

Head, posture, and full-body gestures in dyadic conversations

Luboš Hládek^{1,2}, Bernhard U. Seeber¹

Audio Information Processing, Technical University of Munich, 80333 Munich, Germany

During writing at: Acoustics Research Institute, Austrian Academy of Science, 1010 Vienna, Austria

Corresponding Author: Luboš Hládek - ge67kip@mytum.de

Abstract

When face-to-face communication becomes effortful due to background noise and interfering talkers, the role of visual cues becomes increasingly important for communication success. While previous research has selectively investigated head or hand movements, here we explore the combination of movements of head, hand and the whole body in acoustically adverse conditions. We hypothesize that with increasing background noise level, the frequency of typical conversational movements of hand, head, trunk, and legs increases to support the speakers role while the listeners support their role by increased use of confirmative head gestures and head and trunk movements to increase the signal-to-noise ratio. We conducted a dyadic conversation experiment in which (n=8) normal hearing participants stood freely in an audiovisual virtual environment. The conversational movements were described by a newly developed labeling system for typical conversational movements, and the frequency of individual types was analyzed. Increased levels of background noise led to increased hand-gesture complexity and modulation of head movements without a clear pattern. People leaned forward slightly more and used less head movements during listening than during speaking. Additional analysis of hand-speech synchrony with hypothesized loss of synchrony due to the background noise showed a modest decrease of synchrony in terms of increased standard deviation at moderate sound levels. The results support previous findings in terms of the gesturing frequency, and we found a limited support for the changes in speech-gesture synchrony. The work reveals communication patterns of the whole body and exemplifies interactive communication in context of multimodal adaptation to communication needs.

Introduction

Human face-to-face communication relies not only on acoustic speech signals but also on the integration of sensory inputs with motor production. In daily conversations, the visual cues such as gestures or visual prosody cues make the communication easier and less effortful but it is not uncommon that the visual cues change the gist of message (for instance palms facing down or inwards mean something completely different when the palms face up or outwards) or even become critical for understanding (the thumb is pointing up or the thumb is pointing down). When the signal-to-noise ratio becomes very low, such as in noisy environments or acoustically adverse conditions, the motor expressions, especially lip movements, become the major source of information when people rely on lip reading (Hadley et al., 2019, 2022). In addition to lip reading, which brings strong speech intelligibility benefits (MacLeod & Summerfield, 1987), prosodic head movements (Munhall et al., 2004) help convey rhythm, stress, and structure of the utterance. Thus, people recruit compensation strategies in which the context, knowledge, and prosody are more heavily weighted to help disambiguate the missing acoustic information.

On the other hand, the role of hand gestures in communication in noisy environments is less well understood. The beat gestures carry redundant information, and they are the most typical hand movements in communication (Bergmann et al., 2011). The beat movements, like the head movements, indicate prominence in the phrase and regulate the communication by synchronizing with the prosodic peaks or stressed syllables; for instance, they directly influence the perception of sound, similar to the McGurk effect (Bosker & Peeters, 2021). Additionally, to these communicative functions and in contrast to the head movements, the hands and speech production systems are biomechanically coupled. The hand movements directly influence subglottal pressure and vocal intensity, thus gestures seem not to have only the conceptual role but rather are very physical. The studies showed that hand movements increase peak fundamental frequency, vowel space expansion, and vocal amplitude (Pouw et al., 2020; Pouw & Dixon, 2019), which may have a supporting effect for Lombard speech (Hodgson et al., 2007). In contrast to beat gestures, iconic, metaphoric, and deictic gestures carry explicit meaning (Bergmann et al., 2011), hence they emerge

from different cognitive mechanisms than the beat gestures. Yet, in communication in acoustically adverse conditions, they may serve a supporting role and hence might be more present in situations when the speech intelligibility is degraded.

Studies show that when speakers face degraded auditory feedback or increased task difficulty, gesture–speech synchrony becomes more variable and delayed (De Jonge-Hoekstra et al., 2021; Pouw & Dixon, 2019), which suggests that hand-speech synchrony changes when communication becomes more difficult. However, there are various aspects of the task that may interact with the hand-speech coupling at different stages. According to the ‘ballistic model’, gestures are coupled with speech at an early conceptualization stage of the gesture execution, while the later preparatory movements and the stroke, the main phase, are rather automatically executed. However, more recent studies showed interaction or the coupling is affected at the later stages (Chu & Hagoort, 2014; De Jonge-Hoekstra et al., 2021; Pouw & Dixon, 2019) even at the time of the execution of the stroke when the participants heard either a strong echo, the difficulty of instructions changed, or the task goal was interactively adapted during the task. The evidence thus suggests that gesturing is not driven only by an automatic process linked in one way to the speech production but is adapted even at the time of the execution and depends on the amount of available cognitive resources. Hence, if the interaction is maintained during the ongoing cognitive and biomechanical feedback loop, such as predicted by the entrainment hypothesis (Pouw & Dixon, 2019), the adaptive hand motion behavior during communication will be affected at the preparatory or executive stages.

Recent work on speech-gesture coordination (Trujillo et al., 2021) in a noisy environment found that with the increase in background noise level, speakers produced more gestures. However, the gesture duration, peak velocity, and vertical amplitude did not adapt correspondingly. Other studies indicated that combined visual speech and visible iconic gestures became more informative than visual speech without gestures compared to speech content when communication becomes difficult due to competing noise (Drijvers & Özyürek, 2017). Additionally, the whole-body movement may be affected, people lean towards the speakers to increase signal-to-noise ratio, lean back to adjust the social distance, change posture to indicate the openness or closeness, use palm orientations to indicate the attitude of the affect, and legs may

have a specific contribution to these adaptations. However, it is not clear which of these communicative aspects may be disrupted or enhanced by the cognitively demanding cocktail party situation. Thus, the gesturing is affected by the background noise. Yet, the nature of adaptation may vary with the task, overall level of background noise, individual cognitive capacity – knowledge of the topic, or relationship with the partner.

Despite these findings, the dynamics of gesture production during free face-to-face conversation in noise remain poorly understood. The previous work mainly relies on structured tasks, controlled utterances, or perception of phrases without context, which limit ecological validity. Free face-to-face conversation, by contrast, allows turn-taking, provides meaning with the head movement prosody, and adapts the gestures spontaneously in response to ongoing conversation. Gesture trajectories, synchrony with speech prosody, bodily emotional expansiveness, and posture shifts may all change when communication becomes effortful, yet empirical data on these adaptations in acoustically controlled environments are scarce.

In the present study, we aim to extend previous findings by examining how speakers adapt their hand movements during free, unstructured conversation in an immersive virtual reality environment that allows precise acoustic and visual control. Unlike structured tasks, conversational dialogue enables participants to spontaneously adjust gesture type, head position, timing, and posture in response to communicative needs. We hypothesize that as background noise increases and communication becomes more effortful, gesture–speech synchrony may be affected in terms of temporal alignment between gestures and prosodic anchors in line with the coupled interactive model of hand-speech synchronization (Pouw & Dixon, 2019). Additionally, we expect changes in the distribution of gesture types or complexity with increasing background noise levels (Trujillo et al., 2021) as a fraction of overall speech production time during conversation. The speakers are likely to benefit from more frequent or more complex hand movements, while the listeners may benefit from increased use of head or trunk movements that affect the signal-to-noise ratio. Hence, by analyzing multimodal behavior at multiple noise levels, we aim to characterize how interlocutors strategically deploy gestures as a compensatory mechanism under adverse auditory conditions.

Methods

Participants

Eight people participated in the experiment (four male, four female, 25.2 ± 1.3 years (mean \pm SD), min 23 years, max 27 years). None of them had known hearing problems and their pure-tone hearing thresholds were tested in octave frequencies from 250 Hz - 8kHz using a clinical audiometer (MADSEN Astera², type 1066, Natus Medical Denmark Ap, Denmark). Each threshold was equal to or below 20 dB HL. The participants were recruited from university students and through personal contacts. During experiment, participants had conversation in English which was their second language. Each participant provided written informed consent, and the procedures were approved by the Ethical Committee of the Technical University of Munich (65/18S).

Environment

The experiment took place inside an anechoic chamber (10 m x 6 m x 4 m; l x w x h) with the SOFE setup for virtual reality experiments with high-fidelity spatial audio. The audio system consisted of 61 loudspeakers mounted on a cube-shaped metal frame ($4.39\text{ m} \times 4.39\text{ m} \times 3.4\text{ m}$) – all pointing to the center. Thirty-six loudspeakers (Dynaudio BM6A mkII, Dynaudio, Skanderborg, Denmark) were distributed on a horizontal plane at approximate ear level (0° elevation), the remaining loudspeakers were distributed at elevations below and above the horizontal plane. The digital-to-analog converters (RME 32DA, Audio AG, Haimhausen, Germany) exciting the loudspeakers were controlled by a sound card (RME HDSPe, Audio AG, Haimhausen, Germany) in an external PC. The system further included four acoustically treated (32 dBA combined) video projectors (Barco F50 WQXGA, Barco, Kortrijk, Belgium) projecting on four acoustically transparent screens in front of the loudspeakers, creating a 360° projecting space with a floor area of 16 m^2 . An optical motion tracking system (OptiTrack Prime 17W, NaturalPoint Inc. Corvallis, Oregon, USA) recorded the body movements (v3.01, Motive:Body, Optitrack, NaturalPoint Inc. Corvallis,

Oregon, USA) of the participants at a 358.5 Hz sampling rate which was synchronized with the sound card's word clock output (eSync 2, NaturalPoint Inc. Corvallis, Oregon, USA).

Participants wore head-set microphones (VT 800, Voice Technologies, Switzerland, on DWR-R02DN and DWT-B01N digital wireless transmitters, Sony, Japan) connected to the soundcard via an AD converter (Micstasy, RME, Germany). Additionally, the participants had full-body motion-tracking suits with the Optitrack Entertainment Marker Set installed on them (Core+Passive Fingers - 54 markers). Participants were monitored during the experiment through a closed-circuit camera with an intercom by the experimenter sitting outside in the control room.

Audio-visual virtual environment

During the experiment, participants stood inside the SOFE projection area in audio-visual virtual space simulating the platform of an underground station (Hládek et al., 2021; Hladek & Seeber, 2022; van de Par et al., 2021). The acoustic simulation included reverberation of sources and their own speech ($RT_{60} = 1.68$, $DRR = 2.7$ dB, $EDT = 0.31$ s) and was created using the real-time Simulated Open Field Environment (rtSOFE) (Seeber et al., 2010; Seeber & Clapp, 2017; Seeber & Wang, 2021; T. Wang et al., 2022). The virtual environment was implemented in a game engine (Unreal Engine v5.2) with a custom visual model of the simulated underground station. The visual simulation was modelled according to the real underground station and included a running escalator (without escalator sound), train tracks, information boards, acoustic treatments on the side walls, an elevator, and other surroundings of the actual underground station. The model is freely available on Zenodo (Hladek & Seeber, 2022).

The game engine provided the visual projection, while the acoustic simulation of the reverberated speech of both participants in rtSOFE received positional information from the game engine (Enghofer et al., 2021). The game engine controlled the simulation, and the plugin provided spatial updates for the acoustic simulation. The acoustic simulation was frequently updated according to the position of the participants (serving to update own reverberation profile) provided by the motion tracking system. The first order and higher order reflections up to 100 ms of the simulated impulse response were simulated with

short latency using the image source method. Late reverberation (above 100 ms of the simulated impulse response) was obtained from a multi-channel recording of the environment (Hladek & Seeber, 2022). The details of acoustic modelling could be found elsewhere (Hládek et al., 2021). The speaker reverberation rendering further considered the directivity of human speech (Flanagan, 1960). The pick-up microphone of each speaker was spectrally equalized to the acoustic far-field for each participant prior to the experiment. The acoustic and visual simulation was referenced to the middle of the simulation area which corresponded to the position R1 defined for the underground scene (Hladek & Seeber, 2022; van de Par et al., 2021).

Experimental conditions

The experiment consisted of 30-minute-long conversations between two people standing on the platform of a simulated underground station. Broadband temporally modulated speech-shaped noise (Fastl, 1987) was played at different levels that changed every five minutes between these values: 0 dB SPL (No Noise, i.e. only the background noise of the video projection), 70 dB SPL (Medium), and 80 dB SPL (High). The order was pseudo randomized such that each condition took place twice. Reverberation added 2.3 dB; thus, the presentation levels of the noise conditions were 72.3 dB SPL and 82.3 dB SPL. The level of noise was calibrated by the loudspeaker calibration filters (from 100 Hz – 18 kHz) created with custom scripts, a measurement microphone (MM210, Microtech Gefell, Germany), and a microphone calibrator (Type 4231, Brüel & Kjær, Denmark). The levels were also verified with a portable sound level meter (XL2, NTi Audio, Schaan, Liechtenstein). Given the definitions of the underground scene (van de Par et al., 2021), the noise originated from azimuths 0°, 90°, 180°, -90° and simulated distance 1.6 m (note that the loudspeakers were physically at 2.1 m from the center). positions P1, P4, P7, P10 of the Underground scene 1, which corresponded to Each loudspeaker provided an independent sample of the noise that was generated from a custom-made script. The noise sources were reverberated and created by convolution of noise samples with corresponding multi-channel impulse response of the given position in the underground scene.

Procedures

The task of the experiment was to have a free dyadic conversation for thirty minutes. Before the main experiment, participants received information about the experiment and signed informed consents. Subsequently, pure-tone thresholds of each participant were measured using pure-tone audiometry in a sound-proof booth. Following that, participants moved to the anechoic chamber building where they changed into the full-body motion tracking suits. The tracking markers and head-worn microphones were installed and calibrated by two experimenters.

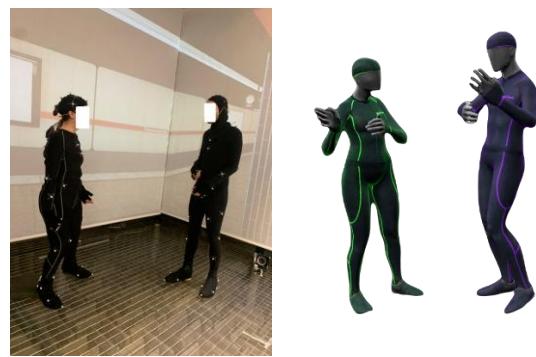


Figure 1 A) Two participants are conversing in the virtual underground station, simulated using rtSOFE inside an anechoic chamber. B) Reconstruction of the avatars (a different moment).

The calibration was done before the start of the experiment. The experimenters asked the participants to stand in the middle of the simulation area and look in the forward direction while one experimenter created the virtual avatar in the motion tracking software. Next, the head-worn microphones were calibrated individually by asking each participant to stand one meter from a calibrated $\frac{1}{2}$ " measurement microphone (MM210, Microtech Gefell, Germany) and speak for 20 seconds while looking directly at the measurement microphone. Based on the power spectra difference between the reference and personal microphone, a finite impulse response filter with an order of 512 taps was designed, which was then applied in the acoustic simulation.

Before the start of the experiment, the experimenter explained to the participants that they should talk freely on any chosen topic but avoid topics that could be potentially harmful. Given that the

preparations took around one hour, the participants already had a chance to get to know each other, hence none of them had problems chatting for another 30 minutes. Then, the experimenter left the room and initiated the experiment. During the experiment, participants continued a free, unscripted conversation in English for 30 minutes on a topic of their choice.

The whole experiment was conducted in one session, for which two participants were invited. The participants did not know each other before the experiment, but mutual knowledge was not an exclusion criterion. The experimenters did not notice problems in the conversation flow, possibly because the participants had already spent the preparation time together in one room and were already acquainted with each other. Another factor that positively contributed to the conversation flow was that some participants already knew each other from school courses, but knowing each other was not a condition for participation in the experiment. Some participants were complete strangers to each other before the experiment. After the experiment, the participants were debriefed regarding the purpose of the experiment.

Gesture annotation system

We developed a novel annotation system to annotate typical movements during conversations with the focus on beat gestures (Elkjær et al., 2022; Hadley et al., 2019; Harrigan, 1985; Kendon, 2004; Lausberg & Sloetjes, 2016; Leonard & Cummins, 2011; McNeill, 1995; Wagner et al., 2014) . The motivation of development of such system instead of using an already established one e.g. NEUROGES (Lausberg & Sloetjes, 2009, 2016) was the focus on gesticulation (co-speech movement without explicit meaning) during conversation of two or three standing participants where all body parts including postures are affected (e.g., turning torso, crossed legs), including leaning forward and backward behaviors, and having simple categories that relate to speech communication.

Body motion was categorized in short time frames separately for the speaker and the listener. The speaking activity labels had only two options: whether the person was speaking or not and was determined using the detectSpeech function (v24.2, MATLAB, Natick, MA) while the output was resampled to the

motion tracking sampling rate. The speech detection process was manually controlled, and misclassified periods of speaker silence due to cross-talk or unwanted noises were removed.

To be able to holistically and comprehensively analyze body movements during conversations, body motion analysis focused not only on head and hand movements, but on the entire body. Different body parts contribute distinctively to our expressions of intentions, states, and emotions, yet the combination creates a unique blend. Hence, we extended our scope from arm movements to head movements, trunk movements, and leg movements to capture the whole scale of movements and behavior in communication situations. This integrated framework enabled us to describe not only specific gestures but also broader postural changes, coordinated movements, and whole-body dynamics. We propose that such a complex approach plays a critical role in interpreting human behavior.

Arm movements and poses

The arm movements are a most prominent component of nonverbal communication (Wagner et al., 2014), conveying more information than speech. As arm contains seven degrees of freedom (Cheng et al., 2011; Z. Wang et al., 2018)—three at the shoulder, one at the elbow, one for forearm rotation, and two at the wrist—the range of possible movements is highly complex. In this study, we exclude finger movements and therefore consider a total of fourteen degrees of freedom across both arms. With the aim of describing typical movements in face-to-face communication situations, our analysis concentrates on the functional roles of hand and arm movements and uses their trajectories as a decisive cue. Based on this approach, we defined a structured set of gesture and posture categories that capture common expressive, emotional, and interactional movements of the body and the limbs.

Arm Movements	Description
Chopping/Clicking	Small, brief, and often rhythmic hand or wrist movements. They are often used to signal emphasis or highlight speech segments. These movements may function as a micro-empathy gesture that may express internal states or affection.

Circular Gesture	One or both hands swipe in a circular trajectory. Circular motions may illustrate continuity or the thinking process or accompany explanatory speech.
Complex Gesture	Both hands produce extended, multi-phase, or coordinated movements whose trajectories are not easily reduced to simple patterns defined by other gestures. These gestures often accompany expressive speech and can convey emphasis, uncertainty, affirmation, or other nuanced meanings.
Repetitive Gesture	Regular, repeated movement (often up-down or forward-back) of one or both hands, commonly serving as beat gestures that align with speech rhythm. In some contexts, repetitive motions may also express impatience or restlessness.
Palm-Inward Swipe	A large hand movement where the palm faces inward (toward body) or downward. This gesture may function as expressive or micro-empathetic gesture and may accompany explanations and express intentions like closeness, disagreement, or orders.
Palm-Upward Swipe	A hand movement similar to the palm-inward swipe, but performed with the palm facing upward. This gesture is widely associated with offering, requesting, questioning, or inviting further explanation. May be part of open posture. It may also mark openness, agreement, or acceptance.
Arm Postures	
Contractive Posture	A closed posture in which the hands rest in front of the body (e.g., arms crossed, hands clasped, or arms held close). The upper body appears more closed than in a neutral standing position.
Expansive Posture	An open posture with the arms extended outward, hands apart, or with the chest and torso widened. The body appears more open than when arms hang naturally.
Static Droop	A posture, usually standing still or ending a gesture to change into this posture, with hands hanging down naturally.
Self-Touching	Touching one's own body (e.g., chin, arms, thighs). Self-touch can signal cognitive processing, discomfort, anxiety, self-soothing, or mild itchiness.

Table 1 Labels of typical arm movements in face-to-face communication and their description.

Head movements

Head movements are a strong back-channel cue to keep the flow of communication. They often function for emphasizing, signaling attentiveness, and providing prosody. Biomechanically, head motion can be described through three rotational degrees of freedom—pitch, yaw, and roll—generated primarily by the rotation of the neck. Based on these fundamental movement axes, the present gesture labeling set is grounded

in the trajectories and communicative functions of head movements, informed by established research on nonverbal behavior and movement kinematics (Kendon, 2004; Sharma et al., 2018).

Head Movements	
Nodding	A common back-channel cue. A quick downward-then-upward movement of the head to signal reception and agreement. Additionally, nodding may express encouragement, or emphasis and can signal that the listener should continue speaking.
Shaking	A quick repetitive side-to-side movement of the head. This gesture typically indicates disagreement (culturally specific); however, it can also express confusion or uncertainty.
Tilting	A rotation of the head around the roll axis (ear-to-shoulder) to the left or right. Head tilts can convey interest, attentiveness, questioning, or emotional nuance, depending on context.
Turning	Rotation of the neck in the yaw axis, it can change the orientation of the face.
Up/Down	Rotation of the neck in the pitch axis to make the face towards up/down for a while. Different from nodding by motion velocity and communication function.
No Movement	The recorded person's head remains stationary or remains still without exhibiting any notable movements.

Table 2 Labels of typical head movements in face-to-face communication and their descriptions.

Trunk movement

Trunk movement involves coordinated motion of the waist, hips, and spine, allowing the upper body to rotate, bend, or lean. Trunk movement enables changes in posture, body orientation, and interpersonal distance during interaction, which is essential for increasing signal-to-noise ratio (Weisser & Buchholz, 2019) or head orientation benefit (Grange & Culling, 2016). The current set of trunk movements was selected according to their communicative functions. The movements may contribute to the expression of affiliation, social distancing, or simply may serve to improve speech intelligibility.

Trunk Movements

Turning Upper Body	A rotation of the trunk around the yaw axis, shifting the torso left or right and changing the orientation of the upper body. The movement may relate to other gestures (deictic) or may signal disengagement.
Leaning Forward	A forward tilt of the upper body is used to reduce the distance between the collocutors. The gesture may be associated with interest, affiliation, or serve to improve speech perception (both as a speaker or listener).
Leaning Backward	A backward tilt of the upper body that increases interpersonal distance. It may express discomfort, distancing, or simply adaptation of interpersonal space.
Leaning Side	A tilt of the upper body to the left or right. This may indicate comfort adjustments, or may signal shift in attention.
No Movement	A static posture with no noticeable upper-body movement during the observation period.

Table 3 Labels of typical trunk movements in face-to-face communication and their description.

Leg movement

Leg movement, like arm movement, can be described in terms of multiple degrees of freedom, with a single leg having seven rotational degrees of freedom across the hip, knee, and ankle joints. In everyday interaction, the communicative functions of leg movements tend to be more constrained than those of the arms. Here, we define a set of leg-movement labels based on the communication function. These behaviors contribute to posture, balance, and interpersonal signaling (Kendon, 2004; Lausberg & Sloetjes, 2016).

Leg Movements	
Standing On Tiptoe	Standing with the toes on the ground and the heels lifted.
Walking In Place	Minor stepping or lifting of the feet without shifting the body's overall position; often used to adjust posture or release restlessness.
Walking	Alternating leg movements that shift the body's center of mass horizontally, causing the person to change physical location.
Leg Raise	Lifting one foot off the ground while standing on the other leg.
Cross	Placing one foot adjacent to or crossing over the other (left crossing right or vice versa).
Full Turn	A rotation of the entire body by repositioning the feet and legs while keeping the upper body oriented until the rotation is complete.
No Movement	A static posture with no noticeable leg movements throughout the observed period.

Table 4 Labels of typical legs movements in face-to-face communication and their descriptions.

Gesture labeling

The motion tracking data served for gestural and temporal analysis. In the first step, we reconstructed virtual avatars from the data point clouds using the Optitrack software (v3.01, Motive:Body, Optitrack, NaturalPoint Inc. Corvallis, Oregon, USA). Videos of the reconstructed avatars and aligned speech served for gesture annotation using annotation software (*ELAN*, 2025; Lausberg & Sloetjes, 2009). When recording the audio, we marked the motion tracking timestamp at the beginning of the recording, which allowed precise temporal alignment of the motion tracking data and the audio signal, hence the annotator listened to the conversation and watched reconstructed videos during the annotation.

The labeling was done by an annotator, who marked the beginning and ending of each gesture, including a pre-stroke hold, post-stroke hold, and retraction phase (Pouw & Dixon, 2019; Wagner et al., 2014). While labeling the four categories (arm gestures and postures were the same category), the annotator listened to the speech and watched the reconstructed video and was instructed to mark the whole phrase as a gesture - not only a stroke (Pouw & Dixon, 2019).

We analyzed the relative time of gesturing, which was determined as a proportion of time during one's own or the other collocutor's speech within a 5-minute interval. The values were transformed into a logit scale to account for the non-normal distribution of proportion values. Because of this transformation, we replaced the value with a value of 0.005 (half a percent). Due to a technical error, the legs and trunk motion were analyzed only for six participants.

Hand-speech synchrony analysis

We analyzed synchrony between the peaks of the pitch and the peaks of the speaker's hand movement velocity in beat gestures using a procedure motivated by the previous work (Pouw & Dixon, 2019) in a subset of arm gestures: *expansive gesture, complex gesture, circular gesture, palm-inward swipe gesture, palm-upward swipe gesture, chopping/clicking gesture, repetitive gesture*, since these represented primarily the beat moves. The pitch traces were extracted using PRAAT (Boersma & Weenink, 2017) and temporally aligned with the motion tracking data. The pitches were extracted for the frequency range 51

Hz – 400 Hz and resampled to the motion tracking sampling frequency. The motion tracking data – hand position vectors – were filtered with a 10-Hz fourth-order Butterworth low-pass filter prior to computation of the first derivative.

The speech-hand synchrony was analyzed from the distribution of the hand-speech delay

$$d = t_{P_{max}} - t_{G_{max}} \text{ [ms]} \quad (\text{Eq. 1})$$

where $t_{P_{max}}$ corresponds to the time of pitch peak, which was a simple maximum of the F0 trace within the gesture duration, and $t_{G_{max}}$ corresponds to the time of the most prominent gesture. The most prominent peak was found for a set of candidate peaks within a two-second interval around the pitch peak (using MATLAB function *findpeaks*). Statistics d was then defined as the signed time difference between the pitch peak and gesture stroke with the highest peak velocity.

$$d = \arg \min_{t \in [t_{P_{max}} - 1, t_{P_{max}} + 1] \cap [dur_G]} |t - t_{P_{max}}| \quad (\text{Eq. 2})$$

such that

$$v_G(t) = \max_{t' \in [t_{P_{max}} - 1, t_{P_{max}} + 1] \cap [dur_G]} \left(\left| \frac{dG(t')}{dt'} \right| \right) \quad (\text{Eq. 3})$$

Here G corresponds to the gesture trace and $v_G(t)$ is the gesture velocity. The two-second interval was chosen based on the distribution ranges observed by (Pouw & Dixon, 2019). We assumed that delays above 1 s do not represent a synchronous relationship.

Results

The relative frequency of the individual gestures was analyzed across the three conditions as a proportion of the time of the given gesture relative to the time when the gesticulator was speaking (Gestures As Speaker) and relative to the time when the co-locutor was speaking (As Listener).

Hand gestures

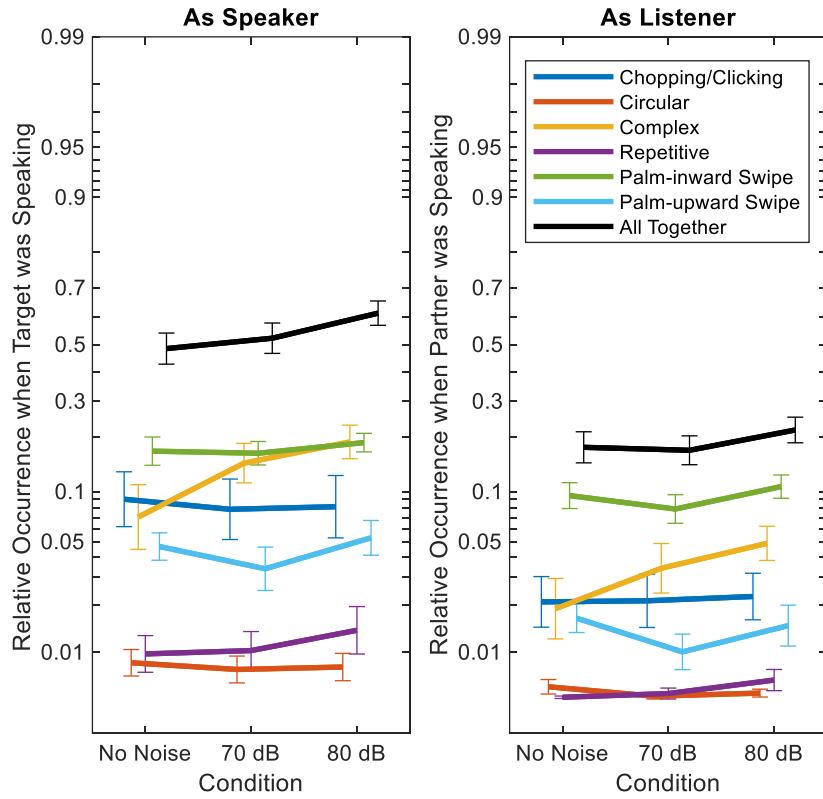


Figure 2 Hand gesture usage as a function of increasing background noise level (x-axis). The y-axis corresponds to the relative amount of time people were gesturing during speaking (Left: As Speaker) or when the conversation partner was speaking (Right: As Listener). Color codes individual gesture types (see legend), the black line corresponds to the relative time when either of the selected gestures (in the legend) were performed. The proportional data were Logit transformed before plotting, and the y-axis was then rescaled to show back-transformed proportions in this and the following graphs. Error bars represent the standard-error of the mean (here and following graphs).

Figure 2 shows the usage of different hand gestures under varying noise levels of background noise in two modes: either during speaking (Left) or listening (Right). The data show proportions of time for each gesture that represented beat gestures. The data were transformed into Logit scale before plotting and for statistical analysis, the y-axis shows proportions (e.g., 0.9 corresponds to 90%).

Statistically significant differences were found between the hand gesture types (Table 5: main effect of Type) modulated by the background noise (Table 5: interaction Condition x Type). Palm-inward swipes and complex gestures were more frequent than palm-upward swipes (pair comparisons using t-test results in $p < 0.05$ after Bonferroni correction), and palm-inward swipes were more frequent than chopping/clicking gestures ($p < 0.05$ after Bonferroni correction) while repetitive and circular gestures (significantly different

from all other gestures at $p<0.05$ after Bonferroni correction) were rarely present. Complex gestures increased with the background noise, while the other gestures were affected to only a small extent.

As expected, people gestured significantly more during speaking than listening (Table 5: main effect of Mode), and this was evident for all observed types of gestures, however, the increase during speaking relative to listening was more pronounced for the complex gestures, than for instance, circular gestures.

Source	F value	p value	p-adjusted	Significance	Partial ω^2
Mode	123.96	0.0000	0.0000	***	0.9461
Type	32.61	0.0000	0.0000	***	0.8187
Mode x Type	10.86	0.0000	0.0003	***	0.5848
Condition x Type	3.02	0.0031	0.0248	*	0.2239

Significance levels: *** $p<0.001$ * $p<0.05$

Table 5 ANOVA of the relative occurrence of hand gestures.

Hand postures

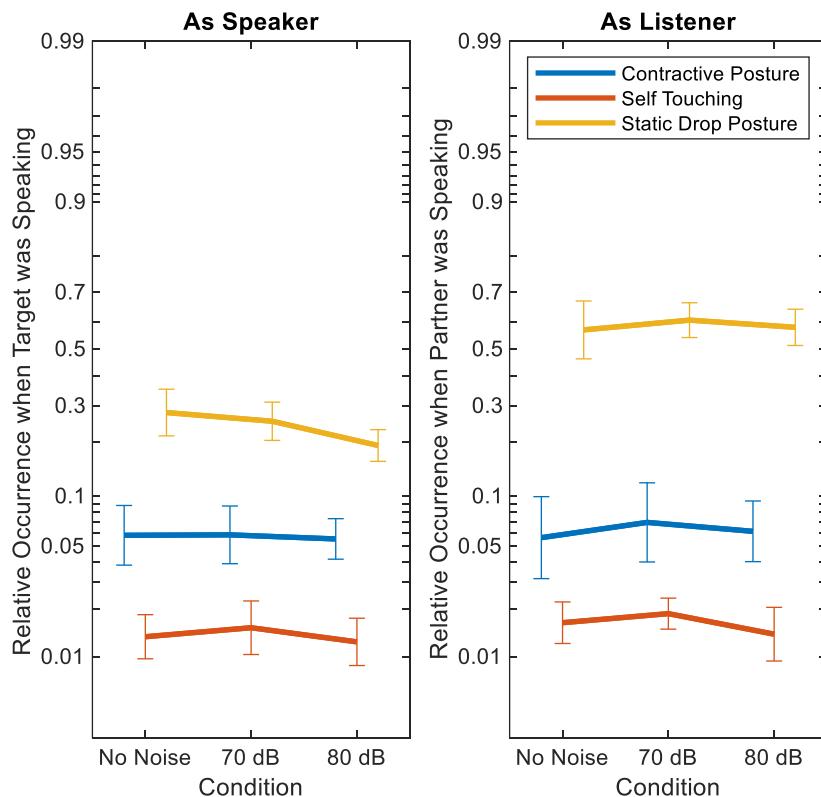


Figure 3 Hand postures as a function of increasing noise level (x-axis).

Figure 3 analyzes hand postures (which were in the category of hand movements) during speaking (left) and listening (right). Color codes posture type (see legend), x-axis shows condition.

The static drop posture was the most common occurring more often than contractive posture or self-touching (Table 6: main effect of Type; pairwise post-hoc comparison using t-test showed statistically significant differences at level $p<0.05$ after Bonferroni correction). Its frequency of occurrence decreased during speaking compared to listening (Table 6: interaction Mode x Type) complementing the increase in gesturing during speaking.

Source	F value	p value	p-adjusted	Significance	Partial ω^2
Mode	16.95	0.0045	0.0045	**	0.6950
Type	32.04	0.0000	0.0000	***	0.8160
Mode x Type	14.77	0.0004	0.0021	**	0.6630

Significance levels: ** <0.01 *** <0.001

Table 6 ANOVA of the relative occurrence of hand postures.

Head gestures

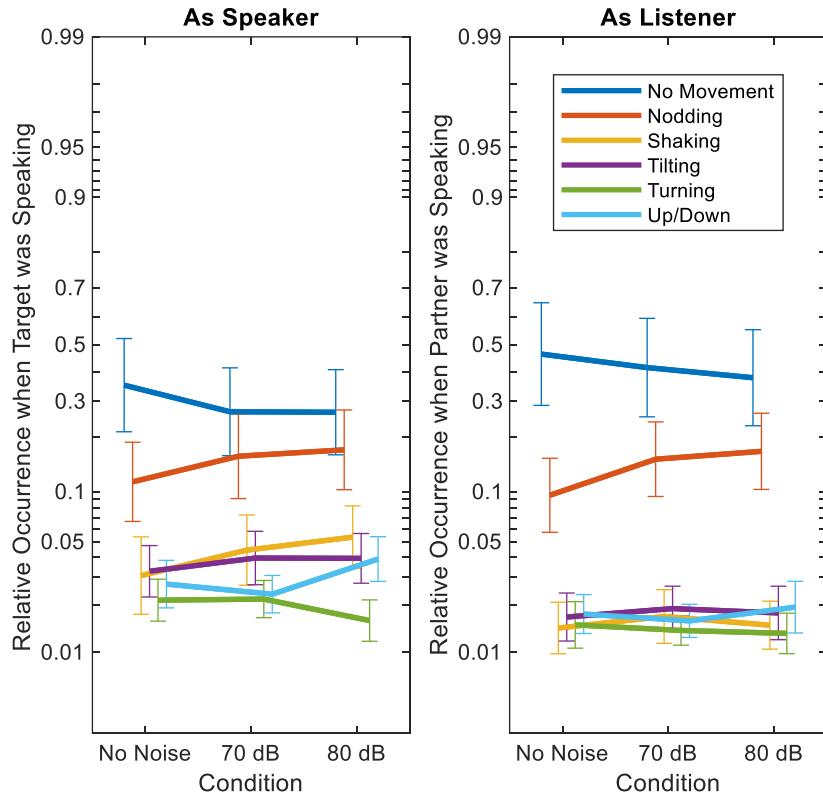


Figure 4 Head gestures as a function of increasing noise level (x-axis). The y-axis corresponds to the relative amount of time people performed the posture during speaking (Left- As Speaker) or when the conversation partner was speaking (As Listener). Color codes posture types (see legend).

Figure 4 shows frequency analysis head postures during speaking (left) and listening (right). Nodding was the most frequent head movement. Other gesture types (shaking, tilting, turning, up/down movements) were rarely present during listening, but more often present during speaking, as evidenced by the statistically significant interaction of Mode and Type (Table 7), while both main effects were also significant. Head gesturing type was also modulated by the presence of the background noise (Table 7: Condition x Type interaction). The interaction was driven by the increase of turning complemented by a decrease in up/down motion at the high background noise level.

Source	F value	p value	p-adjusted	Significance	Partial ω^2
Mode	19.94	0.0029	0.0029	**	0.7301
Type	21.13	0.0000	0.0000	***	0.7420
Mode x Type	9.36	0.0000	0.0011	**	0.5442
Condition x Type	4.00	0.0002	0.0091	**	0.2997

Significance levels: ** <0.01, *** <0.001

Table 7 ANOVA of the relative occurrence of head gestures.

Trunk postures

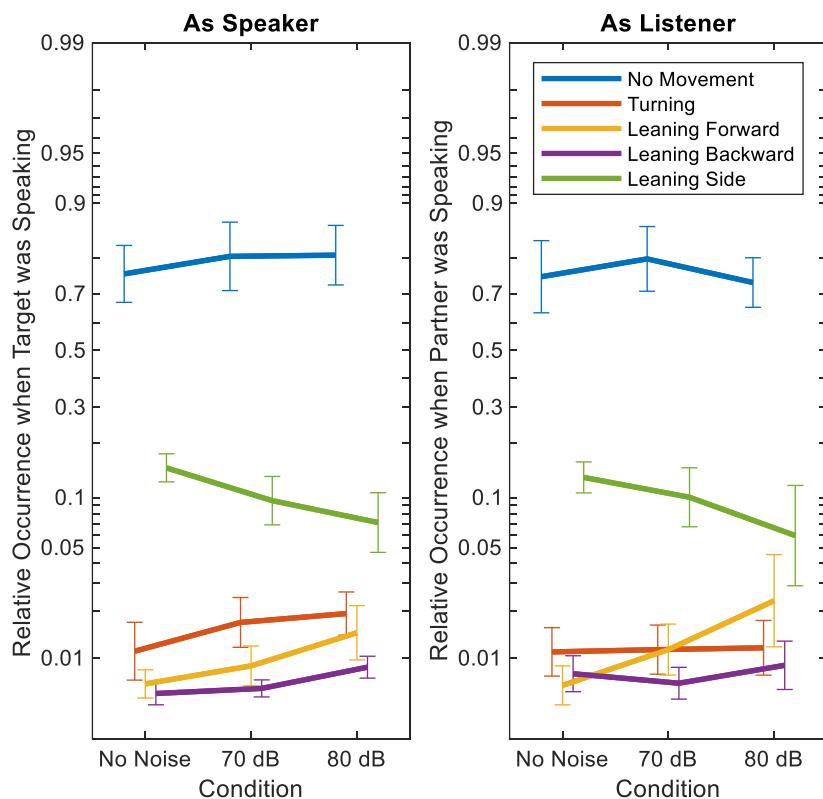


Figure 5 Trunk movements for different posture types (see legend) as a function of increasing noise level (x-axis).

Figure 5 shows the analysis of trunk movements during speaking (left) or listening (right). Leaning side was the most frequent trunk motion during conversation, significantly more than leaning forward and leaning backward (Table 8: main effect of Type, the mentioned pairwise comparisons were statistically

significant at $p < 0.05$ after Bonferroni correction). We saw a tendency to increase the use of leaning behavior (forward or backward) with the increased noise level, but the interaction was not statistically significant (Table 8: Condition x Type).

Source	F value	p value	p-adjusted	Significance	Partial ω^2 (bottom table)
Type	58.58	0.0000	0.0000	***	0.9201
Condition x Type	2.12	0.0561	0.1751	n.s.	0.1833

Significance levels: *** < 0.001

Table 8 ANOVA of the relative occurrence of trunk movements.

Legs movements

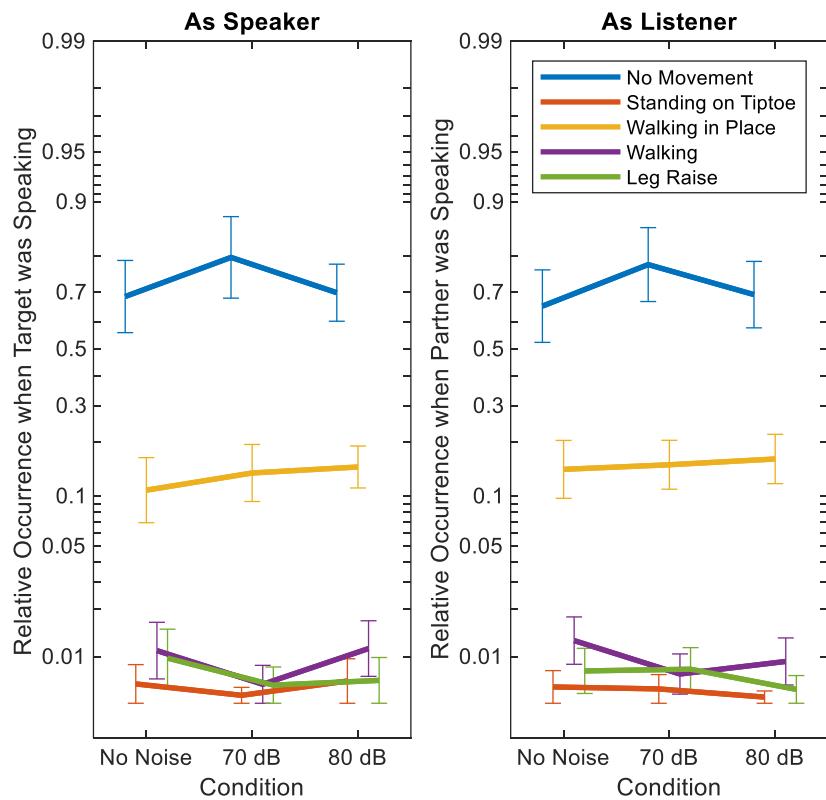


Figure 6 Legs movements (see legend for type) as a function of increasing noise level (x-axis).

Figure 6 shows the analysis of leg movements as a function of the background noise level (x-axis) and mode (As Speaker – left, As Listener - right). Walking in Place was the most common type of movement while the other types were rare (usually less than 1%) (Table 9: main effect of Type). Neither the mode or the background sound level significantly affected leg movement frequency.

Source	F value	p value	p-adjusted	Significance	Partial ω^2
Type	59.11	0.0000	0.0000	***	0.9208
Mode x Condition	3.42	0.0739	0.1035	n.s.	0.3259
Condition x Type	2.10	0.0588	0.1590	n.s.	0.1802

Significance levels: *** <0.001

Table 9 ANOVA of the relative occurrence of legs movements.

Hand-speech synchrony

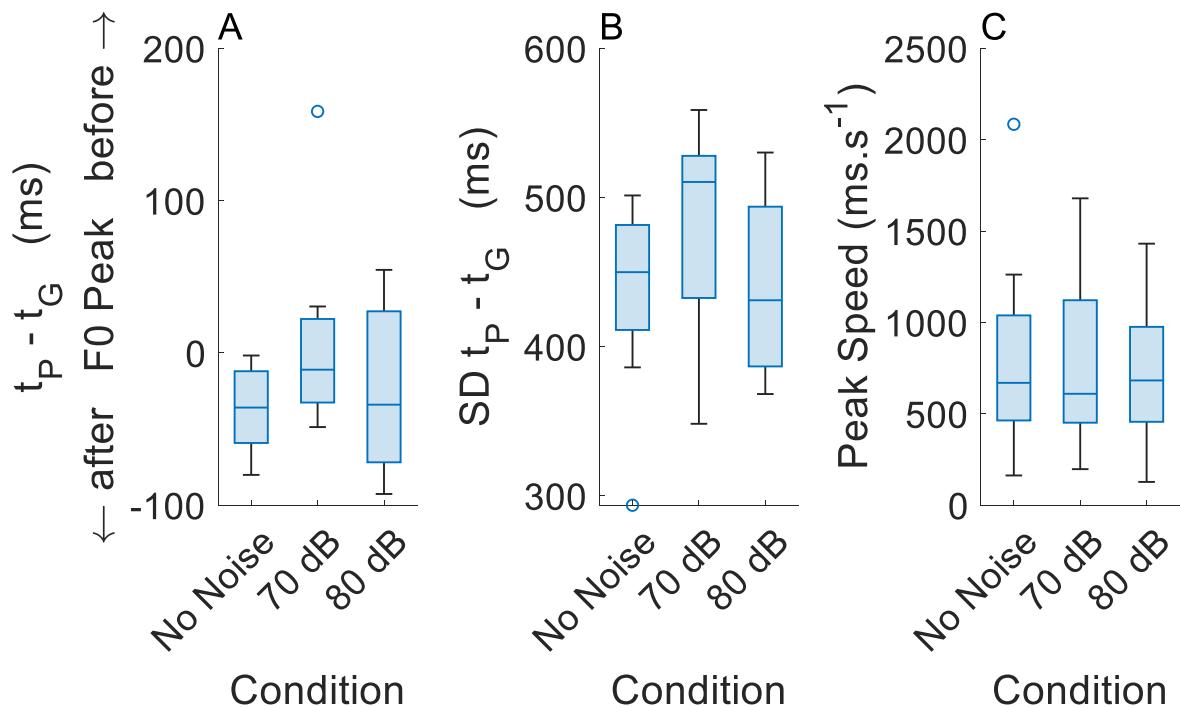


Figure 7 The box-plot analysis of speech-hand synchrony of talker's beat gestures for different background noise levels (x-axis). (A) Across-subject mean value of statistics d. Values above 0 indicate that the gesture precedes the F0 peak, values below 0 indicate that the dominant gesture lags the F0 peak (B) Across-subject value of the standard deviation of statistics d. (C) Peak speed of the dominant movement.

The coupling of speech with hand gestures was tested using the analysis of the distribution of statistics *d*. Figure 7A shows the across-subject distribution of mean values of statistics *d*, panel B shows the across-subject distribution of the standard deviation of statistics *d*, and panel C shows the peak values of the gestural movements that were used in the creation of the statistics *d*. Three box plots are displayed for three conditions of the background noise level in each of the panels.

The timing between hand gesture peaks and speech pitch peaks (synchrony) did not show statistically significant changes across noise levels, though variability approached significance. Peak gesture speed was also unaffected by noise. (Table 10).

Source	F value	p value	p-adjusted	Significance	Partial ω^2
Condition	3.54	0.0572	0.0696	n.s	0.2659

Table 10 ANOVA of the standard deviation of the *d* statistics.

Discussion

The results demonstrate that when background noise during conversation increases, people produce more complex hand gestures. Although speakers use more gestures overall than listeners, the increase in gesture complexity was evident in both roles. In contrast to expectations that listeners might benefit from increased head and trunk movements, listening was generally associated with less frequent gesturing. However, head movements did adapt to rising noise levels: up-down head movements increased during speaking at the highest noise levels, aligning with the observed increase in hand-gesture complexity. Walking in place was the most common leg movement, while other leg-movement types were rare and did not vary with noise level or conversational role.

The analysis of gesture quality, in terms of hand-speech synchrony, did not reveal statistically significant modulation of the *d* statistic. However, there was a trend toward increased variability (standard

deviation) of the d statistic at moderate noise levels, although this effect did not reach statistical significance.

The present data agree with the previous report of Trujillo et al., (2021). Increased levels of background noise led to an increase in gesturing while the quality of gesturing remained rather unaffected in terms of peak velocity. The present study extends the findings to free unstructured conversations rather than structured tasks and hence supports the idea of multimodal Lombard effect. We also showed that gesturing became more complex and investigated whole-body movements, which were not controlled for in the previous study. The conversation function of the complex gesture category was rather broadly defined, but still, it is likely that the speakers provided more supportive information to supplement missing acoustic cues. Even if the current analysis did not reveal a more specific nature of the complex gesture categories, the results seem in line with the increased utility of iconic gestures when background noise levels are increased (Drijvers & Özyürek, 2017).

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 352015383 – SFB 1330 C5. During write-up of the article, Luboš Hládek was supported by the Seal of the Excellence Fellowship of the Austrian Academy of Sciences. Yang Jiao contributed to the development of the annotation system in his Master’s thesis: Automatic Analysis of Body Movements and Gestures in Free Conversations. Technical University of Munich. Cem Kaya helped with annotations and setting up the virtual environment. We thank all participants of the experiments.

References

- Bergmann, K., Aksu, V., & Kopp, S. (2011). The relation of speech and gestures: Temporal synchrony follows semantic synchrony. *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011)*.

Boersma, P., & Weenink, D. (2017). *Praat: Doing phonetics by computer* [Computer software].

<http://www.praat.org/>

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B: Biological Sciences*, 288(1943), 20202419.

<https://doi.org/10.1098/rspb.2020.2419>

Cheng, X., Zhou, Y., Zuo, C., & Fan, X. (2011). Design of an Upper Limb Rehabilitation Robot Based on Medical Theory. *Procedia Engineering*, 15, 688–692. <https://doi.org/10.1016/j.proeng.2011.08.128>

Chu, M., & Hagoort, P. (2014). Synchronization of speech and gesture: Evidence for interaction in action. *Journal of Experimental Psychology: General*, 143(4), 1726–1741.

<https://doi.org/10.1037/a0036281>

De Jonge-Hoekstra, L., Cox, R. F. A., Van der Steen, S., & Dixon, J. A. (2021). Easier Said Than Done? Task Difficulty's Influence on Temporal Alignment, Semantic Similarity, and Complexity Matching Between Gestures and Speech. *Cognitive Science*, 45(6).

<https://doi.org/10.1111/cogs.12989>

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

ELAN (Version 7.0). (2025). [Computer software]. Max Planck Institute for Psycholinguistics, The Language Archive. <https://archive.mpi.nl/tla/elan>

Elkjær, E., Mikkelsen, M. B., Michalak, J., Mennin, D. S., & O'Toole, M. S. (2022). Expansive and Contractive Postures and Movement: A Systematic Review and Meta-Analysis of the Effect of Motor Displays on Affective and Behavioral Responses. *Perspectives on Psychological Science*, 17(1), 276–304. <https://doi.org/10.1177/1745691620919358>

Enghofer, F., Hladek, L., & Seeber, B. U. (2021). An “Unreal” Framework for Creating and Controlling Audio-Visual Scenes for the rtSOFIE. In H. Waubke & P. Balazs (Eds.), *Fortschritte der Akustik—DAGA '21* (pp. 1217–1220). Deutsche Ges. für Akustik e.V. (DEGA).

- Fastl, H. (1987). A background noise for speech audiometry. *Audiol. Acoustics*, 26, 2–13.
- Flanagan, J. L. (1960). Analog Measurements of Sound Radiation from the Mouth. *The Journal of the Acoustical Society of America*, 32(12), 1613–1620. <https://doi.org/10.1121/1.1907972>
- Grange, J. A., & Culling, J. F. (2016). Head orientation benefit to speech intelligibility in noise for cochlear implant users and in realistic listening conditions. *The Journal of the Acoustical Society of America*, 140(6), 4061–4072. <https://doi.org/10.1121/1.4968515>
- Hadley, L. V., Brimijoin, W. O., & Whitmer, W. M. (2019). Speech, movement, and gaze behaviours during dyadic conversation in noise. *Scientific Reports*, 9(1), 10451. <https://doi.org/10.1038/s41598-019-46416-0>
- Hadley, L. V., Naylor, G., & Hamilton, A. F. D. C. (2022). A review of theories and methods in the science of face-to-face social interaction. *Nature Reviews Psychology*, 1(1), 42–54. <https://doi.org/10.1038/s44159-021-00008-w>
- Harrigan, J. A. (1985). Self-touching as an indicator of underlying affect and language processes. *Social Science & Medicine*, 20(11), 1161–1168. [https://doi.org/10.1016/0277-9536\(85\)90193-5](https://doi.org/10.1016/0277-9536(85)90193-5)
- Hládek, L., Ewert, S. D., & Seeber, B. U. (2021). Communication Conditions in Virtual Acoustic Scenes in an Underground Station. *2021 Immersive and 3D Audio: From Architecture to Automotive, I3DA 2021*, 1–8. <https://doi.org/10.1109/I3DA48870.2021.9610843>
- Hladek, L., & Seeber, B. U. (2022). *Underground station environment*. <https://doi.org/10.5281/zenodo.5532643>
- Hodgson, M., Steininger, G., & Razavi, Z. (2007). Measurement and prediction of speech and noise levels and the Lombard effect in eating establishments. *The Journal of the Acoustical Society of America*, 121(4), 2023–2033. <https://doi.org/10.1121/1.2535571>
- Kendon, A. (2004). *Gesture Visible Action as Utterance* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511807572>
- Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods*, 41(3), 841–849. <https://doi.org/10.3758/BRM.41.3.841>

- Lausberg, H., & Sloetjes, H. (2016). The revised NEUROGES–ELAN system: An objective and reliable interdisciplinary analysis tool for nonverbal behavior and gesture. *Behavior Research Methods*, 48(3), 973–993. <https://doi.org/10.3758/s13428-015-0622-z>
- Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10), 1457–1471. <https://doi.org/10.1080/01690965.2010.500218>
- MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, 21(2), 131–141.
- McNeill, D. (1995). *Hand and mind: What gestures reveal about thought*. The Universoty of Chicago Press.
- Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual Prosody and Speech Intelligibility. *Psychological Science*, 15(2), 133–137. <https://doi.org/10.1111/j.0963-7214.2004.01502010.x>
- Pouw, W., & Dixon, J. A. (2019). Entrainment and Modulation of Gesture–Speech Synchrony Under Delayed Auditory Feedback. *Cognitive Science*, 43(3). <https://doi.org/10.1111/cogs.12721>
- Pouw, W., Harrison, S. J., Esteve-Gibert, N., & Dixon, J. A. (2020). Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *The Journal of the Acoustical Society of America*, 148(3), 1231–1247. <https://doi.org/10.1121/10.0001730>
- Seeber, B. U., & Clapp, S. W. (2017). Interactive simulation and free-field auralization of acoustic space with the rtSOFE. *The Journal of the Acoustical Society of America*, 141(5), 3974–3974. <https://doi.org/10.1121/1.4989063>
- Seeber, B. U., Kerber, S., & Hafter, E. R. (2010). A system to simulate and reproduce audio–visual environments for spatial hearing research. *Hearing Research*, 260(1–2), 1–10. <https://doi.org/10.1016/j.heares.2009.11.004>
- Seeber, B. U., & Wang, T. (2021). *Real-time Simulated Open Field Environment (rtSOFE) software package* [Computer software]. <https://doi.org/10.5281/zenodo.5648304>

Sharma, M., Ahmetovic, D., Jeni, L. A., & Kitani, K. M. (2018). Recognizing Visual Signatures of Spontaneous Head Gestures. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 400–408. <https://doi.org/10.1109/WACV.2018.00050>

Trujillo, J., Özyürek, A., Holler, J., & Drijvers, L. (2021). Speakers exhibit a multimodal Lombard effect in noise. *Scientific Reports*, 11(1), 16721. <https://doi.org/10.1038/s41598-021-95791-0>

van de Par, S., Ewert, S. D., Hladek, L., Kirsch, C., Schütze, J., Llorca-Bofí, J., Grimm, G., Hendrikse, M. M. E., Kollmeier, B., & Seeber, B. U. (2021). *Auditory-visual scenes for hearing research*. <http://arxiv.org/abs/2111.01237>

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57, 209–232. <https://doi.org/10.1016/j.specom.2013.09.008>

Wang, T., Lambacher, M., & Seeber, B. U. (2022). Extension of the real-time Simulated Open Field Environment for fast binaural and non-isotropic reverberation rendering. *Fortschritte Der Akustik - DAGA '22*, same issue.

Wang, Z., Cai, Z., Cui, L., & Pang, C. (2018). Structure Design And Analysis Of Kinematics Of An Upper-limbed Rehabilitation Robot. *MATEC Web of Conferences*, 232, 02033. <https://doi.org/10.1051/matecconf/201823202033>

Weisser, A., & Buchholz, J. M. (2019). Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions. *The Journal of the Acoustical Society of America*, 145(1), 349–360. <https://doi.org/10.1121/1.5087567>

Yang, J. (2023). *Automatic Analysis of Body Movements and Gestures in Free Conversations*. Technical University of Munich.