This paper was presented at the 6th Gesture and Speech in Interaction Conference that was held at Paderborn University, Germany from September 11-13, 2019.

To cite this paper:

Pouw, W., Paxton, A., Harrison, S. J., & Dixon, J. A. (2019). Acoustic specification of upper limb movement in voicing. In: Grimminger, A. (Ed.): *Proceedings of the 6th Gesture and Speech in Interaction – GESPIN 6* (pp. 68-74). Paderborn: Universitaetsbibliothek Paderborn. doi:10.17619/UNIPB/1-812

Acoustic specification of upper limb movement in voicing

Wim Pouw[1,2], Alexandra Paxton[1,3], Steven J. Harrison[1], and James A. Dixon[1]

[1]Center for the Ecological Study of Perception and Action, University of Connecticut
[2]Department of Psychology, Education, & Child Studies, Erasmus University Rotterdam
[3]Department of Psychological Sciences, University of Connecticut Affiliation

wimpouw@uconn.edu, alexandra.paxton@uconn.edu, steven.harrison@uconn.edu, james.dixon@uconn.edu

Abstract

Hand gestures communicate through the visual information created by movement. Recently, we found that there are also direct biomechanical effects of high-impetus upper limb movement on voice acoustics. Here we explored whether listeners could detect information about movement in the voicing of another person. In this exploratory study, participants listened to a recorded vocalizer who was simultaneously producing low-(wrist movement) or high-(arm movement) impetus movements at three different tempos. Listeners were asked to synchronize their own movement (wrist or arm movement) with the vocalizer. Listeners coupled with the frequency of the vocalizer arm (but not wrist) movements, and showed phase-coupling with vocalizer arm (but not wrist) movements. However, we found that this synchronization occurred regardless of whether the listener was moving their wrist or arm. This study shows that, in principle, there is acoustic specification of arm movements in voicing, but not wrist movements. These results, if replicated, provide novel insight into the possible interpersonal functions of gesture acoustics, which may lie in communicating bodily states.

1. Introduction

A conundrum in gesture studies is that gestures are often recruited by a gesturer who knows full well that gestures will never visually reach the listener. For example, during phone conversations, we do not stop gesturing (Bavelas, Gerwing, Sutton, & Prevost, 2008). Even speakers with congenital blindness gesture to listeners who also are blind since birth (Iverson & Goldin-Meadow, 2001).

Here we explore the possibility that visual information from gesture is but one of its (communicatively meaningful) products. Recently, we have found that upper-limb movements with relatively high physical impetus produce prominent but non-intentional changes in voice quality (Pouw, Harrison, & Dixon, 2018b). Specifically, we found peaks in the fundamental frequency (F0) and the amplitude envelope of continuous phonation of the vowel /a:/ when participants made high-impact movements that recruited the entire arm but not when producing low-impact wrist movements or when standing still (see Figure 1). Such peaks in phonation were observed at the moment at which posturally destabilizing forces of the arm movements were highest and at which the body counteracted such forces by tensioning of the muscles in anticipatory fashion. These results accommodate findings as observed in naturalistic contexts. Namely, sudden increases in speech intensity and fundamental frequency are key properties that define the prosody of speech, and spontaneous co-speech gestures are known to synchronize with such prosodic aspects of speech (Wagner, Malisz, & Kopp, 2014). Scaling up to natural speech, other work has found that infants' babbling becomes more adult-like in voice quality when infants simultaneously and rhythmically move their arms (Ejiri & Masataka, 2001) and that encouraging gesture production during adults' speech production boosted maximum observed F0 and intensity of speech (Cravotta, Busà, & Prieto, 2018).

We could wonder therefore whether there is information in speech acoustics specifying bodily gestures. Note, Hoetjes et al. (2004) found no statistically significant changes in acoustics when participants were (restrained) from gesturing, nor were listeners able to detect whether

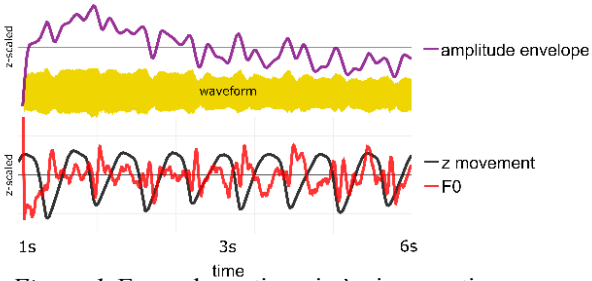


Figure 1. Example motion vis-à-vis acoustics.

someone was gesturing based on listening to their speech. However, we could argue mixed results might be obtained by averaging acoustic metrics over time (cf. Hoetjes et al.) and inferences about acoustics and gesture might be obscured when not taking into account physical impetus of gestures.

Since gestures—especially of the more forcefully beat-like kind (see e.g., https://osf.io/29h8z/)—affect voice acoustics we should assess whether listeners can pick up information about a gesturer's movement. The idea that there is acoustic information that specifies an object or event in the environment is actually non-controversial in ecological psychology of language (Fowler, 1986) and object perception (Carello, Wagman, & Turvey, 2005). Namely, Carol Fowler famously asserted that we do not hear speech sounds that we need to translate into meaningful objects of language perception but that we directly hear the cause of the sound - the articulatory gestures. Evidence for this includes studies on the McGurk effect, in which otherwise ambiguous speech sounds are disambiguated by visually or even manually perceiving the articulatory gesture (Fowler & Dekle, 1991). Furthermore, a line of research in ecological acoustics has shown that properties of objects (e.g., object length, object thickness; relative position) can actually be directly perceived by attuning to acoustic properties of the objects (Carello, Anderson, & Kunkler-Peck, 1998).

The current idea that we can hear bodily gestures is then complementary to these Gibsonian perspectives (Gibson, 2014). However, we are after a direct specification of bodily action in speech acoustics. If such specification exists to some degree, this would open up the investigation into whether bodily gestures' communicative function lies in part in its direct linkage with speech acoustics. We have a long experimental road ahead before we could conclude that gesture acoustics serve such a communicative role in a manner similar to the visual information created by gesture. Indeed, to date there is simply no evidence that humans can hear gesture (Hoetjes, Krahmer, & Swerts, 2014).

In the current exploratory study two participants were asked to make a wrist or arm motion while listening to a recording from a vocalizer, an original participant from Pouw, Harrison, et al. (2018a). The vocalizer was continuously voicing the vowel /a:/ while making a high-impetus arm motion or a low-impetus wrist motion at slow, medium, or fast movement tempos. Arm motions have a higher physical impetus on the body as larger body parts are involved in the movement, as compared to wrist movements. The current participants' task was to synchronize their own movement with the movement of the vocalizer, as they perceive it via the acoustics. The current exploratory study served as a basis for a pre-registration of a planned confirmatory study (see https://osf.io/9843h/). For a comprehensive follow-up study and description of methods see (Pouw, Paxton, Harrison, & Dixon, under review).

2. Methods and results

2.1. Participants and design

Two female graduate students (ages: 22 and 28) participated in the current exploratory study. The current study entailed a full 2 x 2 x 3 within-subject design: a two-level factor 'listener movement' condition (listening while moving wrist vs. listening while moving arm), a two-level factor 'vocalizer movement' condition (wrist movement vocalization vs. arm movement vocalization), and a three-level factor 'tempo' (slowed down vs. self-paced vs. sped up). Note that slowed-down and sped-up versions were derived from the original self-paced movement vocalizer (see procedure below). Participants performed 12 trials, one for each cell of the design (2 x 3 x 2). Each trial consisted of 5 phonation cycles where the vocalizer took a full breadth and phonated until breadth was almost emptied and phonation could not be steadily maintained (Pouw, Harrison, et al., 2018a, 2018b).

2.2. Stimuli

We extracted two trials collected with a participant in a previous study wherein (henceforth referred to as the 'vocalizer'. For both trials, the vocalizer continuously voiced the vowel /a:/ (as in 'cinema') while making repetitive up-and-down upper-limb movements at a self-paced tempo (around 1.8Hz). The movements were made on the sagittal plane with fingers fully extended, with a higher velocity

in the down-phase so as to have a beat-like movement profile. The vocalizer was instructed to try not to let voicing be affected by the movement.

In the extracted wrist-movement vocalization trial, repetitive wrist movements of the dominant right hand were made with no movement around the elbow joint. For this wrist movement, the elbow joint was kept at a 90-degree angle. The wrist movement vocalization trial (listen to the sound-clip here: https://osf.io/rvx3c/) reflected a low-impetus movement relative to second arm-movement vocalization trial (sound-clip available here: https://osf.io/ymqnu/). The arm-movement vocalization trial was produced by the participant moving her lower arm around the elbow joint, while keeping the wrist joint locked at 0 degrees.

To construct the stimuli of different tempos, we first looped the original audio track from the vocalizer (i.e., movement with self-directed speed) 5 times, such that there were 5 voicing episodes with intermittent pauses where the vocalizer took a full breath. This self-paced vocalization track serves as our "normal" tempo stimuli. We then created two additional versions of this vocalization track, one that was artificially slowed down by 20% and one that was artificially sped up by 20%. These transformations were done with AVS Audio Editor (Online Media Technologies Ltd.), which allows for tempo transformation while maintaining the original pitch. We made a set of three vocalization tempos for both the wrist-movement vocalization conditions and arm-movement vocalization conditions. Effectively this resulted in 3 tempo conditions (slow down vs. self-paced vs. sped up). The tempo conditions provide additional information whether participants are sensitive to movement-induced rhythm in voicing, which would be evident in in lower or higher frequency of listener movement for slowed down or sped up condition (respectively) as compared to self-paced tempo condition.

2.3. Motion-tracking equipment

A Polhemus Liberty sampling at 240Hz was used to record movement of the listener (L). Given that upper-limb motions were primarily defined by vertical motion (in the z-dimension), we will only perform synchrony analyses for this dimension. We smoothed z-position traces with a first-order low-pass Butterworth filter of 33 Hz.

Audio Presentation. Participants wore a Samsung Level On EO-PN900BBEGUS headphone (with noise-cancelling deactivated) with a wired connection to the PC. Volume was set at a comfortable level for the participant. The audio was pre-buffered and then played using a custom C++ script that started the audio at the exact moment that the motion tracker started recording. This ensured that the original vocalizer motion-tracking data is completely synchronized with listener motion-tracking data.

Procedure. Each participant (i.e., listener) was asked to stand upright with their elbow in a 90-degree angle. The experimenters then demonstrated the two movement types that the listener needed to make: one wrist movement and one arm movement. Subsequently the participant was informed that they would repeatedly listen to someone voicing, while the vocalizer had been concurrently moving her upper limbs at different speeds (although no additional information was given about speed manipulations). The listener was then told that they would need to synchronize with the movements of the vocalizer, based solely on hearing her voice. Participants briefly practiced the synchronization task with an arm movement vocalization trial of self-paced tempo, while the listener was asked to move their wrist or arm in synchrony. After practicing, participants performed 12 trials (in randomized order) containing all 2 (listener movement) x 2 (vocalizer movement) x 3 (tempo) conditions wherein they heard 5 voicing episodes before going to the next trial.

2.4. Analyses

Spectral Analyses (FFT). We performed spectral analyses (fast Fourier transform or FFT) using R's native stats package (function spectrum) to assess changes in listeners' movement frequency as a function of vocalizer tempo condition.

Relative Phase Analyses (Φ). We performed relative-phase analyses using a simple continuous point-wise estimation method (e.g., Zelic, Kim, & Davis, 2015; see also Kelso, Del Colle, & Schoner, 1990). To calculate Φ we used the equation

$$\varphi = 2\pi\Delta t/T_v$$

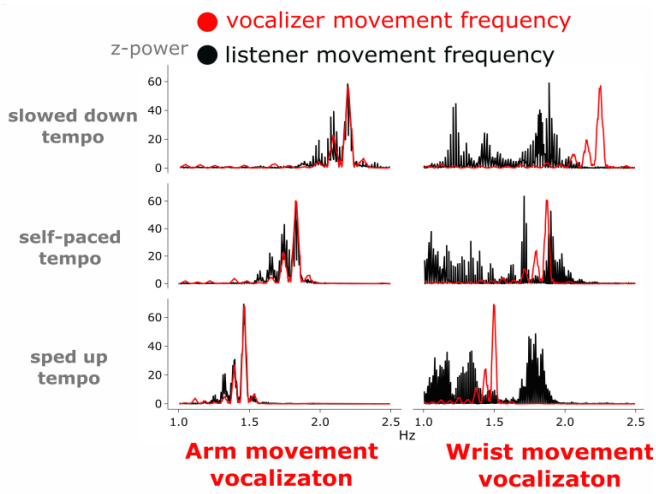
where T_v is the current time interval for the original vocalizer's maximum vertical extension (i.e., the time between each beat of the vocalizer movement). Δt reflects the asynchrony of the moment

of maximum vertical extension of the listener versus the vocalizer. 2π transforms temporal dispersion into the angular dispersion. We converted the equation's output from radians to degrees such that 0 degrees indicated in-phase coordination, -180 degree indicated anti-phase coordination with listener in the lead, and 180 degrees indicated anti-phase coordination with vocalizer in the lead.

2.5. Overview results

There are two hierarchically organized research questions that need to be answered before concluding that there is (some) acoustic specification of upper-limb movements in phonation. Firstly, can participants attune to the rhythmic tempo of the movement? Secondly, if indeed participants are sensitive to rhythm in phonation, are participants able to attune to the exact phasing of the vocalizers' movement in a 1:1 in-phase fashion? Note, a supplementary figure is available at https://osf.io/zngb2/ containing an example time series of the listener (participant 1) against the vocalizer for different movement tempos.

Acoustic Specification of Movement Tempo in Phonation (FFT Analyses). Next, we formally assessed for all the data the degree to which participants were attuning to tempo information in the vocalization. Figure 2 shows the mean spectral results for the arm- vs. wrist-movement vocalization conditions for all three tempos (and regardless of which movement listeners were making). Namely, there were no clear effects for when listeners were trying to synchronize while making wrist- versus arm-movements (see additional plot with listener movement conditions here: https://osf.io/6adm4/). As shown in Figure 2, we found clear evidence for listener-vocalizer tempo-specific movement coupling when the listener heard the clips in which the vocalizer was making arm movements (i.e., arm-movement vocalizations), but not those trials in which the vocalizer was making wrist movements (wrist-movement vocalizations). Thus, for wrist-movement vocalizations, participants seemed to fail to pick up movement tempo information; while listening to arm-movement vocalizations, participants could both adjust tempo of their own arm movements and wrist movements.



Note. FFT results for all movement frequencies (horizontal axis in hertz [Hz]; vertical axis *z*-standardized power for that frequency) for both the vocalizer and the listener movement frequencies. The vocalizer wrist and arm movement frequencies show slower or faster frequencies per tempo condition by design (as we artificially manipulated the tempo for these conditions). The natural frequency of the vocalizer (original tempo condition) was about 1.75 Hz, with a slight faster frequency for when the vocalizer was making a wrist movement. Interestingly, in the arm vocalization conditions, there was clear frequency coupling between listeners' movement with that of the vocalizer. This is indicated by the large overlap of spectral peaks in arm vocalization condition.

Figure 2. Spectral results movement frequencies for vocalizer and listener.

For statistical confirmation of the results obtained in Figure 1, we assessed whether listener's dominant movement frequencies were affected by tempo condition (as well as vocalizer movement condition and listener movement condition). That is, we quantified whether tempo conditions predicted dominant frequency—with higher frequencies for fast tempo conditions and lower frequencies for slow tempo conditions, both as compared to the original tempo. To test this, we extracted the frequency with the highest observed power (i.e., dominant frequency) for each trial. Subsequently we performed `nlme` mixed regressions using participant as a random intercept (adding adding random slopes caused the model not to converge), identifying the best model by comparing model fits at increasing levels of complexity.

Compared to a model predicting the overall mean for dominant frequency, entering tempo condition as a predictor for dominant frequency improved the fit of the model (change in χ^2 [1] = 10.32, p = .006). Adding to the previous model, vocalizer movement condition improved the fit of the model further, change in χ^2 [1] = 6.95, p = .008. Adding the interaction between tempo (3 levels)

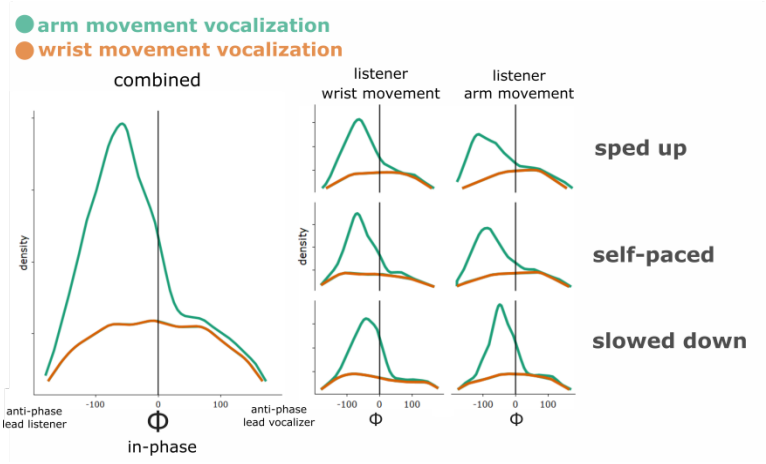
and vocalizer-movement conditions to the previous model further improved the model, change in χ^2 [2] = 7.67, p = .021. Finally, adding listener movement condition to this previous model did not significantly improve predictions of dominant frequency further (change in χ^2 [1] = 2.41, p = .12).

The best-fit model with vocalizer movement and tempo (and their interaction) was assessed with post-hoc comparisons with the `R` package `lsmeans` (using Tukey correction for multiple comparisons). Only the arm movement vocalization condition showed tempo scaling of listener movement with that of the vocalizor, and this was only statistically reliable for the contrast between sped-up vs. slowed down tempo condition. All model results are reported in Table 1.

Table 1
Post-hoc comparisons for frequency scaling per tempo and vocalizer movement condition

Arm Movement Vocalization	Difference estimate	t (df=17)	p-value (corrected)
Sped up - slowed down tempo	0.71Hz	4.89	**<.001**
Sped up - self-paced tempo	0.36Hz	2.52	.054
Self-paced - slowed down tempo	0.34 Hz	-2.37	.073
Wrist Movement Vocalization			
Sped up - slowed down tempo	0.181	1.25	0.443
Sped up - self-paced tempo	0.192	1.33	0.402
Self-paced - slowed down tempo	-0.011	-0.077	0.996

Acoustic Specification of Movement Phasing in Phonation (Relative Phase Analyses: *SD Φ*). Now that we have established that there is frequency-coupling between listener and vocalizer movement (but only for vocalizer arm movement), we assess whether there is also phase-synchronization. Note that it is possible that participants picked up the rhythmic structure in the voicing while being oblivious about the exact phases of the vocalizer's movement. Figure 3, however, clearly shows that there was indeed listener-vocalizer phase-coupling but only for the vocalizer arm movement. Furthermore, there is not perfect in-phase locking but rather a negative mean asynchrony whereby the listener anticipates the vocalizer with about Φ = 50 degrees. Note that negative mean asynchrony is a very common phenomenon in sensorimotor synchronization tasks (Repp, 2005).



Note. The left panel shows the relative phase distributions for all data combined for the vocalizer wrist movement and the vocalizer arm movement condition. On the right-hand side, data are parsed for each tempo and listener movement conditions. The clear peaked relative phase distributions for the vocalizer arm motions indicating phase-coupling for this condition, but we also saw that listeners tended to anticipate vocalizer movement.

Figure 3. Distributions relative phase listener-vocalizer.

We statistically confirmed the phase-coupling results by computing the standard deviation of Φ per trial performed. If phase-coupling is more pronounced, lower SD Φ will be observed (i.e., less variable/more stable phase relations around the average relative phase). We assessed this using

`nlme` mixed regressions (again using participant as the sole random intercept without random slopes, as adding random slopes caused the model to fail to converge).

As compared to a model predicting the overall mean, entering vocalizer movement condition as a predictor for SD Φ led to increased fit of the model (change in $\chi 2$ [1] = 30.62, $p < .001$). Adding tempo condition as an additional predictor did not further improve prediction of relative-phase (change in $\chi 2$ [1] = 1.34, $p = 0.51$). Adding listener movement condition as a predictor for relative phase (next to vocalizer movement) also did not improve the previous model (change in $\chi 2$ [1] = 0.74, $p = 0.378$). Therefore, the resulting best-fit model—which included vocalizer movement condition as the sole fixed effect—revealed that vocalizer arm movement condition had a lower SD Φ of 51 degrees as compared to the vocalizer wrist movement condition, $b = -51.09$, $t[21] = -7.81$, $p < .001$. These findings support our hypothesis that listeners synchronized their movement phasing with phase information in the vocalizer acoustics.

3. Discussion

While preliminary (results require replication), the current exploratory study demonstrates that it is—in principle—possible to glean information about bodily movement from voice acoustics alone. We found that listeners demonstrated both frequency-coupling and phase-coupling of their own movements with that of a vocalizer who was moving at different tempos while producing a single vowel sound. As predicted based on the absence of acoustic effects (Pouw, Harrison, et al., 2018a, 2018b), the vocalizer wrist movements (as opposed to vocalizer arm movements) were not reliably detected by the listeners; no evidence was obtained for frequency- or phase-locking in the wrist movement vocalizations. Although it appears that there must be some information about bodily gestures in phonation acoustics, the current exploratory study is unable to determine how pervasive the couplings might be.

The current results suggest that we do not necessarily hear voicing as only voicing: Intriguingly, we can also detect within voicing the bodily states of the voicer on the basis of acoustic-body invariants. The current research therefore directly aligns with the ecological psychology of language (Fowler, 2010) and the acoustic perception of object geometry (Carello et al., 2005). Our findings may extend this research program with the idea that prosodic contrasts in speech are direct informational sources of bodily tensioned states (including hand gestures). The findings further align with research on other animals, who often modulate their vocal activity so as to appear larger (and more intimidating or appealing) in size (Hardus, Lameira, Van Schaik, & Wich, 2009).

References

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.

Carello, C., Anderson, K. L., & Kunkler-Peck, A. J. (1998). Perception of object length by sound. *Psychological Science*, *9*(3), 211–214.

Carello, C., Wagman, J. B., & Turvey, M. T. (2005). Acoustic specification of object property. *In J. D. Anderson & B. Fisher Anderson (Eds.), Moving*, 79–104.

Cravotta, A., Busà, M. G., & Prieto, P. (2018). Restraining and encouraging the use of hand gestures: Effects on speech. In *9th International Conference on Speech Prosody 2018* (pp. 206–210). ISCA. doi: 10.21437/SpeechProsody.2018-42

Ejiri, K., & Masataka, N. (2001). Co-occurrences of preverbal vocal behavior and motor action in early infancy. *Developmental Science*, *4*(1), 40–48.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*(1), 3–28.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 816–828.

Gibson, J. J. (2014). *The Ecological Approach to Visual Perception : Classic Edition*. Psychology Press.

Hardus, M. E., Lameira, A. R., Van Schaik, C. P., & Wich, S. A. (2009). Tool use in wild orangutans modifies sound production: A functionally deceptive innovation? *Proceedings. Biological Sciences*, *276*(1673), 3689–3694.

Hoetjes, M., Krahmer, E., & Swerts, M. (2014). Does our speech change when we cannot gesture? *Speech Communication*, *57*, 257–267.

Iverson, J. M., & Goldin-Meadow, S. (2001). The resilience of gesture in talk: gesture in blind speakers and listeners. *Developmental Science*, *4*(4), 416–422.

Pouw, W., Paxton, A., Harrison, S., & Dixon, J. A. (under review). Social Resonance: Acoustic information about upper limb movement in voicing. https://psyarxiv.com/ny39e

Pouw, W., Harrison, S. A., & Dixon, J. A. (2018a). The physical basis of gesture-speech synchrony: Exploratory study and pre-registration. https://doi.org/10.31234/osf.io/9fzsv

Pouw, W., Harrison, S. J., & Dixon, J. A. (2018b). Gesture-speech physics: The biomechanical basis of festure-speech synchrony. https://doi.org/10.31234/osf.io/tgua4

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychonomic bulletin & review*, *12*(6), 969-992.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232.