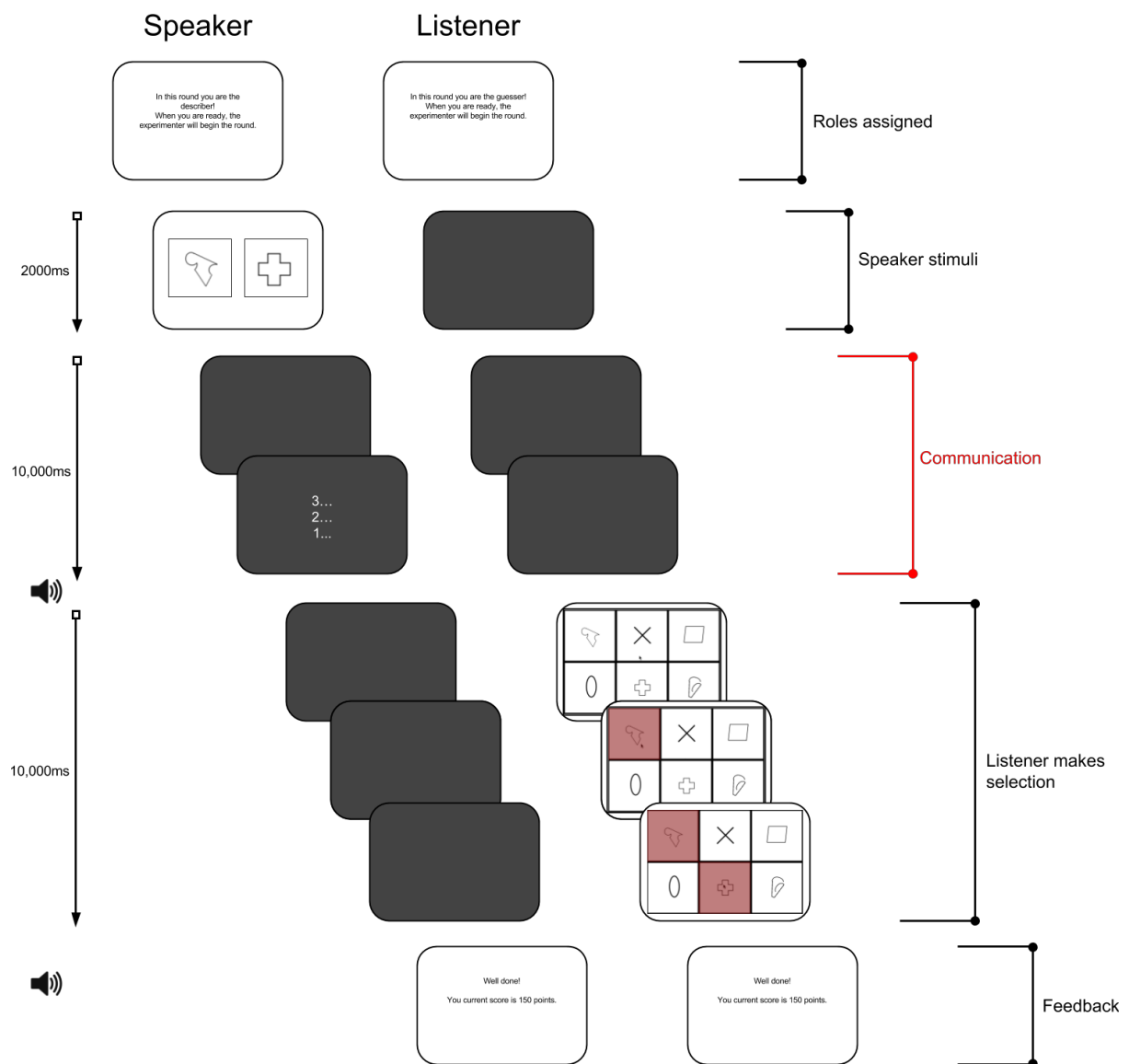


Figure 1. Procedure of a given trial.

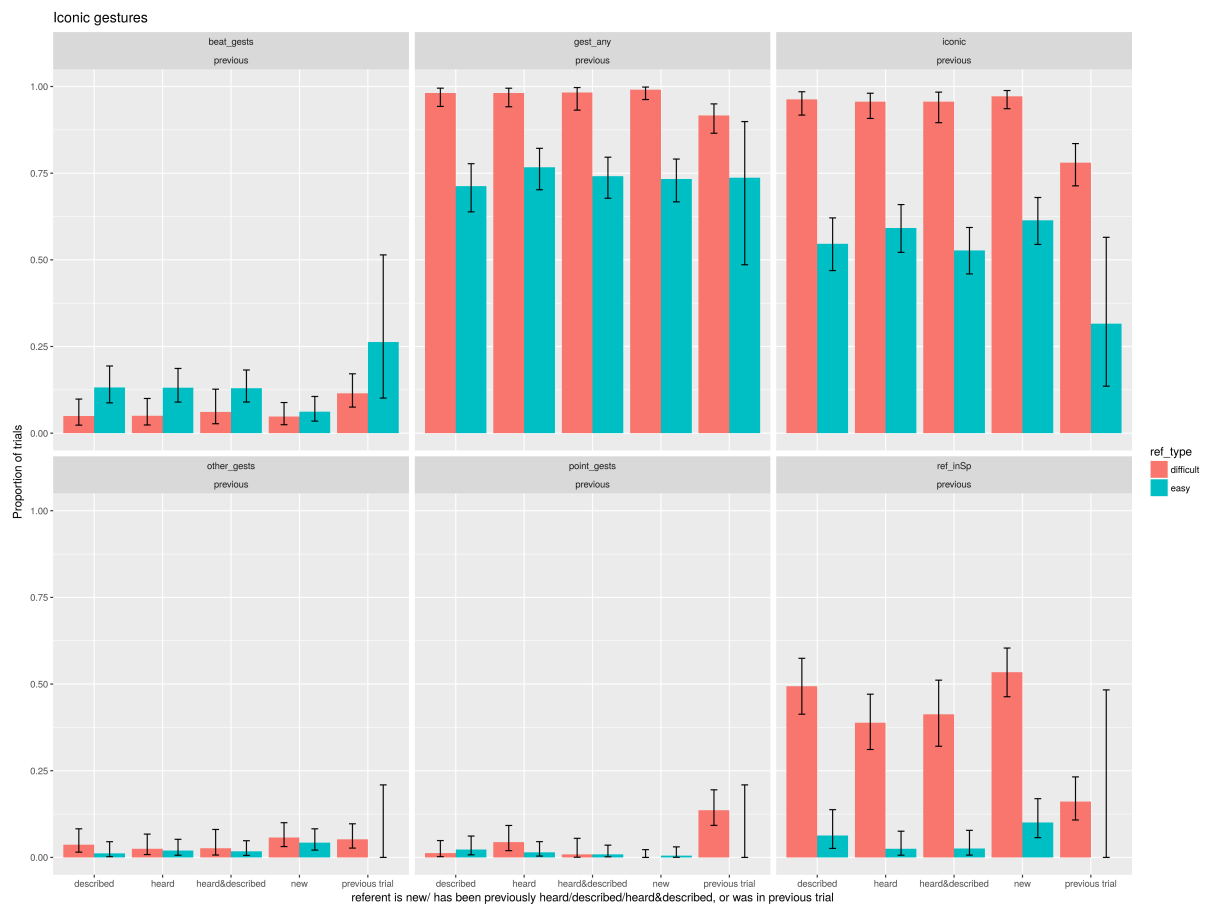


Figure 2. gestures