Sensory Modality of Input Influences the Encoding of Motion Events in Speech But Not Co-Speech Gestures

Ezgi Mamus (ezgi.mamus@mpi.nl)
Centre for Language Studies, Radboud University, Nijmegen, The Netherlands
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Laura J. Speed (l.speed@let.ru.nl)
Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

Aslı Özyürek (asli.ozyurek@mpi.nl)
Centre for Language Studies & Donders Center for Cognition, Radboud University, Nijmegen, The Netherlands
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Asifa Majid (asifa.majid@york.ac.uk)
Department of Psychology, University of York, York, UK

Abstract

Visual and auditory channels have different affordances and this is mirrored in what information is available for linguistic encoding. The visual channel has high spatial acuity, whereas the auditory channel has better temporal acuity. These differences may lead to different conceptualizations of events and affect multimodal language production. Previous studies of motion events typically present visual input to elicit speech and gesture. The present study compared events presented as audio-only, visual-only, or multimodal (visual+audio) input and assessed speech and co-speech gesture for path and manner of motion in Turkish. Speakers with audio-only input mentioned path more and manner less in verbal descriptions, compared to speakers who had visual input. There was no difference in the type or frequency of gestures across conditions, and gestures were dominated by path-only gestures. This suggests that input modality influences speakers' encoding of path and manner of motion events in speech, but not in co-speech gestures.

Keywords: motion events; visual perception; auditory perception; multimodal; spatial language; iconic gestures

Introduction

Vision is widely considered as the primary source of information about space and is the basis of rich mental representations. Vision dominates spatial perception as it has the advantage of providing high spatial acuity for both close and distant space (e.g., Eimer, 2004; Stokes & Biggs, 2015). When presented simultaneously with conflicting non-visual information, the dominance of vision results in cross-modal illusions such as the ventriloquism effect (Alais & Burr, 2004; Howard & Templeton, 1966), although audition is found to dominate in temporal processing, since audition has higher temporal acuity than vision (e.g., Recanzone, 2003; Repp & Penel, 2002).

The dominance of vision is thought to be reflected in language too, especially in Western societies (Levinson & Majid, 2014; Majid et al., 2018; San Roque et al., 2015; Viberg, 1983). Compared to other senses, vision-related words are more frequent and numerous in the languages of the world (San Roque et al., 2015; Winter, Perlman, & Majid, 2018). Nevertheless, in one study of 20 diverse languages, Majid et al. (2018) found that not all languages show highest linguistic codability—i.e., agreement on descriptions of experience—for vision.

Given the qualitative perceptual differences between modalities and the diverse codability of the senses, we ask whether sensory modality of input influences linguistic encoding of spatial information. In the present study, we compared motion events presented as audio-only, visual-only, or multimodal (visual+audio) stimuli and examined both verbal and gestural expressions of path and manner of motion. We examined speech and gesture as each can provide distinct information about the underlying conceptualization of events for language production (e.g., Kita, Alibali, & Chu, 2017; Kita & Özyürek, 2003). Since gestures are considered to arise mainly from visuospatial representations (e.g., Hostetter & Alibali, 2008, 2019), we can determine whether audio and visual input give rise to similar visuospatial representations. So, we ask for the first time whether speakers produce similar types and frequency of gestures to depict spatial information about motion events extracted from visual versus auditory input.

In contrast to the holistic nature of visual information, auditory information is represented sequentially. Spatial cognition and language studies have shown that when people rely exclusively on non-visual information to build spatial representations (such as in blindness), their representations reflect the sequential nature of input (e.g., Iverson, 1999; Iverson & Goldin-Meadow, 1997; Thinus-Blanc & Gaunet, 1997). For example, during a route description task for a familiar spot in their school, blind children describe the path in a more segmented fashion with more landmarks in their speech than sighted children (Iverson & Goldin-Meadow, 1997). Furthermore, when children gave segmented verbal descriptions, regardless of their visual status, they produced

fewer gestures. Iverson and Goldin-Meadow (1997) claimed that gesture frequency decreases with segmented speech due to the process of gesture generation. Speech and gesture arise from an integrated system, and gestures capture a representation as "a global whole" (McNeill, 1992). Therefore, when speech is represented sequentially, it is not well-suited for gesture. These studies suggest there may be differences between visual and non-visual modalities, however it is unclear whether the attested differences in the linguistic encoding of spatial information arise from the long-term effect of blindness or instead are due to the sequential nature of input at encoding.

More generally, the encoding of motion has been studied extensively across languages (e.g., Gennari, Sloman, Malt, & Fitch, 2002; Gullberg, Hendricks, & Hickmann, 2008; Kita & Özyürek, 2003; Papafragou, Massey, & Gleitman, 2002; Slobin, Ibarretxe-Antuñano, Kopecka, & Majid, 2014). Slobin (1996) proposed that speakers learn to encode aspects of events depending on distinctions in their language. One crucial distinction in motion representation is between path and manner (Talmy, 1985). Since path and manner of motion are distinct spatial notions, modality of input could influence their encoding differently in speech and gesture depending on the type of language. No study has systematically investigated this issue as of yet.

Most previous studies present visual stimuli to elicit speech and gesture about motion events, with the exception of one study which used haptic input (Özçalışkan, Lucero, & Goldin-Meadow, 2016) and another which used auditory input (Mamus, Rissman, Majid, & Özyürek, 2019). However, neither of these studies directly tested the role of input modality on linguistic representations of motion events. Özçalışkan et al. (2016) tested blind participants, sighted participants, and sighted but blindfolded participants. Toys such as a house and a crib were used as landmarks and multiple static dolls in different postures were used to create the impression of motion (e.g., a girl running into a house). Participants were instructed to describe the scenes and were explicitly encouraged to gesture at the same time. Blind speakers talked and gestured in a comparable manner to blindfolded and sighted speakers of their language. No direct comparison was made between blindfolded and sighted speakers' verbal and gestural patterns however. Therefore, it remains unclear whether input modality specifically affects event representations from these results.

In a later study, Mamus et al. (2019) created auditory motion events by presenting audio-recordings in a 5+1 surround sound system and tested the effect of blindfolding on verbal expressions of path of motion. They found that blindfolded speakers' path descriptions were more sequential (i.e., segmented with landmarks) than sighted speakers, but all speakers could extract information about the path of motion based on the sounds of events. However, since there was no comparison with visual input, it is unclear whether descriptions were impoverished, richer, or the same as those that would be elicited from sight. Moreover, Mamus et al. (2019) did not explore expressions of manner of motion or gestures during event description. Therefore, a systematic comparison of how sensory input influences expressions of spatial information in language and gesture is required.

Another goal of our study was to experimentally test how people express naturalistic auditory motion events using spontaneous iconic gestures. Iconic gestures are considered an effective tool to convey visuospatial and motor information because they are said to represent such information directly from a mental image (Alibali, 2005; Hostetter & Alibali, 2008, 2019). Several gesture production theories claim that gestures depend on visuospatial imagery and therefore occur more frequently during the expression of spatial and motor information (e.g., de Ruiter, 1998; Hostetter & Alibali, 2008; Kita & Özyürek, 2003; Wesp, Hesse, Keutmann, & Wheaton, 2001). However, it is also claimed that type of language (e.g., motion event typology) influences which aspects of spatial features of events are expressed in speech and gesture (e.g., Kita & Özyürek, 2003).

To date, gesture production has predominantly been studied using visual stimuli (e.g., video-clips, cartoons, line drawings, paintings, and so on; but see, e.g., Iverson, 1999; Özçalışkan et al., 2016). It is possible that focusing on the visual modality might create a modality-specific bias in favor of visuospatial imagery. To our knowledge, no study has addressed whether speakers produce the same type and frequency of gestures when expressing spatial information drawn from non-visual vs. visual modalities.

The present study investigated how Turkish speakers represent spatial information in language based on auditory or visual input. We compared verbal and gestural expressions of path and manner of motion events that were presented as audio-only, visual-only, or multimodal (visual + audio) stimuli. Our main goal was to compare audio-only versus visual-only input, however including a multimodal condition allows a further test of the dominance of vision. If vision alone provides enough information about events, then we would not expect a difference between the visual-only and the multimodal conditions in linguistic expressions of spatial information.

We can make distinct predictions about the encoding of path and manner in speech and gesture as a function of input modality. If the previously attested differences of path information from non-visual input (Iverson, 1999; Iverson & Goldin-Meadow, 1997) are caused by in-the-moment differences in perception, we would predict that participants in the visual conditions would describe motion events in a more global fashion, leading to fewer mentions of path in speech than participants in the audio-only conditions.

It is less clear how input modality would affect manner encoding. To differentiate particular manners such as walk vs. run, vision provides rich information about biomechanical properties, as well as information about speed and direction of motion (e.g., Malt et al., 2014). However, audition is also good at providing temporal information—such as rhythm of a motion (e.g., Recanzone, 2003; Repp & Penel, 2002). It is presently unclear whether visual vs. auditory information

would necessarily lead to different manner encoding of motion events in speech and gesture.

If gestures are generated mainly from visuospatial imagery (e.g., Hostetter & Alibali, 2008, 2019), then gesture frequency for both path and manner should decrease in the audio-only condition compared to visual-only and multimodal conditions. However, if speakers can build comparable spatial representations from auditory input as they do from visual input, then they should produce similar types and frequency of gestures to depict path and manner of motion in all conditions.

However, these predictions about the encoding of path and manner in speech and gesture are based only on stimulus affordances, but can be further influenced by language-specific patterns. Turkish is considered a verb-framed language, which primarily encodes path in the main verb and optionally encodes manner in a subordinated verb or adverbial phrase (Talmy, 1985). As encoding of path is essential in Turkish but manner is an optional element, manner expressions might be more susceptible to effects of input modality than path. This prediction also holds for gesture types. Previous studies showed that Turkish speakers are more likely to produce path gestures than manner gestures to depict motion events, even if both manner and path are expressed in speech (e.g., Kita & Özyürek, 2003; Özyürek, Kita, Allen, Furman, & Brown, 2005). Therefore, gestures of Turkish speakers might also reflect this path framing, overriding any effect of input modality.

Method

Participants

Forty-five native Turkish speakers with normal or corrected-to-normal vision were recruited from Boğaziçi University. Fifteen participants were randomly assigned to each of three conditions: audio-only (*M*=21 years, *SD*=2), visual-only (*M*=22 years, *SD*=3), and multimodal (*M*=21 years, *SD*=2). Participants were tested in a quiet room on Boğaziçi University campus. They all received extra credit in a psychology course for their participation and provided written informed consent in accordance with the guidelines approved by the IRB committees of Boğaziçi and Radboud Universities.

Stimuli

We video- and audio-recorded locomotion and non-locomotion events performed by an actress. Locomotion events were the critical items, whereas non-locomotion events were included as filler items. We created 12 locomotion events by crossing 3 manners (walk, run, and limp) with 4 paths (to, from, into, and out of) in relation to a landmark object (door or elevator). A camera and sound recorder were placed next to the landmark objects. For *to* and *into* events, the actress moved towards landmarks—so the path direction was approaching the sound recorder—and for *from* and *out of* events, the actress was moving away from landmarks—so the path direction was away from the sound recorder.

To create non-locomotion events, the same actress performed three-participant "transitive" actions with different objects (e.g., opening a can, chopping a cucumber), and the video and sound were recorded across from her at a fixed distance. There were 24 trials in total, including 12 locomotion and 12 non-locomotion events.

Procedure

Participants were presented with the events on a laptop using Presentation Software. The events were presented as audio-clips to participants in the audio-only condition, as silent video-clips to the participants in the visual-only condition, and as video+audio clips to the participants in the multimodal condition. All participants wore a headphone during the task.

Participants were asked to describe each event. They were told that another participant would watch their descriptions and watch/listen to the same events and be asked to match descriptions with events. There were no instructions about gesture use. Before the experiment started, they had two practice trials consisting of two non-locomotion events. Participants initiated the next trial at their own pace by pressing a button after they described the event. Descriptions were recorded with a video camera placed approximately 1.5 m away and across from participants. After the description task, participants filled out a demographic questionnaire on a laptop. The duration of the experiment was around 15 minutes.

Coding

Speech for the locomotion events were transcribed and coded by native Turkish speakers using ELAN (Wittenburg et al., 2006). Event descriptions were split into clauses. A clause was defined as a verb and its associated arguments or verb clauses with gerund phrases. Clauses were then coded as relevant if they included locomotion descriptions. For example, a clause including a transitive event such as opening a door or ringing the bell was not coded as relevant to the target event. Finally, each relevant clause was coded according to the type of information it contained: (a) path (trajectory of motion) and (b) manner of motion (how the action is performed)—see (1).

(1)	***Kapı-dan***	***çık-tı***	**/**	***yürü-yerek***
	door-ABL	exit-PST		walk-Connective
		(path verb)		(manner)

'(someone) exited from the door / while walking.'
/ indicates a new clause for the purposes of coding

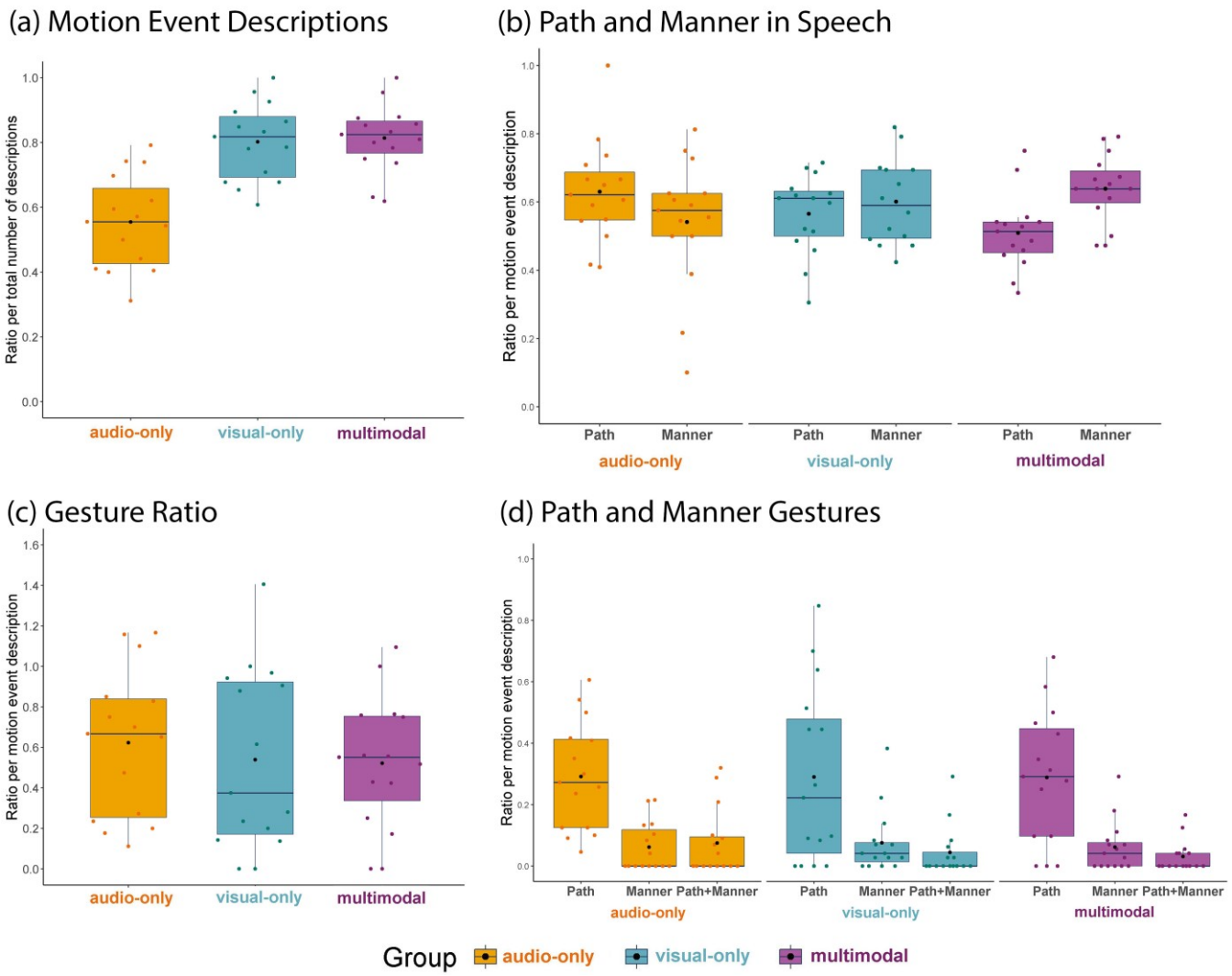

Figure 1. (a) Motion event descriptions. (b) Path and manner in speech. (c) Gesture for motion event descriptions. (d) Path and manner gestures for motion event descriptions. Colorful dots represent the average data for each participant; black dots represent the mean.

Participants' spontaneous iconic gestures were also coded for each target motion event description. Iconic gestures represented trajectory and/or manner of movement and were further classified into (i) path-only, (ii) manner-only, and (iii) path+manner conflated together. Path-only gestures depict a trajectory of movement without representing the manner, and manner only gestures show the style of a movement without representing the trajectory. Path+manner gestures depict both trajectory and manner of movement simultaneously.

Results

To analyze the data, we used one-way ANOVA and linear mixed-effects regression models (Baayen et al., 2008) with random intercepts for Participants and Items, using the packages lme4 (Version 1.1–23; Bates et al., 2015) and lmerTest (Version 3.1–3; Kuznetsova et al., 2017) to retrieve *p*-values in R (Version 3.5.1; R Core Team, 2018). We conducted separate linear mixed effects models on path and manner mentions in speech and gesture. To assess statistical significance of the fixed factors, we used likelihood-ratio tests, comparing models with and without the factors of interest.

Speech

We investigated whether participants differed in how they described path and manner of motion events based on audio-only, visual-only, or multimodal input. We first calculated the ratio of motion event descriptions per participant. For each participant, we divided the total number of motion event descriptions by the total number of descriptions. A one-way ANOVA showed there was a significant difference between input modalities, $F(2,42) = 21.14$, $p < .001$. A post-hoc Tukey test showed that participants in the audio-only (M=0.56, SD=.15) condition had fewer motion event descriptions

compared to both participants in the visual-only (M=0.80, SD=.12) and multimodal (M=0.81, SD=.10) conditions (ps < .001). (See Figure 1a).

To investigate whether participants differed in how they expressed path and manner in speech, we ran an lmer model with the fixed factors of input modality (audio-only, visual-only, multimodal) and type of expression (path vs. manner) using the ratio of mention of path and manner per motion event description as input (see Figure 1b). The model revealed no fixed effect of input modality, χ^2 (2) = 0.78, p = .68, and no fixed effect of type of expression on path and manner mention, χ^2 (1) = 3.66, p = .06. However, the model did reveal an interaction between input modality and type of expression, χ^2 (2) = 16.52, p < .001. Compared to participants in the audio-only condition, participants in the visual-only (β = .147, SE = .053, t = 2.79, p = .005) and multimodal conditions (β = .210, SE = .053, t = 3.99, p < .001) mentioned manner more than path. In other words, participants encoded more manner than path information in the visual conditions (visual-only and multimodal) than auditory condition. There was no difference between participants in the visual-only and multimodal conditions in terms of reference to path vs. manner (β = .063, SE = .052, t = 1.21, p = .23). Moreover, participants in the audio-only condition mentioned path more often than participants in the multimodal condition (β = -.012, SE = .044, t = -2.66, p = .011) but not more than participants in the visual-only condition (β = -.063, SE = .044, t = 1.43, p =.16).

Gesture

We investigated whether participants differed in how they gestured about path and manner of motion events based on input. First, we compared groups in terms of the gesture ratio per motion event descriptions. A one-way ANOVA showed that there was no significant difference in gesture ratio between participants in the audio-only (M=0.62, SD=0.37), visual-only (M=0.54, SD=0.45), and multimodal (M=0.52, SD=0.33) conditions; F(2,42) = 0.3, p = .74 (see Figure 1c).

To further investigate what type of gestures participants produced, we calculated the ratio of iconic (path only, manner only, and path+manner) gestures per motion event description. For these calculations, total counts of path only, manner only, and path+manner gestures were divided by the number of motion event descriptions for each trial. The data were analyzed in the same way as the speech data. We ran an lmer model with fixed factors of input modality (audio-only, visual-only, and multimodal) and type of expression (path-only, manner-only, and path+manner) using the ratio of path and manner gestures per motion event description as input (see Figure 1d). The model revealed a fixed effect of type of expression, χ^2 (2) = 278.54, p < .001. Regardless of the condition, speakers produced more path-only gestures than manner-only (β = -.223, SE = .015, t = -14.61, p <.001) and path+manner gestures (β = -.238, SE = .015, t = -15.61, p < .001). There was no difference between manner-only and path+manner gestures (β = -.015, SE = .015, t = -1.01, p = .31). The model revealed no fixed effect of input modality, χ^2 (1) = 0.16, p = .92, and no interaction between input modality and type of expression on path and manner gestures, χ^2 (4) = 2.13, p = .71.

Discussion

The present study investigated how Turkish speakers represent spatial information in language based on differential sensory input. We compared motion events presented as audio-only, visual-only, or multimodal (visual+audio) stimuli and examined the expressions of path and manner of motion in speech and gesture. We found that speakers produced more motion event descriptions when they watched events with visual input—either multimodal or visual-only—in comparison to when they only listened to events. This shows that speakers provide richer linguistic information about spatial components of motion events when visual input is present. This finding fits the claims that vision dominates in language, at least in the domain of space (e.g., Levinson & Majid, 2014; Majid et al., 2018; San Roque et al., 2015; Viberg, 1983; Winter et al., 2018).

Our data also showed that speakers were able to extract information about both path and manner of motion from auditory input alone. This extends the previous findings of Mamus et al. (2019), suggesting that audition is informative about at least some aspects of manner of motion.

Nevertheless, there was a qualitative difference in linguistic expressions of spatial information drawn from visual vs. auditory input. We found that Turkish speakers were more likely to mention manner than path information in their speech in the visual conditions than auditory condition. This is interesting because, based on the Turkish typology, encoding path is more essential than manner in motion event descriptions (Talmy, 1985). So, this finding may be the result of differences in stimulus affordances: as vision provides more detailed information about manner of motion than audition does, manner of motion might be more salient in visual input, even in a path language. This suggests that the modality of input influences speakers' encoding of spatial event components independently of the well-established tendencies of speaking a particular language (e.g., Slobin, 1996; Talmy, 1985).

This is also the first study that directly tested whether modality of sensory input influences gesture production for the same motion event. Existing theories about the nature of gestures emphasize that gestural representations are mainly visuospatial (e.g., Alibali, 2005; de Ruiter, 1998; Hostetter & Alibali, 2008; Wesp et al., 2001). We found that auditory spatial input can elicit similar types and frequency of gestures as visual input for the expression of path and manner of motion events. Our results provide new insight into the nature of gestures, showing that speakers can build gestural representations from input that is auditory.

Interestingly, we found the difference between path and manner representations across input modalities found in speech was not reflected in gesture. Due to the interaction between speech and gesture, it might be expected that when manner of motion is mentioned more often in speech, manner

gestures should also increase. However, even though speakers in the visual-only and multimodal conditions mentioned manner more often in speech, there was no increase in the frequency of manner gestures. Regardless of the type of input they received, speakers produced more path only gestures than manner only or path+manner gestures. This finding aligns well with the previous literature. It is well-documented that Turkish speakers produce more path gestures than manner or path+manner, in accordance with the language-specific syntactic patterns of speech (e.g., Kita & Özyürek, 2003; Özçalışkan et al., 2016; Özyürek et al., 2005). Gestures not only depict imagistic elements in event representations, but are also shaped by language during speaking (e.g., Kita & Özyürek, 2003; Özyürek et al., 2005; Slobin, 1996).

Although our results imply that modality of input does not affect the gesture of Turkish speakers, results may differ for a satellite-framed language that encodes manner in the main verb—such as English—or an equipollently-framed language—such as Mandarin Chinese (e.g., Brown & Chen, 2013). To better understand whether and how co-speech gesture is influenced by non-visual spatial input, a cross-linguistic investigation is necessary.

Conclusion

The present study examined the role of sensory input modality on the linguistic expression of spatial event components in both speech and co-speech gesture and found they pattern in distinct ways. In comparison to the auditory modality, the visual modality appears to foreground manner more than path in speech, but gestures are generated similarly regardless of input modality. These findings suggest that the modality of input influences speakers' encoding of path and manner of motion events in speech.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology, 14*(3), 257-262.

Alibali, M. W. (2005). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition & Computation, 5*, 307-331.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Brown, A., & Chen, J. (2013). Construal of Manner in speech and gesture in Mandarin, English, and Japanese. *Cognitive Linguistics, 24*(4), 605–631.

de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284-311). Cambridge: Cambridge University Press.

Eimer, M. (2004). Multisensory integration: how visual experience shapes spatial perception. *Current biology, 14*(3), R115-R117.

Gennari, S. P., Sloman, S. A., Malt, B. C., & Fitch, W. T. (2002). Motion events in language and cognition. *Cognition, 83*(1), 49-79.

Gullberg, M., Hendriks, H., & Hickmann, M. (2008). Learning to talk and gesture about motion in French. *First Language, 28*(2), 200-236.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic bulletin & review, 15*(3), 495-514.

Hostetter, A. B., & Alibali, M. W. (2019). Gesture as simulated action: Revisiting the framework. *Psychonomic bulletin & review, 26*(3), 721-752.

Howard, I. P., and Templeton, W. B. (1966). Human Spatial Orientation. Oxford, England: Wiley.

Iverson, J. M. (1999). How to get to the cafeteria: Gesture and speech in blind and sighted children's spatial descriptions. *Developmental psychology, 35*(4), 1132.

Iverson, J. M., & Goldin-Meadow, S. (1997). What's communication got to do with it? Gesture in children blind from birth. *Developmental psychology, 33*(3), 453.

Kita, S., Alibali, M. W., & Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological review, 124*(3), 245.

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26.

Levinson, S. C., & Majid, A. (2014). Differential ineffability and the senses. *Mind & Language, 29*, 407–427.

Mamus, E., Rissman, L., Majid, A., & Özyürek, A. (2019). Effects of blindfolding on verbal and gestural expression of path in auditory motion events. In A. K. Goel, C. M. Seifert, & C. C. Freksa (Eds.), Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019) (pp. 2275-2281). Montreal, QB: Cognitive Science Society.

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O'grady, L., ... & Levinson, S. C. (2018). Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences, 115*(45), 11369-11376.

Malt, B. C., Ameel, E., Imai, M., Gennari, S. P., Saji, N., & Majid, A. (2014). Human locomotion in languages: Constraints on moving and meaning. *Journal of Memory and Language, 74*, 107-123.

McNeill, D. (1992). *Hand and mind: What gesture reveals about thought*. Chicago: University of Chicago Press.

Özçalışkan, Ş., Lucero, C., & Goldin-Meadow, S. (2016). Is seeing gesture necessary to gesture like a native speaker?. *Psychological science, 27*(5), 737-747.

Özyürek, A., Kita, S., Allen, S., Furman, R., & Brown, A. (2005). How does linguistic framing of events influence

co-speech gestures?: Insights from crosslinguistic variations and similarities. *Gesture, 5*(1-2), 219-240.

Papafragou, A., Massey, C., & Gleitman, L. (2002). Shake, rattle, 'n'roll: The representation of motion in language and cognition. *Cognition, 84*(2), 189-219.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundations for Statistical Computing. Available online at: https://www.R-project.org/.

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of neurophysiology, 89*(2), 1078-1093.

Repp, B. H., & Penel, A. (2002). Auditory dominance in temporal processing: new evidence from synchronization with simultaneous visual and auditory sequences. *Journal of Experimental Psychology: Human Perception and Performance, 28*(5), 1085.

San Roque, L., Kendrick, K. H., Norcliffe, E., Brown, P., Defina, R., Dingemanse, M., ... Majid, A. (2015). Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies. *Cognitive Linguistics, 26*, 31–60.

Slobin, D. (1996). From "thought" and "language" to "thinking for speaking." In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge, MA: Cambridge University Press.

Slobin, D. I., Ibarretxe-Antuñano, I., Kopecka, A., & Majid, A. (2014). Manners of human gait: A crosslinguistic event-naming study. *Cognitive Linguistics, 25*(4), 701-741.

Stokes, D., & Biggs, S. (2015). The dominance of the visual. In D. Stokes, M. Matthen, & S. Biggs (Eds.). *Perception and its modalities* (pp. 350–378). Oxford: Oxford University Press.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and semantic description* (pp. 36–149). Cambridge: Cambridge University Press.

Thinus-Blanc, C., & Gaunet, F. (1997). Representation of space in blind persons: vision as a spatial sense?. *Psychological bulletin, 121*(1), 20.

Viberg, Å. (1983). The verbs of perception: A typological study. *Linguistics, 21*, 123–162.

Wesp, R., Hesse, J., Keutmann, D., & Wheaton, K. (2001). Gestures maintain spatial imagery. *American Journal of Psychology, 114*, 591-600.

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition, 179*, 213-220.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), Proceedings of the 5th international conference on language resources and evaluation (pp. 1556–1559). Genoa, Italy: European Language Resources Association.