Approved electronically generated cover-page version 1.0

From Sounds to Actions: how Gestures Depict Auditory Information

A thesis submitted to the University of Manchester for the degree of

MPhil

in the Faculty of Medical and Human Sciences

2014

Clara Cotroneo

School of Psychological Sciences

Word Count: 18,014

Contents

List of Tables

The University of Manchester

Clara Cotroneo MPhil From Sounds to Actions: how Gestures Depict Auditory Information November 2014

Abstract

The *Gesture as Simulated Action* (GSA) Model holds that gestures result from *simulated actions,* which are in turn led by simulations of motor representations and visuospatial representations (Hostetter & Alibali, 2008). The GSA also predicts that, all things being equal, the action component of the mental simulation predicts variations in the gesture rate. Chapter 2 addressed the question of whether audiospatial representations can also lead to gesture production. To test this hypothesis, musical pitch has been chosen because its mental representation is constituted by a spatial and an auditory component, as shown by Connell, Cai, and Holler (2013). Chapter 2 presents two experiments where participants described sequences of sounds that varied in pitch, from higher to lower pitch and vice versa, and an analysis of how this information was conveyed through the gesture modality was conducted. In Study 1, participants verbally described short sequences of musical notes to an addressee. Information about sequences of notes featuring shifts from low to high pitches was conveyed through upward hand movements, and shifts from high to low pitch led to downward hand movements. As a result, gestures conveyed spatial information about pitch such that higher pitches were allocated to higher locations compared to lower pitches. The second experiment aimed at ruling out the possibility that speaking (e.g., labels such as 'high', 'low', 'up', or 'down') shaped this gestural behaviour. Participants were given an articulatory suppression task which required them to hum the sound sequences and therefore prevented them from employing verbal descriptions. The gestural behaviour was consistent with that observed in the previous experiment. In summary, the gestures produced while communicating about sound pitch emerged from the spatial representation of pitch in the auditory modality, and this was independent of the speech production process. Chapter 2 will conclude discussing the implications of these findings on gesture production models, with a focus on the GSA framework.

Chapter 3 tested whether the action component of the mental simulation underlying gestures can alone predict variations in the gesture rate. The GSA claims that "as one moves from visual images through spatial images to motor images, the amount of action simulation increases and so does the likelihood of gesture" (Hostetter & Alibali, 2008, p. 510). Results suggest that the strength of action simulation and visual simulation together might predict variations in the gesture rate. However, the experiment did not have enough power to provide definite evidence for it, or to disentangle the effect of the action component from the effect of the visual component on gesture rate. These results are discussed in relation to future research. Finally, Chapter 4 provides a brief discussion of all the findings of earlier chapters and their theoretical meaning.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the 'Copyright') and she has given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patens Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form a part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade markers and other intellectual property (the 'Intellectual Property') and any reproductions of copyright works in the thesis, for example graphs and tables ('Reproductions), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialization of this thesis, the Copyright and any Intellectual Property and /or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any Relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in the University's policy on Presentation of the Theses.

To my parents and to Phil

Acknowledgements

I would like to express my gratitude to my supervisors, Dr. Louise Connell and Dr. Judith Holler, whose expertise, understanding, and patience, has made it possible for me to write this thesis and to grow personally and professionally. I would like to thank other members of the School of Psychological Sciences at the University of Manchester that have contributed to my improvements, especially Dr. Andrew Stewart and Dr. Thea Cameron-Faulkner, as well as the external examiner Dr. Max Louwerse. I thank also all my colleagues, especially Hilda Osafo Hounkpatin, Cintia Faija, Jeffrey Wood and Samantha Rowbotham for their help and enthusiastic support throughout the whole project. I would also like to thank members of staff, especially Mike Bossons and

Emma Braithwaite and anyone who smiled to me during my time in Coupland Building 1 during these years. And last, but not least, I would also like to thank my family for all the support provided for 28 years now and my old friends and my boyfriend Phil who have been like a second family for me. I also recognize that this research would not have been possible without the financial assistance provided by project grant from

Leverhulme Trust (F/00 120/CA) awarded to Louise Connell and Judith Holler.

Chapter 1: Introduction

The present chapter introduces the theoretical background which motivated the experiments included in the following chapters. The discussion is divided into three main sections. The first section aims at providing an introduction to the phenomenon of co-speech gestures by providing a definition and some examples. The second section of the chapter presents the most influential models of gesture production. It highlights their assumptions and predictions, and discusses similarities and differences. Finally, the chapter closes with an outline of the research that will be presented in the following two chapters.

1.1. The phenomenon of co-speech gestures

While speaking, people frequently produce spontaneous body movements whose meaning often relates to the meaning that is expressed in the accompanying speech: such movements are known as *co-speech gestures* (McNeill, 1992). For example, a speaker may say 'I was drinking my cup of tea' while producing a C-shaped hand gesture that represents the act of holding a cup of tea. Co-speech gestures, henceforth *gestures*, are movements produced with one or more parts of the body, for example the hands or the head, which coordinate in *meaning*, *time* and *function* with the concurrent speech. Because of their tight binding to speech their origins and function have often been studied in conjunction with speech.

With respect to their timing, gestures often synchronize with speech in such a way that the part of the gestural movement that expresses meaning, called the *stroke phase*, co-occurs with or, more rarely, slightly precedes the segment of speech that expresses related semantic content (McNeill, 1992; Nobe, 2000). For example, when one says 'I was drinking a cup of tea', the C-shaped gesture occurs while uttering 'cup'. This example illustrates the semantic relation between speech and gesture:

speech and gesture are co-expressive, in that they convey very similar information, in this case 'holding a cup of tea'. In this example the gesture is said to be *redundant* with respect to the information that it conveys as it does not add any semantic information to the one that is already conveyed in speech (Bavelas, Kenwood, Johnson, & Phillips, 2002). By contrast, *non-redundant* gestures express meaning that is not conveyed in speech; for example when a speaker says 'We were sitting by a round coffee table' while the right hand, palm facing the floor, indicates the height of the coffee table. In this example, the gesture is non-redundant because it conveys additional information that is not expressed in speech, i.e., information about the height of the table. Finally, in addition to being temporally and semantically related, gesture and speech closely cooperate with respect to their discourse and interactional function. Gestures can have a discourse structuring function, for example by stressing those parts of the narrative which are the most salient (McNeill, 2000). Gestures can also play an interactional function, for example by helping to manage roles in conversational exchange like nodding gestures do (Bavelas, Chovil, Lawrie, & Wade, 1992).

The movements considered henceforth are classified as gestures. However, not all body movements are gestures. There are important traits that mark the distinction between gestures and other bodily movements. McNeill (1992) distinguishes gestures from other body movements on the basis of presence/absence of speech. He places gestures along what he calls *Kendon's continuum:* gestures are those movements with a higher degree of conventionalization and language-like structure (McNeill, 1992). To start with, not all bodily movements occur alongside with speech, while gestures generally do. The co-occurrence between speech and gesture is important when distinguishing gestures from other movements of the arms

and hands that express also meaning but that do not co-occur with speech. An example of such movements is represented by *pantomimes*: these are movements of the body that represent meaning but are performed while nothing is being uttered. For example, one may imitate the movement involved when climbing up a tree while being silent because the context requires this. Next along Kendon's continuum are *emblems*, symbolic movements that have a conventional meaning, such as the 'thumbs-up'. These are not considered to be gestures because they have a conventionalized meaning that can be understood when there is no accompanying speech, e.g. the word 'OK'. At the far end of Kendon's continuum are *signs*, bodily movements that are employed in sign languages: these are part of an autonomous linguistic structure and have a codified meaning for users of that sign language. Further, gestures (as well as all other forms of body movement mentioned above) are fundamentally different from non-communicative movements such as scratching one's nose, adjusting one's hair, or playing with a pen while talking (Ekman & Friesen, 1972).

1.2. Classification of co-speech gestures

When it comes to communicating, there is a wide variety of information that gestures can express: for example, actions (e.g., 'running'), spatial relations (e.g., 'being next to'), concrete objects, (e.g., 'cup') or abstract entities (e.g., 'inflation'). By changing some features of the gesture, such as the hand shape, orientation, location, motion direction, and so on, one can expresses different meanings (McNeill, 1992). For example, a left-right movement, flat hand facing the floor may indicate 'flat surface', whilst an upwards movement, flat hand facing the floor may indicate something 'growing upwards'. So, variations in the gestural movements relate to

variations in the meaning and function of the gesture. Given the variety of gestures that one can produce, McNeill (1992) identified patterns that can be used to classify gestures. Accordingly, gestures can be classified into *iconics*, *metaphorics*, *deictics* and *beats*.

Iconic gestures have a formal resemblance with the object or event that they depict: for example, one may produce a C-shaped gesture to represent 'holding a cup' where the physical form of the gesture resembles the physical shape of a hand holding a cup. While iconic gestures represent concrete objects and events, metaphoric gestures are used to represent abstract, non-physical entities, such as 'inflation rises'. Such abstract concepts are depicted by means of a concrete form. For example, the concept of 'inflation rises' can be conveyed through an upward hand movement. Another type of gesture is deictics. Deictic gestures point at physical locations, or at objects that are physically present at the moment of speaking (concrete deictics); alternatively, they point at a location in a fictive space that is taken to represent the location of an entity that is not present at the moment of speaking (abstract deictics). The prototypical deictic gesture is the one that is performed by extending the index finger. Iconics, methaporics and deictics all are *representational gestures*, in that they visually represent a referent that can be either a concrete or an abstract entity (Alibali, Heath, & Myers, 2001). While some representational gestures may perform referential or pragmatic functions, *nonrepresentational gestures* exclusively perform pragmatic functions. This category contains beat gestures, which are rhythmic hand movements performed with a flat, often lax, hand-shape aligning with some syllables or words that are being stressed in speech (McNeill, 1992).

1.3 Co-speech Gesture production models

The previous section described the relationship between speech and gestures as being characterized by a tight binding with respect to their meaning, timing and function. The observation of these characteristics led to the claim that speech and gesture form an inseparable unit (McNeill, 1992). This claim finds empirical support especially in studies on language development and language impairments (e.g., Hill, 2001; Iverson & Goldin-Meadow, 2005). First, speech and gesture develop in parallel in children (McNeill, 1985). Studies on language development show that gestures emerge together with early words, and communicative gestures even precede and anticipate the production of the first words (e.g., Iverson, Capirci & Caselli, 1994). Second, language impairments resulting from brain damages are mirrored by impairments in the production of hand movements (Kimura & Archibald, 1974). Taken together, this evidence supports the idea that "gesture and speech might form a single, integrated system in which the two modalities work together to convey meaning" (Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001, p. 516).

The claim that speech and gesture emerge from the same cognitive system implies that the same thought is laid out not only through speaking but also through gestures. In other words, the same thought comes to be expressed in two different modalities, the speech modality and the gesture modality. Several models of gesture productions share this assumption. The Growth Point Theory (GPT) holds that both speech and gesture emerge from the same *unit of thought* (McNeill, 1992). Units of thought, called *growth points*, combine imagistic content, often visuospatial and actional, with linguistic categorical content (McNeill, 1992). Growth points generate from the so-called *fields of opposition* between conceptual contents: a growth point is formed when a thought emerges from background context. The following example

illustrates the formation of growth points and the ensuing emergence of speech and gestures. The speaker is describing a cartoon and says:

2a) 'and so he tries to get in the building again'

2b) '[**he crawls up a pipe**]'

In this example the speaker describes a sequence from the cartoon 'Tweety Bird' (McNeill, 1992, p. 199). The sequence shows Sylvester the cat trying *different* ways to get to Tweety the bird. After several attempts, the cat tries to reach the bird by crawling up the pipe. The field of opposition here is created by the context itself: 'crawling up' is another way – a *new* one - to try and get to Tweety. In other words, 'crawling up' is a distinct action from the other actions that the cat previously performed and, as such, it stands out against the background of all the other actions aimed at reaching the bird. The cat performing a new action represents the field of opposition from which growth points generate. The sentence marked in bold and within brackets indicates the point where a gesture is performed. The gesture consists of an upwards hand movement with the hands folded into a hollow oblong shape representing the *interior* of the pipe, or the cat *inside* the pipe. In McNeill's theory, therefore, the context plays a major role in determining what is going to be expressed in speech and gestures.

McNeill's theory represents a sound attempt at studying the gesture in conjunction with speech. However, it presents two main limitations. First, the GPT focuses on gestures which are produced alongside speech, leaving untold what originates originate gestures that do not occur alongside speech. Second, as to the content of the perceptive representations which underlie gesture, the GPT gives relevance to visual content over other perceptive contents, such as auditory ones. Such bias might be due

to the fact that a significant part the Growth Point Theory relies on data that consist of coded descriptions of cartoon stories. As a consequence, the theory focuses on the gestures which encode visual and spatial information embedded into a cartoon.

The idea of an *interface representation* between speech and gesture is at the core of Kita and Özyürek's model for gesture production (Kita & Özyürek, 2003). The so-called Information Packaging Hypothesis (IPH) shares with the Growth Point Theory the assumption of a dialectic between gestures and speech. The IPH holds that gestures emerge from spatio-motoric representations, while speech emerges from linguistic (analytic) content. The two models differ with respect to one fundamental assumption. The GPT assumes that gestures simply *represent* thought – in particular spatial and motoric thinking. The Information Packaging Hypothesis, instead, assumes that gesture not only *express* thought, but they also *organize* it. Gesture helps organizing thought so that it makes it easier to verbalize. By assuming that gesture influences speech by organizing information, the model predicts that the information that is going to be encoded in gestures influences what is going to be expressed in speech (Kita & Özyürek, 2003). Therefore, performing gestures plays a substantial role in Kita's model: gesture expresses thoughts and influences it. On the one hand, gesture arises from actional thinking, for example when one is thinking about the interacting with objects and/or moving one's own body (Kita & Özyürek, 2003). On the other hand, performing a gesture helps speaking by packaging information for speaking. The IPH somehow expands the GPT by clarifying further how speech and gesture influence each other.

The idea that producing a gesture somehow facilitates speech production also characterizes the Lexical Access Model (LAM) according to which gesture facilitates lexical access (Rauscher et al., 1996). The LAM theorizes that propositional content

in the working memory is conceptualized and verbalized; and, spatio-dynamic features of a concept are selected and inform a motor program for the planning and execution of a gesture. Crucially, this model assumes that when the speaker struggles with word retrieval, producing a gesture should help him with the task. Producing a gesture is assumed to activate spatio-dynamic features of a concept and activate the lexical affiliate of that concept. In other words, producing gestures facilitates speech through cross-modal priming which occurs when speech becomes difficult. The Lexical Access model differs from the Information Packaging Hypothesis in one main theoretical assumption. According to the LAM, gesture influences speech only when a speaker is struggling with the retrieval of a word. Therefore, the Lexical Access model assumes a one-way interaction between speech and gestures, which occurs only when speech becomes difficult. On the contrary, the IPH assumes a constant interaction between speech and gestures.

The three models presented above do not mark a difference between the role of spatial thinking and actional thinking in gesture production. That is, they do not make an explicit claim on whether thinking about actions and thinking about spatial information have a different impact on gesture rate. In other words, representing actions leads to producing gestures as likely as representing visuospatial information. As we shall see briefly, a difference between these two representational contents is at the centre of the Gesture as Simulated Action Model for gesture production.

1.4 Simulated actions and gesture production

Recent developments in gesture studies further advance our understanding of the mental representations underlying co-speech gestures. The Gesture as Simulated Action Model (GSA) claims that gesture production is the result of *embodied*

representations. The central assumption is that language is embodied, and "cognition depends crucially on having a body with particular perceptual and motor capabilities and the types of experiences that such a body affords" (Iverson & Thelen, 1999, p. 19).

To start with, embodied theories of language production hold that the conceptual system, that is the system representing knowledge about the world, organizes knowledge throughout the different brain modalities (Barsalou, 1999; Glenberg & Kaschak, 2002; Zwaan, 2004). For example, the concept 'dog' is represented in the visual modality, where information about what the dog looks like is captured; in the auditory modality, where information about its bark is recorded; and in the tactile modality, where information about how its fur feels is recorded (Connell & Lynott, 2010). Once a multimodal representation of the concept 'dog' is in place, it provides representational support across the spectrum of cognitive tasks, among which are language production and language comprehension.

The representational support provided by multimodal representations consists of the instantiation of modality-specific states during language production and comprehension. The mental states that arise during online experience, for example when seeing a dog, are partially re-activated in its absence, for example when speaking of it. The reactivation occurs because modality-specific states that arise during online experience are partially recorded and then later used to represent (or *simulate*) these situations/concepts during off-line processing, such as when remembering, speaking and thinking. This means that the raw material of language is the multimodal representation of concepts (Barsalou, 2010). Embodied theories of language production have found extensive empirical support (e.g., Barsalou, 2010; Fischer & Zwaan, 2008; Hauk, Johnsrude, & Pulvermüller, 2004). Studies show that

language production is supported by modality-specific activations. For example, uttering a word which denotes an action, such as 'pick' activates the same motor areas that are activated when performing the action of 'picking' (Hauk et al., 2004). The phenomenon by which verbal descriptions of actions activate motor areas involved in performing those actions is named motor resonance (Taylor & Zwaan, 2008). A similar phenomenon is assumed to take place also with perceptions: for example, the same neural substrates supporting visual perception are also activated to recreate the experience of seeing an object in its absence (Kosslyn, Thompson, & Alpert, 1997).

The Gesture as Simulated Action Model proposes that co-speech gestures also arise from simulating multimodal representations (Hostetter & Alibali, 2008). Accordingly, gesture arises from the same perceptual and motor simulations that underlie speaking. The fundamental difference between the GSA and the models that I described previously is that the GSA gives action simulations a role of prominence in gesture production. While the models presented previously would predict that spatial or actional thinking are equally likely to lead to gesture production, the GSA assumes that representing actions is more likely to give rise to gestures than visuospatial representations. The GSA claims that gestures arise from the representation of actions or perceptions that are linked to action. With respect to the process that leads from multimodal representations to the production of gestures, the GSA assumes that gestures arise from simulated actions and from simulated perceptions that then lead to simulated actions. For example, when a speaker is talking about holding a cup and producing a 'holding-like' gesture, the gesture arises from representing the action 'holding a cup'. One of the predictions of this view is that a speaker may gesture more when his conceptual representation is strongly about actions. All things being equal, the factor that affects the rate at which one gestures is

the extent to which action information is represented. According to the GSA, more action activation means more gesture production because accordingly, activation spreads from pre-motor to motor areas and therefore results in gestures being performed. Evidence for this claim is that speakers produce more gestures when describing actions than when describing visual perceptions (Hostetter & Alibali, 2010).

The GSA outlines several different conditions under which perceptual information may evoke an action simulation: when a visual image changes over time, changes visual perspective, moves, has action affordances, or when the act of perceiving the object itself requires movement. Therefore, motor simulation is the necessary condition for a gesture to arise and motor activation occurs when a) one simulates actions and b) one simulates visual perceptions that are informed by actions. On the GSA view, seeing is already an action: simulating actions activates pre-motor areas and if activation spreads to motor areas a gesture will be produced. Activation is more likely to spread when thinking about actions: for example, thinking about the shape of a cake involves motor activation (action activation) because perceiving the cake's contours requires eye movement, and motor activation which arises from seeing in turns informs actions.

1.5 Thesis Overview

The Gesture as Simulated Action (GSA) Model (Hostetter & Alibali, 2008) holds that gestures result from *simulated actions,* which are in turn led by simulations of motor representations and perceptual representations. To start with, the GSA framework holds that the perceptual and motor simulations that underlie speech production also underlie the production of gestures: for example, when speaking of a cake, simulating visual representations (e.g., seeing a cake) and motor representations

(e.g., holding a piece of cake) may lead to the simulation of actions, which may result in the production of gestures. In particular, when the simulation of motor representations crosses a certain threshold, a movement is produced. A second assumption is that, the more action-oriented the representation, the higher the likelihood that gestures will be produced. For example, speaking about holding a cake should lead to a higher rate of gesture than speaking about the colour of a cake (Hostetter & Alibali, 2008).

Following this claim, this work investigates the role played by a) audiospatial representations and b) visual and action representations in the production of gestures depicting sound-related information. In particular, the research presented in this thesis investigates three main assumptions. First, the GSA concentrates on the link between the simulation of visuospatial representations and the production of actions. The first experiment (Chapter 2) focused on how and whether the simulation of audiospatial representations, in particular the simulation of auditory pitch, can also lead to gesture production. Second, it has been assumed that speech and gesture emerge from the same underlying mental representation, in such a way that what is conveyed in speech influences what is conveyed in gesture (e.g., McNeill, 1992; 2005; Kendon, 2004; Kita, 2000). For example, describing the notes of a melody as being 'higher' or 'lower' may be driven by the use of words such as 'higher' or 'lower'. The second experiment (Chapter 2) tested this assumption. Finally, the GSA predicts that describing something related to action, such as holding a cake, would lead to more gesticulation than describing something less tied to action, such as the smell of a cake. This hypothesis is based on the assumption that the degree of action activation explains differences in the gesture rate. Experiment 3 (Chapter 3) tested this hypothesis by manipulating the degree of action information elicited by the

representation of different sounds: for example, describing the sound produced by someone bouncing a basketball should lead, as predicted by the GSA, to a higher gesture rate than describing the sound produced by the wind.

Experiment 1 investigated the nature and the role of the audiospatial representations involved in the production of co-speech gestures depicting information about auditory pitch. The GSA framework (Hostetter & Alibali, 2008) concentrates on how one type of perceptual simulation, i.e., visuospatial, leads to actions. Seeing an object, such as a cake, informs the motor system about what movements are involved to perceive and/or interact with the cake, which in turn leads to motor representations, and thus to gesture production (Hostetter & Alibali, 2008). By assuming this, the GSA leaves untold how other perceptual simulations, i.e., those that are not visuospatial can also lead to gesture production. The objective of the first study is to determine whether audiospatial representations lead to the production of gestures. To answer this question, Experiment 1 used different pitched sounds as target stimuli. Pitch variation is particularly suitable for investigating whether spatial representations other than visual lead to gesture production, given that previous research has determined that the representation of pitch is audiospatial in nature (see Connell, Cai & Holler, 2013). Following these findings, the first experiment involved participants in a verbal description task which required them to describe sequences of musical notes that varied in pitch. When describing the sequences of sounds, participants produced upwards hand movements to represent shifts from lower to higher pitches and downwards hand movements to represent shifts in the opposite pattern, thus indicating that audiospatial representations also lead to co-speech gesture production. Experiment 2 further investigated these findings. The experiment was used to rule out the possibility that encoding spatial information through speech (e.g.,

words such as “higher/lower”) may have driven the pattern obtained in Experiment 1. Participants completed an articulatory suppression task that required them to hum the sound sequences used in the first experiment. If gestures encode spatial information about pitch independently from speech, then speakers’ gestural depictions of the sounds should stay the same when just humming the melodic sequences instead of describing them. For example, shifts from lower to higher pitch should lead to upward hand movements. Results showed that patterns of gestural depictions remained the same as in Experiment 1, suggesting that gestural encoding of audiospatial information is independent from linguistic encoding.

Experiment 3 explored the cognitive factors that affect the rate at which a speaker gestures, focusing on the effect of different simulations on gesture production. Previous research suggests that, all things being equal, stronger action simulation is associated with a higher gesture rate, whereas weaker action simulation leads to a lower gesture rate (e.g., Hostetter & Alibali, 2010). For example, thinking about walking in the rain should lead to a greater production of gestures than thinking about only looking at the rain from a window. Experiment 3 tested this hypothesis by measuring gesture rate in a task that manipulates the strength of the action component of a simulation. Participants described 30 sounds that varied with respect to how strongly their representation involved action. Sounds could be high in both action and visual strength (e.g., someone brushing their teeth), high in visual strength but low in action (e.g., fire crackling), or low in both visual and action strength, (e.g., synthesized scrambled sound). Results did not give definite support to the GSA assumption that the strength of action simulation alone predicts variations in gesture rate. They seem to suggest that together action strength and visual strength might

predict gesture rate. This result might be due to the experimental design will be discussed in reference to further research.

In conclusion, taken together the results presented in this dissertation seem to suggest that gesture production results from a more complex set of mechanisms than the one proposed by the GSA. First, it shows that action simulations may also result from audiospatial simulations. This result urges a revision of the GSA; the framework needs to address more specific cases when considering what perceptual simulations can lead to gesture production. Second, it shows that the gestural encoding of audiospatial information is independent from the verbal encoding. Third, the degree to which the component of the mental representation is about action or visuospatial perceptions seem to predict variations in the gesture rate. However, further research is needed to disentangle the effect of action representations from the effect of spatial representations on gesture rate.

Chapter 2

Gesture and auditory pitch

2.1. Introduction

The present chapter presents two studies which aims at addressing the question of whether and how information other than visuospatial leads to the production of gestures. According to the GSA, gestures are produced as the result of simulated actions, which in turn occur in a variety of instances. Firstly, simulated actions occur when simulating seeing an object as visually perceiving an object involves body movement, such as eye movement. For example seeing a cake may involve eye movement to follow the cake's contour. This is an example of a (visual) perceptual simulation that is informed by actions. In this case, a gesture may be produced which depicts the eye movement around the cake's contour. Secondly, when imagining how an object would look from a different perspective. Changes in perspective may involve the representation of physical motion that is needed to perform the change in perspective. The resulting gesture might represent the movement required to represent how the object would look from a different spatial location. Thirdly, performing mental transformation of visual objects might involve the representation of physical motion, such as that required to manipulate the object itself. So, imagining rotating a three-dimensional object might lead to the production of a gesture in which the index finger and thumb are opposed as if they were holding the object. Fourthly, representing spatial properties of objects may activate information about how to interact with them. For example, imagining spatial information about the shape and size of a cup activates information about how to hold it, the so-called object *affordances* (Ellis & Tucker, 2000). In this case, a gesture can

be produced which mimics the action of holding the cup. Finally, simulated actions occur when imagining one's body in motion (e.g., when imagining landing a triple axel). In each of these cases, simulating seeing or acting activates a motor program. When activation in the premotor areas spreads to motor areas, a gesture is produced. In particular, according to the GSA a gesture is produced when the amount of action simulation exceeds a certain threshold.[1]

So, the GSA claims that gestures are produced as a result of perceptual simulations that are either informed by actions or that lead to actions. One thing to be noticed is that the GSA seems to classify the representation of any body movements as the representation of an action. Crucially, the GSA seems to suggest that the route to representing actions passes through visual perceptions. In its current form, the model assumes that "although mental images can correspond to any of the senses, *visual mental imagery* and *motor mental imagery* have the clearest relationships with perception and action" (Hostetter & Alibali, 2008, p. 499). The suggestion is that, of all perceptual modalities, the visual modality is the one with the strongest link to action. This assumption does not take into account possible cases when actions may be elicited by other perceptual simulations. The current study has been designed to clarify this point by addressing the question of whether spatial representations that are not visual can also elicit gesture production.

To further investigate this issue, the two studies presented in this chapter will make use of auditory stimuli, specifically different pitched sounds. The rationale behind this choice is that "when people hear a musical note, its pitch is not just represented in the auditory modality. Rather, its representation is audiospatial, in that it comprises both an auditory and a spatial representation of the note's frequency"

[1] The gesture threshold as described within the GSA will be discussed in Chapter 3.

(Connell et al., 2013, p. 129). While The GSA focuses on how visuospatial representations lead to the production of gestures, the present study investigates whether audiospatial representations can lead to gesture production.

With regard to auditory pitch, this can be defined as the attribute of auditory perception which makes us perceive sounds as being high or low. Numerous studies have shown that the dimension of auditory pitch and the dimension of vertical space are associated (Lidji, Kolinsky, Lochy, & Morais, 2007; Rusconi, Kwan, Giordano, Umilta, & Butterworth, 2006). For example, when people hear individual sounds that differ in pitch and are asked to point at the location in space where they think the sound originates, they point to higher locations for higher pitched sounds relative to lower pitched sounds, despite the sound source being in a fixed location (Pratt, 1930). Nonmusicians respond faster to high tones when pressing a key in an upper location and to lower tones when pressing a key in a lower location (Rusconi et al., 2006). The same effect appears even when they are asked to identify the musical instrument that produced the note rather than make an explicit pitch judgment. These results suggest that there is a connection between the domain of pitch and the domain of vertical space. To examine whether the representation of pitch is *obligatorily* spatial in the vertical axis, Connell, Cai and Holler (2013) asked participants to judge whether the pitch of a target note was higher, lower or the same as a previous cue note. Crucially, target notes were presented in a video where the actor sang the note while performing upward or downward hand movements. Results showed that seeing concurrent motion in the vertical axis biased participants' pitch discrimination so that upward hand movements made notes appear higher in pitch and downward hand movements made notes appear lower in pitch. The effect disappeared when participants were given a spatial memory load, but was unaffected by a verbal memory load. These findings

demonstrate that musical pitch processing is intrinsically audiospatial, meaning that spatial information is intrinsic to pitch representation. The evidence discussed adds to the body of literature showing that pitch processing and spatial processing requires overlapping areas (e.g., Arnott & Alain, 2011; Degerman, Rinne, Särkkä, Salmi, & Alho, 2008). Such results are compatible with the view that there is a supra or multi-modal spatial representation system that integrates spatial representations that occur across modalities (e.g., (Bryant, 1992; Giudice, Betty, & Loomis, 2011; Renier et al., 2009). In its current formulation, the GSA would not make any empirical prediction about *whether* the audiospatial representation of pitch could give rise to gestures. The present chapter addresses this question. In two experiments, I asked participants to describe single notes and sequences of two to four notes to an addressee. The notes differed with respect to their pitch, and all of the sequences had a different melodic contour. In Experiment 1, the participant was asked to describe sequences of notes and was not allowed to sing. If the audiospatial nature of pitch representation gives rise to gestures, then we predict that pitch would be depicted in gestures so that ascendant contours (i.e., from a lower to a higher pitch) will be expressed with upward hand movements and descendent contours (i.e., from a higher to a lower pitch) will be expressed with downward hand movements. The second experiment was designed to eliminate the possibility that verbal labelling (e.g., the use of words such as 'high' or 'low'), would influence the gestural behaviour. Here participants were given an articulatory suppression task, namely humming the sound sequences. If gesture expresses spatial information about pitch *independently* of the linguistic processing, then we expect that the gestural depiction of auditory pitch would still emerge with the same patterns even when language processing is inhibited and participants are involved in a non-linguistic task.

2.2. Experiment 1

2.2.1. Participants

Twenty pairs of participants (16 female pairs) took part in Experiment 1. All participants were students or staff from the University of Manchester. The average age was 22 years. All participants were right-handed and native speakers of English and gave written informed consent to participate in the study and to be video recorded. Each participant completed a questionnaire to determine their degree of musical experience, such as their history of playing an instrument, singing in a choir or reading music notations. All these experiences may influence the gestural behaviours such that, for example, experience with playing an instrument may result in mimicking the action of playing when describing the musical sequences. Two participants (and their addressees) were replaced for having extensive experience in at least one domain of musical training that could have influenced their pitch representation (e.g., experience of playing a piano leads to horizontal pitch mapping (Lidji et al., 2007), and their gestural behaviour (e.g., mimicking the act of playing the instrument). Additionally, another participant (and their addressee) was excluded because during debrief at the end of the experiment she guessed that the focus of the experiment could have been non-verbal communication.

2.2.2. Stimuli

Stimuli consisted of short sequences of musical notes sung by multiple human voice choirs (male and female). Human voices were chosen over musical instruments to prevent unwanted influences on gestural behaviour, such as participants mimicking playing an instrument instead of describing the changes in pitch of the notes that constituted the sequence. Multiple voices were preferred over single voices in order to avoid participants describing the nature of the singer (e.g., 'it's a woman singing').

Each note lasted 0.75 sec, and there were no pauses between notes in a sequence (i.e., stimuli lasted from 0.75 to 3 sec). The fundamental frequencies of notes varied from 110.0 Hz to 293.7 Hz (A2-D4 in music notation), and were selected to reflect the range of male and female vocal frequencies. Sounds were created using Garage Band Software for Macintosh. Sound duration was edited using Audacity Software

(http://audacity.sourceforge.net/).

The stimuli materials were organized into two blocks of eight trials each. Each trial consisted of a single note or sequences of two, three or four notes, and all of the sequences had a different melodic contour. Each block contained two trials of one note, two trials of two notes, two trials of three notes and two trials of four notes. The order of presentation of the sequences was not randomized. Instead, trials were ordered from the single-note sequences to the four-note sequences with the aim of increasing the level of the complexity of the stimulus that the participants had to describe. The sound files were embedded into a Microsoft PowerPoint presentation so as to allow participants to move from one sequence to the next in a self-directed manner. Each PowerPoint slide had a blank page with a number in order to inform participants of their progression through the slides. Each laptop had a set of headphones plugged in.

2.2.3. Procedure

Participants were told that they would take part in a study that investigated how people communicate musical sounds to an addressee. Neither ‘hand movements’ nor ‘gestures’ were mentioned in the instructions. Participants entered the lab in pairs and were invited to sit down opposite each other in two chairs placed across a low coffee table. Participants were randomly allocated to the role of describer or

addressee. The two participants from each pair were then randomly assigned to one of the two separate blocks of eight stimulus trials. Stimuli were presented over headphones and describers heard each sequence twice before describing it to their addressee. Participants wore the headphones from the beginning to the end of the experiment. They were instructed to adjust the volume to a comfortable level before the experiment began to ensure that sounds played to the wearer could not be heard by the respective other participant. A laptop was located to the right hand side of each participant. The describer was asked to say the number of the slide aloud before describing the sequence. Both participants (describer and addressee) were asked to listen to each sequence simultaneously. Sound sequences were presented one at a time, and participants pressed a button on the laptop to play the sound. After a 1-sec delay the sound sequence was automatically repeated. After the second play, the describer was asked to describe the sound stimulus to their addressee. Participants were told that the sound sequences they had heard could be the same or different from each other. The addressee's task was to indicate whether the sequence s/he heard was the same as or different from the one described by the other person. The addressees provided their judgment on a sheet of paper that was only visible to them in order to avoid this information serving as feedback to the describer, which may consequently have influenced their gestural behaviour over and above the nature of the sound stimulus. Before the start of the experiment, the describer and the addressee were reminded that all of their sequences were multi-voice sung notes and that on each trial they had the same number of notes, while the sounds differed in pitch. This information was provided to them in order to avoid describers focusing on several characteristics of the sequence other than pitch (e.g., number of notes). Before the experiment started the pairs of participants had two practice trials so that they could

adjust the volume and familiarize with the task. Participants were video-recorded unobtrusively on a split screen using two wall-mounted cameras, and were informed about the presence of the cameras before the experiment started, in line with the British Psychological Society's (2009) Code of Ethics and Conduct. The experimenter stayed in the adjacent control room for the duration of the experiment in order to not be perceived as an over hearer. At the end of the experiment, participants received feedback on their performance, were thanked, debriefed and compensated with £5 or course experiment credits. The entire experiment lasted approximately 25 minutes.

2.2.4. Coding

For this experiment, the coding procedure involved four main steps.

Coding speech. This involved a *verbatim* transcription of the speech produced during the verbal description.

Coding gesture. This coding step involved gesture identification and classification. First, I identified all movements of the hands, arms and other parts of the body, such as the head, and distinguished those that represented meaningful information from non-gestural movements. Then, I classified gestures for their gesture type. Gestures were classified for type into two categories: *representational* gestures and *nonrepresentational gestures* (Alibali et al., 2001). In this study, representational gestures either represented the spatial location of each individual sound, or they depicted the melodic contour of an entire sequence. For example, a three note sequence with the first and last note of the same pitch and the middle note higher, could be represented either using an individual gesture for each of the notes, for

example three pointing gestures; or with one gesture representing the sequence as a whole, with the two hands making a triangle (thumbs and index fingers in contact). Non-representational gestures were those gestures that served a function other than that being representative, and had a pragmatic function. They were for example nodding gestures expressing agreement or mutual understanding between describer and addressee, or beat-like movements that marked some phases of speech. Nonrepresentational gestures were coded but not included in the analysis as their analysis was beyond the scope of the present research.

Coding stimuli. The third step focused on the stimuli being described and involved annotating the number of sounds that constituted each sequence and annotating the frequency of each of the notes constituting the sequence. This way the coding could make visible the relationship between the pitch shift and the location of each individual gesture.

Coding gesture location. Finally, I identified the spatial location of all of the gestures representing two successive notes and compared their location with respect to one another. For example, given a two-note sequence where the first note had a lower pitch than the second note, I coded whether the second gesture occurred higher, lower or in the same location along the vertical axis compared to the first gesture. All of the coding was carried out using the multimedia linguistic annotation tool Elan Linguistic Annotator, Version 4.1.0 (http://www.lat-mpi.eu/tools/elan/).

Reliability. A second coder who was blind to the experimental hypotheses coded gestures in a random sample of video files, corresponding to the 25% of the total number of gestures (157). She was asked to identify those body movements that fall into the category of gestures, to distinguish between representational and

nonrepresentational gestures, and to compare the location of each gesture in a sequence. In order to avoid biases in the coding that may arise from attending to the content of speech (e.g., words such as 'high/low'), the second coder was not allowed to play the sound of the video nor was she aware that the participant was describing sequences of sound. Inter-rater-reliability was calculated and resulted in Cohen's k = .79 (96.8% agreement) for gesture identification, k = .82 (96.4% agreement) for type, and k = .89 (94.9% agreement) for location. These results indicate a substantial level of agreement for gesture identification and gesture type and an almost perfect level of agreement between the two independent coders for gesture location (Landis & Koch, 1977).

2.2.5. Analysis

Following gesture coding, I isolated and excluded from the analysis a) gestures that represented individual sounds, i.e., sounds that were not part of a sequence,[2] b) gestures that depicted the first sound of a sequence, and b) the gestures that represented the entire sequence in one gesture only: for example a gesture performed with two hands in the shape of a triangle representing a 3-notes sequence. I excluded gestures that represented the first sound of a sequence because it was not possible to compare their location along the vertical space with preceding and subsequent gestures. I excluded gestures that depicted entire sequences of notes because they were a *holistic* rather than *discrete* depiction of the sequence; therefore, single tones of the sequence were not depicted individually. I then calculated the percentages of the gestures had and isomorphic correspondences with the pitch that they represented (Table 1).

[2] The set of stimuli included 1-2-3-4 sound sequences. Only gestures representing 2-3-4 sound sequences were analysed.

2.2.6. Results and Discussion

The hypothesis for this experiment related to the effects of pitch height on gesture location: higher pitches (compared to the previous pitch) should lead to gesture occurring in a higher location along the vertical axis (compared to the previous gesture); lower pitches should lead to gestures in a lower location, and; no difference in the pitch height should lead to gestures occurring in the same location along the vertical axis. To test this hypothesis, a 3X3 analysis of variance (Anova) was conducted. The factors of the analysis were: the pitch height (higher/same/lower) and the gesture location (higher/same/lower). Results showed a significant effect of pitch height on gesture location as expected, $F(1,18) = 62.24$, $p = .000$, $\eta^2 = .776$. To explore further these effect, three one-way Anova were performed: they showed that higher pitched sounds led to most gestures to occur in a higher space, $F(2,54) = 42.02$, $p < .001$; same pitched sounds led to most gestures to occur in the same space, $F(2,54) = 56.62$, $p < .001$ and lower pitched sounds led to gestures to occur in a lower location, $F(2.54) = 32.53$, $p < .001$. Table 2 shows the Means and Standard Deviation each of the interactions.

Table 1. *Descriptive statistics for congruency between pitch shift (higher/same/lower) and gesture location (higher/same/lower) with matches indicated in bold.*

Gesture location	**Pitch shift**			**Total N. Gestures**
	Higher	Same	Lower	
Higher	97.0%	4.4%	3.5%	152
Same	1.0.%	94.8%	0.5%	25
Lower	2.0%	0.8%	96.0%	88

Table 2. *Mean and Standard Deviation for each Anova*

Pitch Height	Gesture Location	Mean	Std. Deviation
Higher	Higher	7.79	5.170
	Same	.05	.229
	Lower	.11	.315
Same	Higher	.26	.452
	Same	1.37	.496
	Lower	.05	.229
Lower	Higher	.26	.452
	Same	.05	.229
	Lower	4.47	2.836

2.3. Experiment 2

The gestures produced in Experiment 1 are co-speech gestures (McNeill, 1992) as they occur alongside verbal descriptions and relate to the semantic content of the concurrent speech. Some might argue that the reason why describers conveyed spatial information about pitch in gestures by moving their hands up and down along the vertical axis is due to the gesture being tightly linked to and coordinated with the linguistic content of speech. For example, the particular linguistic units in descriptions of the note sequences (e.g., use of the adjectives 'high/low', of verbs

such as ‘to rise/fall’, ‘to increase/decrease’, or of adverbs such as ‘up/down’) might have activated spatial representations in their own right. As a result, a gesture might depict spatial information about pitch merely as a result of the linguistic encoding that takes place during the speech production process rather than as a direct result of the audiospatial representation of pitch. In Experiment 2, we examined this possibility by instructing participants not to speak. Instead, participants were asked to communicate the note sequences by humming or singing, which not only removed a strategic need for speech planning but also acted as an articulatory suppression mechanism to block access to verbal labels. If gesture simply encoded spatial information because linguistic items evoked spatial representations of some sort, then we would expect the gestural behaviour to differ from that observed in Experiment 1, namely the elimination of vertical gestures to depict changes in pitch. On the other hand, if the isomorphic correspondence between pitch and gestures along the vertical space stays the same across the two experiments, this means that the spatial content reflected in gestures does not derive from the linguistic items employed in speech.

2.3.1. Participants

Twenty new pairs of participants took part in the experiment (9 female pairs). All participants were students or staff from the University of Manchester. The average age was 27 years. All participants were right-handed native English speakers. All participants gave written informed consent to participate in the study and to be video recorded.

2.3.2. Stimuli and Procedure

The stimuli used were identical to Experiment 1. The procedure was identical to Experiment 1, except that participants were explicitly instructed to communicate

the sequences to their addressee by humming and were asked not to describe the musical sequences in words.

2.3.3. Coding and Analysis

Gestures were coded as per Experiment 1. However, instead of coding gestures that co-occurred with speech, we coded gestures that occurred during humming (i.e., 'co-humming' gestures). As speech was absent, there was no speech transcription.

2.3.4. Reliability

An independent second coder who was blind to the experimental hypothesis coded a random sample of files that corresponded to the 25% of the data. Inter-coderagreement resulted in Cohen's k = .89 (94.9% agreement) for gesture identification, k = .63 for gesture type (93.8% agreement), and k = .92 for gesture location (96.2% agreement), overall indicating almost perfect agreement (Landis & Koch, 1977).

2.3.5. Results and discussion

The prediction for this experiment was that the pitch height (higher/same/low) would lead to changes in the location of the gesture along the vertical axis *independently* of speech. To explore this hypothesis, a 3x3 Anova was conducted. The two factors included in the analysis were the pitch height (higher/same/lower) and the gesture location (higher/same/lower). Results showed a significant effect of pitch height on gesture location, $F(1,17) = 83.78$, $p = .000$, $\eta^2 = .915$. These results were further investigated through three on-way Anova. The first Anova further confirmed that higher pitched sounds led to a significantly higher number of gestures

occur in a higher location, $F(2,54) = 69.61$, $p < .001$; same pitched sounds led to gestures occurring the same spatial location along the vertical axes, $F(2,54) = 70.63$, $p < .001$; finally, lower pitched led gestures to occur in a lower spatial location, $F(2,51) = 52.13$, $p < .001$. As a result, in the absence of speech, gestures still depict spatial information about auditory pitch maintaining the same pattern as when speaking.

Table 3. *Descriptive statistics for congruency between pitch shift (higher/same/lower) and gesture location (higher/same/lower), with matches indicated in bold.*

Gesture Location	Pitch shift			Total Gestures
	Higher	Same	Lower	
Higher	96.0%	3.0%	5.0%	61
Same	1.0%	97.0%	1.0%	21
Lower	3.0%	0.0%	94.0%	42

Table 4. *Mean and Standard For each Deviation Anova.*

Pitch Height	**Gesture Location**	**Mean**	**Std. Deviation**
Higher	Higher	2.94	1.474
	Same	.06	.236
	Lower	.06	.236
Same	Higher	.06	.236
	Same	1.06	.416

	Lower	.06	.236
Lower	Higher	.17	.383
	Same	.06	.236
	Lower	2.22	.647

2.4. General Discussion

The aim of the present two studies was to determine whether the spatial component of auditory pitch would give rise to gestures, and whether the observed gestural depiction would depend upon the ongoing speech, meaning that uttering words that indicate spatial locations, such as 'high' would influence the gestural depiction, for example, by leading to upwards hand gestures. In Experiment 1, when describing in words shifts from lower to higher pitch, participants produced upward gestural movements, thus indicating that higher pitches were allocated to higher spatial locations than lower pitches. On the other hand, shifts from higher to lower pitch lead to downward movements. This suggests that gestures express information about musical pitch such that this is represented spatially along the vertical axis. These patterns also held in Experiment 2, when participants were prevented from speaking, therefore excluding the possibility that linguistic labels produced during speech (i.e., words such as 'high/low') influenced the gestural depictions. In summary, to the question of whether spatial representations other than can give rise to gesture, the answer is positive and the path that leads from audiospatial representations to gesture production seems not to be mediated by linguistic processing.

The results are in line with the general claim that perceptual simulations can trigger gesture production. However, the studies presented here differ from previous research in that they introduce the idea that a different type of spatial representation, that of audiospatial, can also lead to gesture production. The empirical prediction of the Gesture as Simulated Action Model is that gestures are produced as a result of simulated actions. And, crucially, simulated actions are either re-enactment of actions, such as 'walking' or are they triggered by representations that are linked to action, in particular visuospatial. Therefore, it is not possible to derive a clear empirical prediction from the GSA framework with respect to what happens in the case in which audiospatial information is being simulated. To bridge the gap, from the studies presented it can be concluded that action simulations can result also from audiospatial representations. Therefore, further studies are needed to research the nature of the mechanism that leads from audiospatial representations to gesture production. According to the GSA, the perceptual route to the simulation of actions is through visuospatial perceptions.

Next to the GSA, other existing models for gesture production differ with respect to what role they assume speech process plays in gesture production. In Experiment 2 participants expressed information about pitch in gestures by maintaining a vertical spatial representation. This finding shows that gesture expresses spatial information about pitch independently from speech. This result is not compatible with the assumption that speech is necessary for gesture production, as suggested in the Lexical Access Hypothesis (Butterworth & Hadar, 1989). This would predict that gestures generate from the semantic content of selected lexical items. For example, if when describing a pitch shift from lower to higher pitch a speaker uttered 'it goes up' and produced an iconic gesture depicting an upwards trajectory, the iconic

gesture is supposed to result from the semantic features that are part of the lexical item ‘up’. In Experiment 2, participants were prevented from lexical planning as they were given an articulatory suppression task. This means that verbal planning was inhibited and as a consequence no lexical item was been chosen. This indicates that gesture production is independent from speech production, a finding that is not compatible with the view that gesture production is a by-product of the speech production process.

Similarly to the Lexical Access Hypothesis, other models of gesture production hold that speech and gesture are closely related to one another in such a way that the semantic content of a gesture is also influenced by the content of speech (e.g., Kita &

Özyürek, 2003). The core idea for these models is that gesture and speech systems interact in such a way that what is going to be expressed in gesture depends, at least in part, on how the message is encoded in speech. In other words, speech constrains and shapes gestures. However, Kita & Özyürek’s (2003) theory also accounts for the fact that gesture can occur in the absence of speech: “co-speech gestures are generated from an action generation mechanism that is highly coordinated with, but independent from, the speech production system” (Kita & Özyürek, 2003, p. 722). For this reason their theory is compatible with the finding that gestures that encode spatial information about auditory pitch maintain the same pattern (e.g., higher gestures in a higher spatial location) in the absence of speech, suggesting a certain level of independence of gestures from speech. So, Experiment 2 supports the idea that speech and gesture production are the result of two distinct (but interacting) processes. However, one consideration needs to be made about the fact that in the second experiment were produced less gestures than in the first experiment. We can speculate

that two may be the reasons for this difference in the production of gestures. The first reason may be that Experiment 2 might have involved participants in a task which would make them more socially aware, i.e., singing. As a result, they might have been less natural and spontaneous in their behaviour. The second reason might be that speaking and gesturing are so closely related that "When a speaker is engaged in gesturing, the oral and manual system may become increasingly entrained, so that each subsequent utterance may increase its likelihood of being accompanied by a gesture" (Hostetter & Alibali, 2008, p. 506). Accordingly, participants produced less gestures precisely because they were not speaking.

Finally, by using audiospatial representations, findings from Experiment 1 suggest that gestures can arise from simulations of spatial information other than visuospatial. This result has been followed up by a second study that aimed to assess whether the gestural behaviour observed in Experiment 1 (e.g., higher pith/higher gesture) was determined by linguistic processes. Results showed the spatial representation of pitch in gesture was unaffected by the speech production. The fact that gesture results from audiospatial representations has one important implication for the GSA with respect to the role that the motor component of the mental simulation has in gesture production. It may be the case that a) it is not necessary that the motor component is part of the perceptual simulation to start gesture production or that b) there is a link between audiospatial representations and the activation of action simulations. Future research needs to investigate what mechanisms lead to the production of gestures in response to perceptual simulations that are not informed by action. This means that it is necessary to investigate the link between spatial simulations – which may be represented in any modality – and gesture production.

One avenue for future research is to further investigate the link between supramodal spatial representations and actions. A considerable amount of literature has put forth the intriguing, and seemly likely, hypothesis that there are supra-modal, i.e. nonmodality specific, spatial representations (Farah, Wong, Monheit, & Morrow, 1989). If this is the case, then there may be audiospatial or olfactory-spatial representations that lead to motor activation and therefore to the production of gestures. The GSA claims that thinking of the smell of a cake would not lead to gesture production, and the assumption underlying this claim is that representing the smell does not involve movement (e.g., the nose does not move) and is not tied to action. However, leaving aside the fact that thinking about the smell of a cake may indeed lead to motor activation, for example when representing the act of chewing, there is another important point to make: also olfactory representations may have a spatial component. For example one may represent where a smell comes from. If a spatial representation of olfactory experiences exists, then as it is for the case of auditory pitch, that may be the spatial representation leading to gesture. It is suggested that future research moves towards this possibility by investigating supra-modal spatial representations.

Chapter 3

From Sound to Action: When Does the Representation of Sounds Affects the Production of Co-speech Gestures?

3.1. Introduction

The previous chapter addressed the question of *whether* audiospatial representations would lead to gesture production. The present chapter investigates *how* the simulation underlying gesture depicting sounds affect gesture production. The GSA framework claims that the following factors account for differences in gesture rate. The first factors is whether the speaker deems gesturing appropriate to a certain social context. The second factor is the strength of the connections between pre-motor and motor cortical areas. The third factor is the concurrent engagement of the motor system for speaking. The last factor is the strength of action simulation. Crucially, the GSA assumes that, all other things being equal, the latter factor (strength of action simulation) determines differences in the gesture rate: the stronger the action activation, the more likely it will surpass the *gesture threshold* and lead the production of gestures (Hostetter & Alibali, 2008). It is central to the GSA the claim that action simulation determines the production of gestures, and thus the stronger the representation of actions, the higher the likelihood that a gesture will be produced: "action characteristics such as how an object moves or how it is physically manipulated should be more likely to lead to gesture than spatial properties like size or shape because the simulations that underlie motor imagery are particularly likely to strongly activate the motor system" (Hostetter, 2014, p. 1469). Therefore, the GSA assumes that the simulation of physical movement – called simulation of actions – activates pre-motor areas and, if it spreads to motor areas it will result in the production of gestures. The stronger the simulation of actions, the more likely

activation will spread from pre-motor to motor areas and result in the production of gestures.

Following this claim, the GSA organizes perceptual and motor simulations into a hierarchy ranging from those representations that are more strongly tied to actions, and thus are more likely to elicit gestures, to those less strongly tied to actions. For example, explaining 'how to land a triple axel', should lead to higher gesture rate than talking about objects visually perceived, such as 'cake' because "while images may lead to action simulations, these simulations may be less highly activated than those that occur when speakers are thinking more directly about actions, as is the case with motor imagery" (Hostetter & Alibali, 2010, p. 247). In turn, speaking about the shape of a cake should lead to fewer gestures than talking about its colour. This is because seeing a cake involves simulating its affordances (e.g., 'to be held') and its spatial properties (e.g., shape), which involve activation of motor areas, (e.g., eye movement).

Conversely, the representation of perceptual simulations less tied to actions (e.g., 'yellowness') does not readily result in action simulation. Similarly, simulating olfactory experiences, such as the smell of a cake, does not involve simulated actions. The GSA assumptions received some empirical support. With respect to the effect of visual representations on gesture production, it seems that these lead to a higher gesture production than other perceptual representations, such as auditory or olfactory. For example, describing graphic designs resulted into more gesticulation compared to describing synthesized sounds or tea tastes (Krauss, Dushay, Chen, & Rauscher, 1995). Further, participants gestured more when describing more *imageable* objects, i.e., those that are easy to imagine in their absence, such as a

sculpture, compared to when they discussed less imageable topics, such as economy (Feyereisen & Havard, 1999).

A study designed to investigate the relationship between representing actions and producing gestures presented participants with one of two conditions. In one condition participants were presented with configurations of dots and asked to observe them only before describing them. In a different condition participants had to observed and physically reproduce the configurations of dots before describing them. Participants gestured more when describing configurations that they had manually reproduced, rather than only observed (Hostetter & Alibali, 2010). These results indicate that representing actions leads to a higher gesture rate than representing spatial properties. Explaining 'how to do something', such as how to wrap a box, leads to a higher gesture rate than talking about visual and abstract targets (Feyereisen & Havard, 1999). Taken together, the results of these studies seem to confirm the GSA assumption that differences in the gesture rate result from differences in the strength of action activation.

However, previous literature has also identified other factors that may affect gesture production. For example, a speaker's familiarity with the stimulus that s/he is describing, the ease with which the stimulus can be described and the speaker's spatial abilities, all affect gesture production (Bavelas et al., 2002; Hostetter, Alibali, & Kita, 2007; Rauscher, Krauss, & Chen, 1996). When asked to describe unfamiliar stimuli, such as a 19th century dress, participants produced more non-redundant gestures, i.e, those depicting information that is not conveyed in the accompanying speech) (Bavelas et al., 2002). The linguistic content also seems to determine differences in the gesture rate, with more gestures produced when using words that are more difficult to retrieve (Beattie & Shovelton, 2000).

Gestures more often appear concurrently to spatial than non-spatial words (Morsella & Krauss, 2004). This result should not surprises given that gestures express and reflect spatial thinking. However, because gesture reflects spatial thinking, the speakers' spatial abilities might influence the gesture rate. To test this possibility, a study asked participants to solve a mental rotation task and then describe how they arrived at their solutions. Participants who had lower spatial abilities produced more static gestures than those with higher spatial abilities (Göksun, Goldin-Meadow, Newcombe, & Shipley, 2013). Therefore, it seems that greater spatial abilities are associated with reduced gesture production.

Summing up previous results, there is a body of research showing that the following factors predict variations in gesture rate: a) the speaker's *familiarity* with the stimulus, b) the stimulus verbal *encodability*, or *describability* c) how easy is to imagine the stimulus in its absence, or *imageability* d) the speaker's spatial abilities. Crucially, the GSA argues that, all things being equal, the action component of the mental representation is the main predictor for variations in the gesture rate. That is, the *strength of action* simulation will affect gesture rate. Gesture arise more readily from simulations of actions and of visuospatial perceptions that are tied to actions. The present study was designed to study the power of these predictors on gesture rate. The study asked participants to describe a set of sounds. Sounds varied in relation to the degree with which their mental representation was linked to action, allowing to test whether there is a difference between the effect of action strength and visual strength on gesture rate. For example, when hearing the sound of someone knocking at the door, visual and motor information involved in seeing and knocking on the door is represented, together with auditory information about the sound of 'knocking'. Conversely, when hearing white noise, auditory information is strongly represented

but action information is not. The advantage of using sounds here is twofold: first, neither visual and action representations are explicitly conveyed in the stimulus (as they would be if one used pictures or videos), thus reducing the possibility of participants imitating the actions involved in producing a certain sound or describing other elements of the image. Second and most importantly, by using sounds it is easy to manipulate the extent of action strength involved in the mental simulation. For example, simulating 'bouncing a basketball' involves high action simulation and high visual simulation; simulating the sound of 'stream' involves high visual simulation but low action simulation; simulating a phase-scrambled version of a sound involves low action simulation and low visual simulation.

3.2. Experiment 3

3.2.1. Participants

Twenty pairs of participants took part in this study. All pairs of participants were same gender (12 female pairs) and participants were not previously acquainted. The average age was 23. All participants were right-handed, native speakers of English and none had any language, hearing or mobility impairments. All participants were students at the University of Manchester who took part to the study in exchange for course credits or £5. They were naive as to the purpose of the experiment, which was advertised as an investigation of how people communicate about sounds to an addressee.

3.2.2. Materials

In order to select the set of sounds to be used for this study, we asked participants to rate individua

Pilot study 1. The aim of this first pilot study was twofold. First, the study sought to establish which sounds from a set of 44 were identifiable. Identifiability of a sound refers to whether the sound can be recognized and discriminated from similar sounds. Identifying a sound involves also being able to tell what may cause the sound. For example, if participants heard the sound produced by someone scratching his/her skin or head, the sound was considered as 'correctly identified' when the action ('scratching') and the agent and patient of the action ('person') were identified as the sound cause. Correct identification of the sound meant that the participants had a representation of the context in which the sound was generated. By contrast, if someone mistake the same sound ('scratching') as the sound produced when someone brushes the floor, the answer would be considered incorrect, and the sound non-identified, because it meant that the participant was not able to discriminate between the two sounds.

Identifiability was chosen as necessary condition for a sound to be included because it bears on the representation of the sound and thus on gestural behaviour. The consequence of misrepresenting a sound is that one may have a completely different mental representation of the context in which the sound is produced, giving rise to gestures that differ with respect to what they represent. This would make it impossible to compare two gestures, for example the gesture representing someone brushing the floor and the gesture representing someone scratching their head, both used to describe the sounds of 'scratching'. Therefore, incorrectly identified sounds were not included in the study.

The second aim of this pilot study was to collect ratings for familiarity, imageability, visual strength, action strength and describability for each sound because these could be possible predictors of the gesture rate. In order to collect these

ratings, 20 participants, all native speakers of English from the University of Manchester, were presented with 44 sounds. The order of presentation of the sounds was randomized while the order of the ratings was fixed. Sounds were presented through inner-ear headphones on a Mac Book using Superlab. First, each participant was presented with one sound. The sound automatically played twice and after the second play participants were asked to identify it by naming what was making the sound. They needed to type their answers into an empty box that appeared at the centre of the screen. Following this, instructions to rate the sounds for familiarity, imageability, describability, visual strength and action strength appeared in order one at the time. *Familiarity and imageability* (how difficult/easy it is to bring a mental image of the sound to mind) of each sound were rated using an electronic 7-point Likert scales ranging from 1 (Unfamiliar/Very familiar) to 7 (Very difficult/Very easy). To measure *visual strength* and *action strength* of the sounds, participants were asked to rate the extent to which they experienced the concept of a sound by *seeing* (visual strength) or *by performing an action* (action strength), using a five-point Likert scale (1 = Not at all to 5 = Greatly).

The rating instructions for visual strength and perceptual strength were adapted from Lynott and Connell's ratings of perceptual strength in five modalities (2013). Finally, participants were asked to rate how easy they would find it to *describe the sound in words* to another person, using a scale that ranged from 1 (Very Difficult) to 6 (Very easy).

For the gesture production study, only stimuli that were correctly identified by 80%-100% of the participants were selected. The sounds included in the study were classified into three main groups of sounds, depending on what caused them. The first group of sounds included those that are produced by the human body (e.g., ‘hands

clapping' or 'finger clicking'); the second included those produced by the human body interacting with an object (e.g., 'typing on computer keyboard'); the third included those not produced by the human body (e.g., 'wind' or 'rain'). This allowed two main categories of sounds to be selected for the gesture production study 1) those whose representation involves a high degree of visual activation and action activation, such as the sound o 'typing on computer keyboard sound' and, 2) those sounds whose representation involves a high degree of visual activation, but a low degree of action activation, such as the sound of a running stream.

Pilot study 2. The aim of Pilot Study 2 was similar to that of Pilot Study 1, except that it concerned unidentifiable sounds. The purpose of this study was to have a selection of sounds that were new to the hearer, and to collect the ratings for each sound. The experimental design required the use of a third category of sounds whose representation involved a low degree of visual activation and a low degree of action activation. Unidentifiable sounds were obtained by editing the sounds used in Pilot Study 1. To create phase-scrambled versions of the sounds used in study 1, speed and tempo were modified. Twenty different participants, all native speakers of English were presented with the new set of 30 sounds and rated them as in Pilot Study 1. Only sounds for which between 0% and 40% of participants identified the original sound were included in the gesture production study. The range of values was chosen to give the regression a continuous range of values.

From Pilot study 1 and 2 were selected 24 sounds, identified by 80% of the participants, and 6 sounds unidentifiable. For three of the six unidentifiable sounds between the 30% and 40% of participants tried to guess, and for the remaining three

between the 10-20% of participants tried to guess. The full list of sounds is available in Appendix A.

3.2.3. Apparatus

The speaker was recorded using a Canon MD160 camcorder mounted on a tall tripod located in front of the participant. The camera captured the describer's body from above their head to their knees and part of the table in front of them. The camera was connected through cable to a TV that was located in the room where the addressee was asked to stay. Sounds were presented to describers through inner-ear headphones plugged into a 13-inch Mac Book, model A1342. The presentation order was randomized by using Superlab 4.0 Software for Mac. All sounds were edited using

Audacity Software for Mac.

3.2.4. Procedure

The pairs of participants were invited to enter the room and take a seat in one of two chairs across a low coffee table. They were introduced to each other and asked to read the participant information sheet. Then, they were asked to read the consent form; by signing the consent form, all participants consented to be video-recorded. The participant information sheet provided them with information about the aim and the structure of the experiment. It was explained that that the aim of the study was to investigate how people communicate about sounds to an addressee when the addressee cannot hear the sounds, and how people understand this information. They were also explained that they would be either describers of the sounds or addressees for the descriptions.

After the participants had signed the consent form, I ran them through the instructions and explained them that the study consisted of two parts. In the first part of the study, the participants were randomly assigned to the role of describer or to the role of addressee by flipping a coin. The describer's task was to describe 30 short sounds, one at the time, to the addressee. The addressee's task was to listen carefully to all the descriptions, because later they would be asked questions about the content of the descriptions. During this first part of the experiment the participants were in two different rooms. The describer described the sounds to the addressee in front of a tripod camera in the experimental room. The tripod camera was connected on a TV located in the observation room, where the addressee would watch the describer.

Each trial started when the describer pressed the Spacebar. The sound automatically played twice and then a message appeared on screen prompting the participant to describe the sound. The describer was encouraged to take as much time as they needed to describe the sound and to move to the next sound once they had completed the description. To make the participants feel as comfortable as possible with the experiment, I showed both the describer and the addressee the rooms in which they would be. Also, in order to familiarise with the task, with the camera and the equipment, the pairs of participants had two practice trials, before the experiment started.

In the second part of the study the describers completed a questionnaire in which they were asked to rate overall the set of sounds that they had described by providing a single rating for the 30 sounds together. Ratings were the same as those used in Pilot studies 1 and 2. Addressees were asked to complete the questionnaire and also to answer questions about the descriptions. Finally, the participants solved a

pen and paper mental rotation task and a computerized Corsi-block task to test their spatial skills.

3.2.5. Coding Speech

Once data were collected, the speech of each participant for each sound was transcribed. The speech transcriptions included: 1) false starts, 2) filled pauses such as 'hem', 'hum', 3) repetitions, 4) onomatopoeic sounds, such as the 'zipping up' sound and, 5) meta-descriptive statements, such as "I am not sure how to describe that". However, false starts, filled pauses and meta-descriptive statements were *not* included in the analysis, nor were the accompanying gestures. False starts and filled pauses were excluded because they are words uttered during the verbal descriptions, but nevertheless they are not relevant in conveying semantic content about what is being described, instead, they serve a more interactive or cognitive function. Likewise, metadescriptive statements do not convey information about the sound being described. Rather, they only say something about the description itself, for example whether it is easy or hard to describe a sound, but not about the sound itself. Because the aim of the study was to see how gestures represent sounds, such statements and the accompanying gestures were not of use for the purpose and therefore were excluded. The descriptions for each individual sound were transcribed separately. A typical example of someone describing a sound is: 'it was someone typing on a computer keyboard' whilst hands move mimicking a typing movement.

3.2.6. Coding gestures

The next step of the coding procedure was to identify the gestures that occurred during the sound descriptions, by distinguishing *gestural* from *non-gestural movements*. As with the studies reported in Chapter 2, gestures were classified into

nonrepresentational gestures, e.g., beat-like movements that play a pragmatic function and *representational gestures*, e.g., gestures that convey semantic meaning (Alibali et al., 2001). Representational gestures were further classified into *iconic* and *metaphoric* (McNeill, 1992). Those hand and body movements that depicted concrete objects, such as a 'ball', and actions, as 'bouncing' were classified as iconic. Iconic gestures that depicted sounds mostly represented the event that gave rise to the sound, the agent and patient, for example 'person' hitting 'ball'. Metaphoric gestures, which depict abstract entities, represented qualitative properties of the sound, such as its timbre. An example of a metaphoric gesture is the participant saying 'it was a deep sound' while producing a downwards hand movement.

3.2.7. Reliability

To establish reliability of the coding, a second observer who was blind to the purpose of the experiment coded the 25% of the data (653 gestures). She was asked to distinguish between gestures and non-gestural body movements and to divide gestures into representational and non-representational gestures. Calculations for the inter-rateragreement for gesture identification was k = .89 (agreement on 95.1% of the observations) and for gesture type it was k = .90 (agreement on 97.1% of the observations). According to Landis and Koch (1977), these results indicate an excellent agreement.

3.2.8. Design and analysis

The aim of the study was to investigate a) which variables best account for variations in the gesture rate when people describe sounds and b) whether action strength and visual strength are the best predictors. A within-subjects design was used with all the participants presented with the same sound stimuli. There were eight

independent variables that have been shown to affect gesture rate in previous studies: 1) how identifiable, 2) familiar, 3) imageable, and 4) describable the sound was to the participants; the level of 5) visual and 6) action experience the cause sound; and the participant's 7) spatial memory and 8) spatial abilities. Spatial memory skills were measured using a computer-based version of the Corse-Block task that yields a measure of the short-term memory. Spatial ability was measured by using a paper-based version of a Mental Rotation task (Shepard & Metzler, 1971) that measures the ability to imagine how a three-dimensional object would look like from a different perspective. The analysis was concerned with which variables best predicted the observed changes in the rate of gestures produced per stimulus. Gesture rate was calculated as number of gestures per 100 words and calculated separately per each sound item.

3.3. Results

First, I calculated the correlation between the participants' scores in the Corsi-block and mental rotation tasks with the gesture rate. There was no correlation between the participants' gesture rate and their performance in the Mental Rotation Task (r = -.131, n = 20, p = .583) or their Corsi-Block score (r = -.047, n = 20, p = .845).

Second, in order to test the predictive power of vision and action strength on gesture rate, I ran a multiple regression analysis that included Familiarity, Imageability and Describability ratings as predictors and the gesture rate as dependent variable (correlations are reported in Table 6). The analysis resulted statistically significant, $F(3,26) = 7.05$, p = .001, indicating that the three predictors together predict well the gesture rate and the model accounted for the 44.9% of the variance in the gesture rate (R^2 = .449). Further, an analysis of the coefficients for each of the

three predictors revealed that nor Familiarity (β = -.22, p = .794), Imageability (β = 79, p = .185), or Describability (β = .09, p = .891) alone significantly predicted gesture rate.

As the analysis aimed at testing the predicting power of action and visual strength on gesture rate, it included a third step where standardized residuals from the preliminary regression were further analysed to test whether action strength and vision strength had independent predicting power. A second regression analysis used Action Strength and vision Strength as predictors. The analysis did not result statistically significant, $F(1,27) = .116$, p =.891, R^2 = .017 and neither of the two predictors was statistically significant as shown in Table 5. It is to be noticed that the predictors were highly correlated with one another, meaning that they were not independent on each other. The high correlation may be due to the fact that how familiar, imageable or easy to describe is an item may also determined by the person's physical experience with it. This means that action strength and vision strength on their own could predict variation in the gesture rate, as visual and actional experience with a sound may determine, for example, how imageable the item is.

Table 5. Values for the two predictors Visual Strength and Action Strength

Predictor	β	p
Visual Strenght	.075	.712
Action Strength	-.134	.511

Table 6.*Correlation coefficients for Identifiability, Familiarity, Imageability, Visual Strength, Action Strength and Describability*.

Predictor	1	2	3	4	5	6
1.Identifiability						
2. Familiarity	0.932**					
3.Imageability	0.905**	0.970**				
4.Visual Strength	0.826**	0.910**	0.936**			
5.Action Strength	0.620**	0.604**	0.572**	0.319		
6.Describability	0.941**	0.979**	0.953**	0.887**	0.605**	

Note. **significant at $p < .01$

Given the lack of power of the two main predictors, i.e., action strength and visual strength, a follow-up analysis investigated whether the two predictors would have any predicting power on gesture rate. The new regression analysis only included Visual Strength and Action Strength as predictors. The prediction model with the two predictors Perceptual Strength and Action Strength, was statistically significant, $F(2,27) = 8.398$, $p = .001$) and accounted for the 34% (R^2 =.384, Adjusted R^2 = .338) of the variance in the gesture rate. Variance in the gesture rate was most accurately predicted by Action Strength ($\beta = 0.46$) than by Vision Strength ($\beta = 0.29$), (see Table 7).

Table 7. *Significant values for the two predictors Visual Strength and Action Strength*

Predictor	β	p
Visual Strength	.292	$p = .039$
Action Strength	.461	$p = .003$

3.4. Discussion

The current study shows that descriptions of a sound are likely to be accompanied by greater g encodability and imageability of a target all predict variations in the gesture rate when describing the target (e.g., Bavelas et al., 2002; Hostetter, Alibali, & Kita, 2007; Rauscher, Krauss, & Chen, 1996). However, unlike previous studies, current results have shown that none of these three predictors can independently predict gesture rate. One of the reasons for this could be that it is difficult to disentangle the effect of each of these predictors. Let us take as an example a stimulus which is familiar, such as the sound of a door banging. A speaker who is familiar with the sound, i.e., has heard a door banging many times, will find it easier to imagine it in its absence.

On the other hand, unfamiliar stimuli, such as unidentifiable sounds, might lead to lower gesture production. An explanation for this result can be that gestures emerge from a combination of perceptual and motor experiences: for example a 'typing' gesture arises from representing the actual experience of typing. In the case of unidentifiable sounds, one has no previous experience with them and therefore their mental representation is excluded from the speakers' representations repertoire. If the

hypothesis that lack of physical experience with a stimulus affects gesture rate because representational support is not readily provided, then the effect of other predictors, such as low familiarity, can ultimately be related to the lack of individual's physical experience.

This hypothesis has been investigated in this study, mainly designed to test the hypothesis that the action component of the mental representation is the best predictor for variations in the gesture rate. As previous studies have shown, the higher the activation strength of the action and visual components of the simulation, the higher the gesture rate (Feyereisen & Havard, 1999; Krauss et al., 1995). More specifically, the strength of the action component of the mental representation is the best predictor for variation in the gesture rate (Hostetter & Alibali, 2010). Results from the second analysis for Experiment 3 seem to suggest that simulations of actions and simulations of visual perceptions both affect the rate at which one gestures. However, the experimental design was not powerful enough to confirm this hypothesis or to disentangle the effect of the action component of the simulation from the visual component of the simulation on the gesture rate.

Experiment 3 and Experiment 1 suggest that further research is needed in particular to test whether the spatial component of the mental simulation has an effect on the gesture rate independently on the action component. Answering this question would also allow to establish whether the GSA model makes better prediction than other models of gesture production, such as the GPT, the IPH, the LA model. Whereas these models claim that action representations and spatial representations are equally likely to give rise to gestures, according to the GSA, the action component is the best predictor: gestures are produced when actions are simulated and simulated

actions occur when one simulates performing an action or when one simulates perceptual representations that are linked to motor representations (Hostetter & Alibali, 2008). Looking more closely into this issue would also allow to test the idea that motor representations can be driven by perceptual content (Prinz, 1997; Schütz-Bosbach & Prinz, 2007). The idea has found some empirical evidence in gesture rate research. For example, it has been found that that iconic gestures are produced when processing words whose content is not related to actions (Bach, Griffiths, Weigelt, & Tipper, 2010). This conclusion suggests that different mechanisms from the one described within the GSA should be explored that can give rise to gesture production and more details are needed when describing the mechanisms involved that lead from spatial representations (not necessarily visuospatial) to the production of actions.

Chapter 4

Conclusions

This thesis focused on gestures depicting information about sounds. In particular, it investigated how gestures represent auditory pitch (Chapter 2) and which aspects of the mental representation underpinning gestures affect gesture rate (Chapter 3). Previous research on gesture production has largely neglected gestures depicting sounds. By focusing on the gestural depiction of sounds, the present thesis has addressed three issues specific to gesture production: 1) whether audiospatial representations of sound pitch elicit gesture production 2) whether the concurrent verbal encoding determine the gestural depiction 3) which cognitive factors predict differences in the rate at which one gestures about sounds. Findings from the three studies presented hopefully contributed to our understanding of gesture production.

First, within the field of gesture and embodied cognition, gesture production has been linked to the simulation of the action component of the mental representation as outlined by the GSA framework (Hostetter & Alibali, 2008). According to the GSA, the action content of the mental representation elicits gesture production. Crucially, the perceptual component of the mental representation elicits gesture production only when its activation is tied to motor simulations. According to the GSA, the conditions under which simulated perceptions evoke simulated actions did not include audiospatial perceptions. Chapter 2 addressed the question of whether audiospatial representations trigger gesture production by asking participants to describe melodies made of different pitched tones. Pitch was chosen because its representation involves an auditory and a spatial component (e.g., Connell et al., 2013). Results showed that participants produced gestures to convey information

about pitch shifts in such a way that higher pitches were allocated to higher spatial locations compared to lower pitches. This result suggests that audiospatial representations lead to producing gestures that express spatial information about pitch. This result fundamentally undermines the assumption made by the GSA that gesture production requires the simulation of actions through visuospatial representations. As a consequence, the GSA should be expanded in order to include mechanisms that account for the production of gestures triggered by audiospatial representations, for example by taking into account the hypothesis of a supra-modal spatial representation system (Farah et al., 1989). Earlier models of gesture production, such as the Growth Point Theory (McNeill, 1992) or the Interface Model (Kita & Özyürek, 2003) claimed that the spatial component of the concept would lead to gesture production; these model offer better predictions than the GSA because they do not commit to a representational format and therefore they do not exclude that audiospatial representation may also lead to gesture production.

The second study (Chapter 2) addressed the question of whether the spatial representations giving rise to gesture depicting different pitched notes were elicited by the linguistic encoding, for example the use of words such as 'high/low'. The hypothesis that the verbal encoding might affect gesture production echoes an issue that is central to models of gesture and speech production, i.e., the relationship between gesture and speech. In order to address this possibility, a second experiment was designed where participants were prevented from speaking, thus inhibiting speech planning and articulation. Because gestures represented pitch with the same patterns, e.g., high pitches were represented as the result of an upwards hand movement, it can be concluded that the linguistic encoding was not responsible for the gestural encoding. This means that gesture production is an interconnected but

autonomous process from speech, which challenges the idea that what is represented in gestures is influenced by what is expressed in speech (e.g., Kita & Özyürek, 2003; McNeill, 1992) and the idea that the gesture production depends upon speech production (e.g., Krauss et al., 1995). Finally, it has been assumed that the action component of the mental simulation has a greater effect on gesture rate than the spatial component (Hostetter & Alibali, 2008). If it is true that the action component of the mental representation is responsible for greater gesture production, than gesturing should become more frequent when communicating about actions. Chapter 3 tested this assumption by analysing the rate of gestures produced while describing sounds whose representation has a stronger or weaker action component. Results from the study suggested that both the strength of the action component and visual component together affect gesture rate. However, the study did not provide definite evidence to the assumption that action simulations better explain variation in gesture production, such that the higher the action strength of the simulation the more gesticulation. Further studies will be needed that investigate the role of action and spatial representations independently on each other in gesture production. Findings from Experiment 1 and Experiment 3 suggest that research should investigate more thoroughly the role of spatial representations in gesture production. In conclusion, the three studies presented here have contributed to exploring the mechanisms underlying the production of gestures. Taken together they suggest that gestures can be produced as the result the simulation of a) the spatial component of the mental representation of auditory pitch, b) independently on the linguistic encoding and c) that action and visual component of the mental representation might be the best predictors for variations in the gesture rate.

Appendix A

Applause, kiss, footsteps, snorting, burping, breathing heavily, punch, ball bouncing, dialling on touch screen telephone, door shutting, drumming, pouring drink, teeth brushing, typing on computer keyboard, turn page, tearing paper, zipper, thunder, tree falling to ground, howling wind, heavy rain, stream, ocean waves, water dripping.

References

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language, 44*, 169-188.

Arnott, S. R., & Alain, C. (2011). The auditory dorsal pathway: Orienting vision. *Neuroscience & Biobehavioral Reviews, 35*, 2162-2173.

Bach, P., Griffiths, D., Weigelt, M., & Tipper, S. P. (2010). Gesturing meaning: Non-action words activate the motor system. *Frontiers in Human Neuroscience, 4*, 214.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences, 22*, 637-660.

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science, 2*, 716-724.

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes, 15*, 469-489.

Bavelas, J., Kenwood, C., Johnson, T., & Phillips, B. (2002). An experimental study of when and how speakers use gestures to communicate. *Gesture, 2*, 1-17.

Beattie, G., & Shovelton, H. (2000). Iconic hand gestures and the predictability of words in context in spontaneous speech. *British Journal of Psychology, 91*, 473-491.

Bryant, D. J. (1992). A spatial representation system in humans. *Psycholoquy, 3*, space 1.

Butterworth, B., & Hadar, U. (1989). Gesture, speech, and computational stages: A reply to

McNeill. *Psychological Review, 96*, 168-74.

Connell, L., Cai, Z. G., & Holler, J. (2013). Do you see what I'm singing? Movement biases pitch perception. *Brain and Cognition, 81*, 124-130.

Connell, L., & Lynott, D. (2010). Look but don't touch: Tactile disadvantage in processing modality-specific words. *Cognition, 115*, 1-9.

Degerman, A., Rinne, T., Särkkä, A., Salmi, J., & Alho, K. (2008). Selective attention to sound location or pitch studied with event_related brain potentials and magnetic fields. *European Journal of Neuroscience, 27*, 3329-3341.

Ekman, P., & Friesen, W. V. (1972). Hand movements. *Journal of Communication, 22*, 353-374.

Ellis, R., & Tucker, M. (2000). Micro_affordance: The potentiation of components of action by seen objects. *British Journal of Psychology, 91*, 451-471.

Farah, M. J., Wong, A. B., Monheit, M. A., & Morrow, L. A. (1989). Parietal lobe mechanisms of spatial attention: Modality-specific or supramodal? *Neuropsychologia, 27*, 461-470.

Feyereisen, P., & Havard, I. (1999). Mental imagery and production of hand gestures while speaking in younger and older adults. *Journal of Nonverbal Behavior, 23*, 153-171.

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *The Quarterly Journal of Experimental Psychology, 61*(6), 825-850.

Giudice, N. A., Betty, M. R., & Loomis, J. M. (2011). Functional equivalence of spatial images from touch and vision: Evidence from spatial updating in blind and sighted individuals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 621.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review, 9*, 558-565.

Göksun, T., Goldin-Meadow, S., Newcombe, N., & Shipley, T. (2013). Individual differences in mental rotation: What does gesture tell us? *Cognitive Processing, 14*, 153-162.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science, 12*, 516-522.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron, 41*, 301-307.

Hill, E. L. (2001). Non-specific nature of specific language impairment: A review of the literature with regard to concomitant motor impairments. *International Journal of Language & Communication Disorders, 36*, 149-171.

Hostetter, A. B. (2014). Action Attenuates the Effect of Visibility on Gesture Rates. *Cognitive Science, 38*, 1468-1481.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as Simulated Action. *Psychonomic Bulletin & Review, 15*, 495-514.

Hostetter, A. B., & Alibali, M. W. (2010). Language, gesture, action! A test of the gesture as simulated action framework. *Journal of Memory and Language, 63,* 245-257.

Hostetter, A. B., Alibali, M. W., & Kita, S. (2007). I see it in my hands' eye: Representational gestures reflect conceptual demands. *Language and Cognitive Processes, 22,* 313-336.

Iverson, J. M., Capirci, O., & Caselli, M. C. (1994). From communication to language in two modalities. *Cognitive Development, 9,* 23-43.

Iverson, J. M., & Thelen, E. (1999). Hand, mouth and brain. The dynamic emergence of speech and gesture. *Journal of Consciousness Studies, 6,* 19-40.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science, 16,* 367-371.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kimura, D., & Archibald, Y. (1974). Motor functions of the left hemisphere. *Brain, 97,* 337-350.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), Language and gesture (pp. 162-185). Cambridge: Cambridge University Press.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic 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.

Kosslyn, S. M., Thompson, W. L., & Alpert, N. M. (1997). Neural systems shared by visual imagery and visual perception: A positron emission tomography study. *NeuroImage, 6*, 320-334.

Krauss, R. M., Dushay, R. A., Chen, Y., & Rauscher, F. (1995). The communicative value of conversational hand gesture. *Journal of Experimental Social Psychology, 31*, 533-552.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Lidji, P., Kolinsky, R., Lochy, A., & Morais, J. (2007). Spatial associations for musical stimuli: A piano in the head? *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 1189.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods, 45*, 516-526.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review, 92*, 350.

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

McNeill, D. (2000). *Language and gesture.* Cambridge University Press.

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.

Morsella, E., & Krauss, R. M. (2004). The role of gestures in spatial working memory and speech. *The American Journal of Psychology,117*, 411-424.

Nobe, S. (2000). Where do most spontaneous representational gestures actually occur with respect to speech. In *Language and Gesture,* David McNeill, ed. Cambridge: Cambridge University Press. pp. 186-198.

Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology, 13*, 278.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9*, 129-154.

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science, 7*, 226-231.

Renier, L. A., Anurova, I., De Volder, A. G., Carlson, S., VanMeter, J., & Rauschecker, J. P. (2009). Multisensory integration of sounds and vibrotactile stimuli in processing streams for "what" and "where". *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*, 10950-10960.

Rusconi, E., Kwan, B., Giordano, B. L., Umilta, C., & Butterworth, B. (2006). Spatial representation of pitch height: The SMARC effect. *Cognition, 99*, 113-129.

Schütz-Bosbach, S., & Prinz, W. (2007). Perceptual resonance: Action-induced modulation of perception. *Trends in Cognitive Sciences*, *11*, 349-355.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701-703.

Taylor, L. J., & Zwaan, R. A. (2008). Motor resonance and linguistic focus. *The Quarterly Journal of Experimental Psychology, 61*, 896-904.

Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of Learning and Motivation, 44*, 35-62.