PERCEPTION OF EMOTION IN SOCIAL INTERACTIONS FROM BODY MOVEMENT AND VOICE

LUKASZ PIWEK M.A., M.SC.

A thesis submitted for the degree of
Doctor of Philosophy (*Ph.D.*)

School of Psychology
College of Science and Engineering
University of Glasgow

December 2013

This thesis is dedicated to the loving memory of Krzysztof Papierkowski. Words simply cannot describe his passion, energy and inspiration which he passed on to many generations of people, including myself. His legacy will always be remembered.

Rest in Peace Krzysztof (1944 – 2013).

ACKNOWLEDGMENTS

It has been an intensive four years of my life and this thesis would not have been possible without the many people who have supported me both academically and personally. I would like to sincerely thank my supervisor Prof. Frank Pollick for his continuous support and help over the many years we have worked together since my undergraduate studies. I thank him for his unending patience, guidance, knowledge and, above all, the incredible opportunities without which I wouldn't have been able to reach my goals and dreams. My wholehearted gratitude goes to Dr Karin Petrini for her incredible energy and immense support in shaping raw ideas into concrete solutions. Karin taught and inspired me whenever I became stuck and she always had time to help me with any aspect of analysis or intellectual grinding of problems. Karin gave me the best possible perspective of what it is to be a researcher. I would also like to express gratitude to Dr David Simmons for his continuous support at every stage of my academic journey, for always believing in me, and for the amazing opportunities I received thanks to him. Dr Phil McAleer, Dr Scott Love, Dr Oliver Garrod, Dr Lawrie McKay and Dr Martin Lages are all thanked for their epic research chats, useful guides and inspiration. I thank Emanuele De Luca and Caterine Arrabal for being there for me in times when I needed them most, as well as for the best combinations of Italian food and philosophical discussions. Love goes to my best friends, Slawomir Pol, Pawel Mroz, Rafal Gosieniecki and Wojciech Glowacki; you all inspired not only this thesis but the direction of my life itself. I would also like to thank The Economic and Social Research Council for providing financial support for my research.

Love and thanks to both of my wonderful moms, Barbara Piwek and Ania Papierkowska.

Finally, I would never have got to the end of this journey without the constant love and support of my partner Ola Papierkowska. To you, I have nothing but unending love and thanks.

DECLARATION

I declare that this thesis is my own work carried out under the normal terms of supervision.

Lukasz Piwek

ABSTRACT

The central theme of this thesis was to examine different aspects related to the observation and judgement of emotions from the body movement and voice of two actors engaged in social interaction. There were four major goals related to this theme. The first goal was to create a novel stimulus set for the study of emotional social interactions. The second was to validate the created stimulus set by examining emotion perception in ways similar to that done with single actor displays. The third goal was to examine the effect of degrading visual and auditory information on the perception of emotional social interactions. The final goal was focused on the multimodal integration of emotional signals from body movement and voice. Initially, a stimulus set was created that incorporated body movement and dialogue between two actors in brief, natural interactions that were happy, angry or neutral at different levels of intensity. The stimulus set was captured using a Vicon motion and voice capture system and included a group of nine professional and non-professional actors. This resulted in a corpus of 756 dyadic, multimodal, emotional interactions. A series of experiments were conducted presenting participants with visual point-light displays, auditory voice dialogues or combinations of both visual and auditory displays. Observers could accurately identify happy and angry interactions from dyadic displays and voice. The intensity of expressions influenced the accuracy of the emotional identification but only for angry rather than happy displays. After validation of the stimulus set, a subset was selected for further studies. Various methods of auditory and visual distortion were tested separately for each modality to examine the effect of those distortions on recognition of emotions from body movement and voice. Results for dyadic point-light displays followed similar findings from single actor displays that inversion and scrambling decreased the overall accuracy of emotion judgements. An effect of viewpoint was also found, indicating that observation of interaction from a side viewpoint was easier for emotion detection than observation of interaction from an oblique viewpoint. In the case of voice, methods of brown noise and low-pass filtering were shown to degrade emotion identification. However, with both visual and auditory methods of distortion, participants were still able to identify emotions above the level of chance, suggesting high sensitivity to emotional cues in a social context. In the final set of studies, the stimulus set was used in a

multimodal context to examine the perception of emotion from movement and voice in dyadic social interactions. It was repeatedly found that voice dominated body movement as a cue to emotions when observing social interactions. Participants were less accurate and slower in emotion discrimination when they were making judgements from body movement only, compared to conditions when movement was combined with dialogue or when dialogue was presented on its own. Even when participants watched emotionally mismatched displays with combined movement and voice, they predominantly oriented their responses towards the voice rather than movement. This auditory dominance persisted even when the reliability of the auditory signal was degraded with brown noise or low-pass filtering, although visual information had some effect on judgements of emotion when it was combined with a degraded auditory signal. These results suggest that when judging emotions from observed social interactions, we rely primarily on vocal cues from conversation rather than visual cues from body movement.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Figure 1: Wordcloud visualizing 150 most frequent words used in this thesis. The larger the size, the more frequently specific word has been used. The wordcloud was created using *R* programming language (R Core Team, 2005) with conditions to remove punctuation, symbols, numbers and stopwords.

1

INTRODUCTION

1.1 THE PUB EXAMPLE

Let us imagine it is a Friday evening in the life of a fictional character. Let us call him John. John has just finished work and his friends have called him to tell him that they are meeting at the local pub. John decided to join them and some time later he enters the pub. The scene he sees may be similar to the one shown in the image below:

Even in its static depiction, it is a very complex scene; dimmed lights worsen visibility and you can imagine the loud music and the numerous conversations from multiple sources which worsen auditory reception. There are approximately 12 people in sight; they shout, laugh and gesticulate. Initially stunned by the contrast between the breezy outside world and the crowded social space of the pub, John starts searching for his friends. He immediately spots two of them discussing something in the corner of the pub and he very quickly makes a judgement that one of them is angry. Such a scene in the pub may look like this:

Let us take this specific situation as a starting point for this thesis - a person looking at an emotional social interaction between two people, in sensory conditions which are not optimal. Let us consider how we can find out why John came to the decision that his friend was angry. Which cues are the most important for him to make emotional judgements in such a context? What specific factors made him decide that it was anger rather than happiness? What error is John likely to make in his judgement of emotions in such conditions? If John made a wrong judgement of emotions, what is the reason for it? If he clearly saw the body movement but only heard snippets of the conversation above the crowd noise, is the body movement going to be a driving signal for his judgement of anger?

One way to address these questions is to create a controlled experimental environment that emulates such natural pub conditions. It is a difficult task due to the complexity of this scene and the richness of the visual and auditory nonverbal cues, as well as a number

of other factors that are vital in our interpretation of emotions. For instance, visual nonverbal cues in emotion perception can include facial expression, head movement, posture, body and hand movements, self- and other- touching, leg position and movements, interpersonal gaze, directness of interpersonal orientation, interpersonal distance and synchrony or mimicry between people (Knapp & Hall, 2009). Auditory nonverbal cues in emotion perception can include discrete vocal sounds (e.g. sighs), the amount of speech, disfluency in speech, interruptions and pauses, but also prosodic cues[1] such as variations in pitch, loudness or speed (Scherer & Oshinsky, 1977). Additionally, there are a number of other factors such as gender (Hall *et al.*, 2000), attractiveness (Mehrabian & Blum, 1997), ethnic origin (Matsumoto, 1993), or linguistic fluency (Lattner *et al.*, 2005) that also affect emotion perception during interpersonal communication.

We cannot take all these cues and factors into consideration in our attempt to understand how John made the decision that his friend was angry, but clearly some of these cues and factors are more important than others in driving his judgement of emotions. In fact, the majority of the existing studies on emotional perception focus only on three channels of emotional expression: the face, the voice and body movement.

The human face is one of the most studied object categories in visual neuroscience (Dering *et al.*, 2011). It is well established in scientific literature that facial expression is central to the perception of emotions. However, there are many situations in which we are unable to clearly see the faces of the people we watch, such as in the pub example above. Even if the scene when John was looking at two people and making an emotion judgement is much less complex than the initial one when he entered the pub, there is still a lot happening. John cannot see his friends' facial expressions clearly due to poor lighting and distance and he cannot hear much of what they are saying because there is a lot of noise from the music and the other people talking around him. John can just about see their movements and gestures behind all the other customers passing by. This leads us to question how we can study perception of body movement in isolation from facial expression.

To begin with, an interaction between two people that is presented on image above can be simplified to show only movement information, such as in Figure 2 below:

1 Prosody may reflect various features of the speaker or the utterance: the emotional state of the speaker; the form of the utterance; emphasis, contrast, and focus; or other elements of language that may not be encoded by grammar or choice of vocabulary (Fernandez & Cairns, 2010).

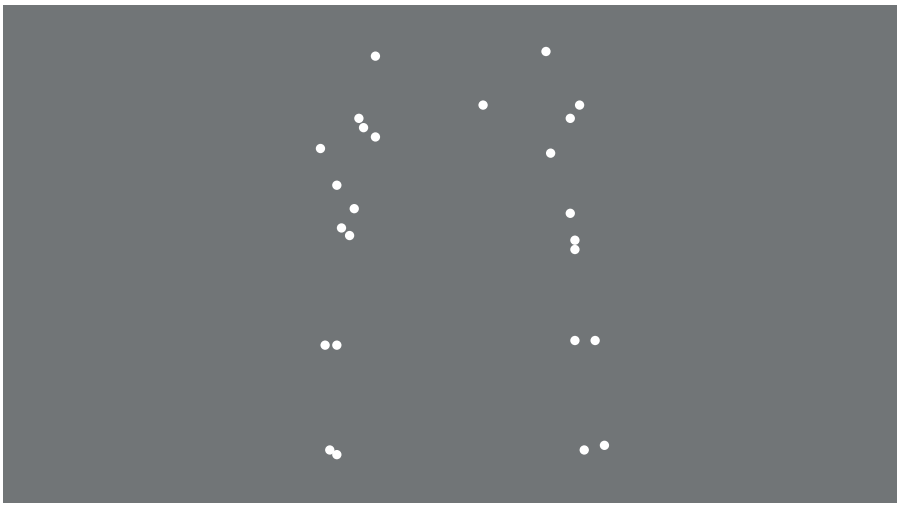


Figure 2: Example of dyadic point-light display.

Such a method of representing movement separately from other cues like clothing or body shape is one of the most common approaches in the study of human motion and is commonly referred to as the point-light display method (Johansson, 1973). We will discuss this method in more detail in the later Section 1.4 of this chapter as well as in Section 2.1.2 of Chapter 2.

The second issue is that, in the context of the pub example, John is observing others interacting rather than participating in the interaction himself. A number of studies have shown clear differences in how people perceive a single actor compared to perceiving a dyadic interaction between two people (e.g. Neri *et al.*, 2006; Centelles *et al.*, 2011). However, only a small group of researchers have utilized dyadic point-light displays (Clarke *et al.*, 2005; Manera *et al.*, 2011). There are no stimulus sets with such dyadic displays for the study of emotional interactions between people, and no stimulus sets that combine visual point-light displays with auditory voice dialogues.

Therefore, the goal of this thesis is to establish a new framework for the study of emotional social interaction in the multisensory context. As highlighted above, it is difficult to create a full, controlled replication of social situations and to understand how emotional judgements emerge in complex social scenes. It is necessary to break such social situations down into smaller components and examine more closely how each element of the percept contributes to the emotion judgements in a social context. This thesis focuses its methodological scope on social situations such as the one described in the pub example with the following principles:

- observer witnesses emotionally positive (happy) or negative (angry) interaction between two agents;
- such emotional interaction may have low, medium or high levels of intensity;
- such emotional interaction lasts around three seconds and consists of only two sentences of dialogue exchanged between two agents;
- observer cannot see the faces of the interacting agents clearly but he can see their body movements and hear their voices;
- there are various external conditions that can distort visual or auditory clarity of perceived interaction (e.g. noise coming from a crowd or the obstruction of view by other people passing by).

These principles guide the design of the new stimulus set and the consecutive experiments throughout this thesis. Before we describe the detailed motivation behind the studies presented in this thesis, we will first introduce existing relevant research in the field of emotion perception. We begin with a brief introduction of emotions with specific focus on happiness and anger. We then describe the role of social context in emotional perception and present research on body movement and voice as channels to communicate emotion. We finish with a general overview of studies on multisensory integration and highlight the structure and goals of the thesis.

To conclude this section of the Introduction, the main idea behind the pub example is that our social reality in everyday life resembles a complex, multisensory, emotional scene. We rarely perceive people in a perceptually clear and unobstructed context. In order to study perception of emotions in dyadic social interactions in the multisensory context, a new stimulus set is needed that will be flexible enough to enable us to manipulate visual and auditory cues, but simple enough to reduce the enormous complexity of social scenes. This thesis describes the creation and validation process of such a stimulus set and some experiments highlighting a potential research application for this stimulus set.

1.2 WHY HAPPINESS AND ANGER?

Theories on the experience of emotions date back to ancient times and the philosophical accounts of Hippocrates and Aristotle (Stearns, 1995). However, defining emotions is still a controversial topic. Nesse (1990, p. 263) provided a very good summary of this:

> More than 100 years later, each theorist still starts afresh. Agreement remains elusive even about basic issues. What are the emotions? Plutchik (1980) lists 27 different definitions. How many basic emotions are there? Each theorist has a different list. Does each emotion have an opposite? Some say yes; others, no. Which aspect of emotions is primary? Some say physiology; others, cognition; others, behaviour; and some say no single aspect is primary. Why do emotions all have hedonic valence? There is disagreement. And finally, what functions do the emotions serve? Some authors emphasize motivation; others, communication; and still others, cognition. There is no consensus on the answers to these major questions about the emotions.

In spite of these issues highlighted by Nesse (1990), there is some agreement that emotions can be described as discrete and consistent responses to internal or external events that have a particular significance for the organism (Nesse, 1990; Scherer, 2005). Emotions are brief in duration and consist of a coordinated set of responses, which may include verbal, physiological, behavioural, and neural mechanisms (Fox, 2008). There is substantial evidence to demonstrate that emotions have a physiological basis. For example, Bechara (2000) has shown that damage to the prefrontal cortex is inversely associated with patients' abilities to process emotion normally. Significant advances in our understanding of the neural bases of emotional processing have also been made in recent decades. Overall, studies of humans and other animals highlight the key role of the amygdala in the detection and evaluation of stimuli with affective value (see Armony, 2012 for review). Nonetheless, contradictory findings have also been reported, especially in terms of the exact role of this structure in the processing of different emotions, giving rise to different neural models of emotion. For instance, although the amygdala has traditionally been considered as exclusively involved in the processing of fear and anger, more recent work suggests that it may be important for processing other types of emotions, and even non-emotional information (Sander *et al.*, 2003; Armony, 2012) .

Most emotions have clear behavioural characteristics. For example, high intensity anger has very obvious behavioural signals that are expressed via the voice, the face and body movement. Many of these signals seem to be universal across cultures. Such a notion goes back to Darwin (1872) who argued that emotions evolved via natural selection and therefore have universal cross-cultural counterparts. In this evolutionary account, emotions are crucial in animal communi-

cation and aid their survival. Drawing on Darwin's intuition and observations about the universality of emotional expressions, Ekman & Friesen (1971) found that certain emotions appeared to be recognized even in cultures that were preliterate and could not have learned associations for facial expressions through media[2]. In their pioneering studies in the field of facial emotion perception, Ekman & Friesen (1971) precisely defined those groups of muscles that allow people to produce specific emotional expressions. Since then, Ekman's *Facial Action Coding System* has been broadly applied as a guideline for studying facial emotion perception with well defined expressions being anger, disgust, fear, happiness, sadness and surprise (Ekman & Friesen, 1978). For example, Ekman & Friesen (1978) demonstrated that an expression of happiness is characterised by cheek rise and lip corner pull while an expression of anger is characterised by lower brow, upper lid rise, and lip tightening (see example in Figure 3).

(a) Expression of happiness. (b) Expression of anger.

Figure 3: Example of facial emotions expression of happiness and anger from Ekman & Friesen (2003).

The classification of emotions has been subject to ongoing scientific debate. Most theorists agree with Ekman's notion of 'basic' or 'default' emotions that are a baseline for other expressions, but there is little consensus for what should be defined as 'basic'. The differences lie mainly in the inclusion requirements for specific emotions. For ex-

2 Some researchers question universality of emotions; for example see Jack *et al.* (2009) for studies arguing for cultural differences in emotional facial expressions between Western Caucasians and East Asians.

ample, Frijda (1986) has defined desire, happiness, interest, surprise, wonder and sorrow as basic emotions, arguing that their inclusion is based on forms of action readiness. In contrast, Gray (1982) included rage, terror, anxiety and joy as being 'hardwired', basic emotions. However, nearly all researchers include anger, happiness/joy, sadness and fear as basic emotions (Ortony & Turner, 1990).

The scope of this thesis focuses only on two emotions - happiness and anger. There are a number of reasons why we decided to focus on only these two. Happiness and anger are relatively prototypical emotions in human experience and are easily available when people introspect about experienced emotions. For instance, Scherer & Tannenbaum (1986) asked participants to report the most recent situation that evoked strong emotional feelings and to describe the pattern of their reactions. Only happiness and anger were reported as relatively pure feeling states. Most other reports were emotion blends, with anger/sadness and sadness/fear occurring most frequently (Scherer & Tannenbaum, 1986). In short, people have a good internal cognitive representation of happiness and anger.

Other reasons for using only happy and angry emotional interactions relate to the difficulty of actors performing or acting other types of emotions. We wanted to avoid reactive emotions such as surprise or disgust because they are associated with specific reactive movements and are difficult to perform (Konijn, 2000). A number of studies have shown that actors find angry and happy emotional expressions easy to convey in various scenarios and observers can easily recognize such expressions (Pollick *et al.*, 2001, 2002; Ma *et al.*, 2004). By contrast, people easily confuse and have difficulty in conveying other 'basic' emotional expressions such as disgust, fear or surprise (Knutson, 1996).

Another reason why we focused on happiness and anger was because we wanted to explore the intensity of expressions *within*, rather than *between* emotions. On an everyday basis, we do not experience or watch anger and happiness on just a single level of intensity. Emotions vary, shift and change in intensity depending on the situation, and these shifts introduce an entirely new quality to the emotional social interaction. For such an exploration of *within* emotion intensity, happiness and anger are perfect candidates because both emotions are easy to map on the arousal spectrum. For anger - from low arousal irritation to high arousal rage; and likewise for happiness - from low arousal pleasure to high arousal elated joy (Dael *et al.*, 2012). We wanted to explore how people detect different levels of arousal or

emotional intensity when they observe and hear interaction between two agents.

From the perspective of voice expression of happiness and anger, these two emotional states share many common vocal indicators: the pitch level is high and the pitch variability broad, the tempo is fast and the loudness is high (Scherer, 1986). Likewise, kinematic similarities can be found for bodily expressions of happiness and anger such as speed in the use of arm gestures (Pollick *et al.*, 2001). Despite sharing these common characteristics, both happiness and anger are easily detected by observers, indicating that people are very sensitive to cues related to these two emotions. At the same time, there are a number of differences that have been observed between happy and angry expressions in voices and body movement. By default, anger and happiness represent the opposite valence of the affective expression. Ekman (1994) has argued that the antecedents for happiness are pleasure, praise, relief and excitement. Unlike love and sadness, happiness may not necessarily be focused on a particular person or event (Averill & More, 2000). Antecedents for anger include threats and insults or an incident that violates one's own values (Ekman, 1994). Overall, anger is easier to detect from body movement than happiness (Ikeda & Watanabe, 2009; Shiffrar, 2011), although happy voices were found to elicit more activation than angry voices in numerous brain areas, suggesting a particularly salient role for vocal expressions of happiness (Johnstone *et al.*, 2006). On the whole, happiness and anger were ideal choices in the study context of multisensory social interactions.

1.3 OBSERVING EMOTIONAL SOCIAL INTERACTIONS

Humans are social in nature. From the evolutionary perspective, human brain functionality reflects the complexity of social environment[3]. This social function of the human brain is also reflected in the existence of a specific type of neurons called 'mirror neurons'. Mirror neurons fire both when an animal acts and when the animal observes the same action performed by another animal (Rizzolatti & Craighero, 2004). These neurons were initially discovered in monkey brains but have since been investigated in the human brain using functional magnetic resonance imaging (fMRI) (Gallese *et al.*, 2004).

3 One example is Dunbar's theory (1992) that neocortex size in humans constrains the size of the social group with which we can interact. Based on those assumptions, supported by historical demographics and social networks analysis, Dunbar (1992) argues that the number of people we can *consciously* keep in our social scope is around 150.

The function and existence of mirror neurons is still controversial but there is good evidence to suggest that these neurons play a significant role in social interactions. Theorists argue that mirror neurons are essential in a number of social processes including learning via imitation (Rizzolatti *et al.*, 2001), gestural communication (Armstrong *et al.*, 1995; Corballis, 2002) and speech evolution (see Rizzolatti & Craighero, 2004, for a review). Another popular hypothesis is that mirror neurons support a brain system that plays a role in the automatic *prediction* of social outcomes of actions (see Brown & Brune, 2012, for a review). If this were the case, it would literally signify that humans are inherently hardwired for social interaction by their brain structure.

While humans are social in nature, emotions serve as crucial components of social communication. As mentioned previously, from the evolutionary perspective emotions have evolved in a social context and should therefore be beneficial for social survival. Social survival is a complex endeavour because it requires a balance between cooperation on the one hand and competition on the other. Fischer & Manstead (2008) argue that emotions are important to social survival because the emotions we experience and express help us to form and maintain social relationships and establish or maintain a social position relative to others. Additionally, emotions have a clear communicative function. Ekman (1999) postulated that emotions are expressed in order to communicate to others, informing the observer that something important is happening. The communicatory function of emotions occurs from an early age. This can be observed by the way infants will look towards their caregiver for assurance that a new object can be approached. If the caregiver smiles, the child will approach the object but if the caregiver expresses fear, disgust or anger the child will retreat (Klinnert *et al.*, 1986). Happiness (sharing positive experiences with others) or sadness (seeking help and support from others) serves the affiliation function, while social distancing functions can be observed in anger (seeking to change another person) and contempt (seeking to exclude another person)(Fischer & Manstead, 2008). Rather than thinking of emotions as intrapersonal states, they can be thought of as a dynamic interpersonal process that occurs between the individual and the environment (Campos *et al.*, 1989). Therefore, the study of emotions at the social level must be of interest as emotions have interpersonal outcomes. The social context of human communication and the social nature of emotional exchange are therefore central topics of this thesis research project.

In the context of the social nature of emotional exchange, it is somewhat surprising that the majority of studies into perception of emotions involving faces, voices or body movement have used displays with a single agent. In fact, there are a growing number of studies indicating substantial differences between the situation where we watch a single person compared to the situation where we watch two people interacting. Social processes change fundamental aspects of visual and auditory perception (Scherer, 2003; Shiffrar, 2011). Neri *et al.* (2006) demonstrated that observers could efficiently use information detected from one of the agents to predict an action or response from the other agent. They established that when participants observed fighting and dancing actions, these meaningful interactions enhanced the visual discrimination of agents. Manera *et al.* (2011) presented participants with point-light depictions of two agents either communicating or acting independently from each other. They showed that the communicative gestures of one agent could serve as a predictor for the expected actions of the respondent, even if no physical contact between the agents was implied. In another study, also using communicative and non-communicative depictions of interactions, Manera *et al.* (2013) showed that the communicative gestures of one agent could serve not only to predict *what* the second agent would do, but also *when* his or her action would take place.

Neuroimaging studies have also revealed overlapping neural circuitry involved in the perception of emotions, social cues, and human movement (e.g. Beauchamp *et al.*, 2004; Kreifelts *et al.*, 2007). Centelles *et al.* (2011) presented participants with both social and non-social versions of dyadic point-light interactions. A social version depicted one agent pointing to something on the ground when facing the other agent, while in the non-social version agents simply acted independently (e.g. one jumped, another raised a leg). They used the fMRI method to determine which brain regions were recruited during the observation of the two interacting agents. While the mirror neuron system and mentalizing networks[4] were rarely concurrently active, the authors found that both of these networks were needed to catch the social intentions carried by whole-body movement. This adds to the argument that observation and understanding of multiagent in-

4 Brain structures frequently referred to as mentalizing networks (devoted to mentalizing processes) typically include the left temporo-parietal junction, the right anterior superior temporal sulci and the dorsal part of the medial prefrontal cortex. The areas frequently referred to as mirror systems (the action observation/execution matching networks) typically include the inferior frontal gyrus, the premotor cortices, bilateral intra-parietal sulci and the right superior parietal gyrus (Saxe, 2006; Gallagher & Frith, 2003; Frith, 2007)

teractions is connected to the broader spectrum of information processing and represents a complex phenomenon in human perception, which is different from the context when people observe a single person.

The problem of using single rather than multiagent stimuli was also raised in a broader context by Risko *et al.* (2012, p. 1) in their recent review:

> This issue [i.e. single versus multiagent stimuli] has recently surfaced in the context of research on social neuroscience given its reliance on stimuli more akin to simple, static representations of socially relevant stimuli than an actual live social interaction in attempting to map the social brain. One of the critical assumptions driving social neuroscience is that the knowledge gained about the social brain using the former class of stimuli will generalize to the richer scenarios associated with everyday social cognition.

Risko *et al.* (2012) argue that there is a need to use stimuli that better represent a natural social environment. They acknowledge that this is a challenge because such an approach brings methodological problems related to the complexity of social stimuli and control over confounding variables. Nevertheless, Risko *et al.* (2012) suggest that the effort is worth the price and one constructive approach would be to compare stimuli that range in their approximation to a real social interaction. Indeed, one of the major goals of this thesis is to replicate some of the approaches used in the study of perception of emotions from body movement and voice but with stimuli that approximate short, real-life interactions between two people.

1.4 BODY MOVEMENT AND VOICE AS EMOTIONAL AND SOCIAL STIMULI

Darwin's classic paper (1872) highlights that both face and voice are critically important in emotion production. However, research into emotion perception from voice has long been neglected. Scherer (1986) argues that this neglect was largely due to methodological limitations in the past, such as the difficulty of storing the sound for analysis before the advent of audio recorders, the problem of graphic representation of speech sound, and the distinction between the linguistic and paralinguistic domains. In recent years, speech scientists

and engineers have started to devote more attention to speaker emotions (Scherer, 1986).

The use of voice stimuli in this thesis builds on the principle that voice is an inherently social stimuli. Voice evolved as an effective method of complex communication and is considered by some as one of the most important factors in the evolution of human intelligence. When we observe people interacting, they are typically involved in some form of verbal exchange. In general, people can easily detect emotions from specific auditory cues in vocal expressions (for a review on emotional voice perception, Scherer, 2003). Basic emotions can be recognised in the voice independent of verbal information (Scherer, 2003), as can more abstract categories of language communication such as sarcasm (Bryant & Fox Tree, 2005). Prosodic features of speech have general and robust effects on listeners' emotions and evaluation of vocalizers (Scherer, 1986). Features such as pitch, loudness, duration and spectral properties often form a stereotyped configuration that relates in systematic ways to emotional categories (Cosmides, 1983). For instance, vocalizations conveying happiness tend to be high in average pitch, high in pitch variability, loud and fast. Sad vocalizations tend to be low average pitch, low in pitch variability, soft and slow (Scherer, 2003). Overall, a large change in pitch and duration was shown to contribute most to the transmission of emotions, whereas loudness seems to be the least important factor (Frick, 1985; Murray & Arnott, 1993). Universals in human vocal production have been proposed as well, including those in infant directed speech, which serve in part to convey affective information, such as approval or disapproval (Bryant & Barrett, 2007). Similar acoustic patterns in infant directed speech have been identified in all cultures studied to date and are likely to be universal (Bryant & Barrett, 2007, 2008). Voices are rich in information about a person's identity, affective state and intention (Belin *et al.*, 2004; Campanella & Belin, 2007).

Of even higher importance than voice is the ability to produce body movements that convey emotional and social signals. Some researchers argue that the primary cognitive function of humans is the capacity to produce adaptable and complex movements and that the complexity of the brain reflects the ability for complex drive and suppression of future movements (Wolpert *et al.*, 2003, 2011). While such a view may be perceived as simplistic, it is clear that body movement is a basis for most communicative acts in social interaction. In evolutionary terms, Rizzolatti & Arbib (1998) argue that body movement is an important feature of humans' phylogenetic past as a precur-

sor to language acquisition. They suggest that early hominids could have described an object or an event by using specific configurations of arm movements that over time became accompanied by vocalizations. In this context, Rizzolatti & Arbib (1998) contend that it was unlikely that facial movement was a precursor to language acquisition for two reasons: firstly, facial communication is limited to a few agents and secondly, actions directed towards objects are inherent in body language. Therefore, speech is likely to have evolved from body movement due to the flexibility and potential expressiveness of arm movements (Rizzolatti & Arbib, 1998). For body movement to be a cue would also signify that humans can pick up information that can be utilized when other communication channels are not easily available. Perception of body movement is fundamental as it is likely to be the first available information when approaching or being approached by another person, or observing interactions between other people from a distance.

Compared to studies on emotional expression in face and voice, expression of body movement has been a less explored mode of communication. This is not surprising as body movement involves incredibly complex mechanisms. The average adult human has 206 joints that have a great degree of freedom to produce countless movement combinations (Gray, 2007). In such a context, it is very difficult to create a system for the categorization of emotional expression of body movement similar to the action unit coding system created by Ekman & Friesen (1978) which exists for facial expressions.

In the research community, point-light displays have been frequently used to study the perception of body movement, mainly because this method allows the study of body movement in isolation from other contextual cues such as clothing, facial expression or body shape (Johansson, 1973). The point-light technique preserves the natural relative movements of body parts, but eliminates most morphological and contextual cues (see Figure xx earlier in this Chapter). A large number of studies have shown that observers can recognize specific actions (Dittrich, 1993; Vanrie *et al.*, 2004), gender (Mather & Murdoch, 1994; Troje, 2002), age (Montepare & Zebrowitz-McArthur, 1988), identity (Cutting & Kozlowski, 1977; Hill & Pollick, 2000) and affect (Dittrich *et al.*, 1996; Pollick *et al.*, 2001; Atkinson *et al.*, 2004; Clarke *et al.*, 2005) from just a set of point-lights representing the main joints of human movement. For example, Pollick *et al.* (2001) used point-light displays of knocking, lifting and drinking arm movements (i.e. knocking on a door angrily) to demonstrate that participants can perceive a range of affects (e.g. fear, anger, tiredness) from

these actions. Dittrich *et al.* (1996) established that surprise, fear, anger, disgust, grief and joy could be identified when portrayed by dancers in point-light displays. Clarke *et al.* (2005) showed that a range of emotions can be recognized in the biological display of actors engaged in dialogue, which is in line with previous research.

1.5 AUDIO-VISUAL INTEGRATION OF EMOTIONAL SIGNALS

From our introduction so far regarding the separate studies on perception of voices and body movement, it is clear that humans can detect emotions from these cues, but it is less clear how people integrate such independent signals. The majority of studies on multisensory aspects of affect perception use faces and voices and show strong bidirectional links between vision and audition (e.g. Massaro & Egan, 1996; de Gelder & Vroomen, 2000; Ethofer *et al.*, 2006). For instance, de Gelder & Vroomen (2000) presented participants with static photographs of emotional faces combined with short vocal verbalizations. The important aspect was that emotions were presented in a discordant manner; for example, a sad face combined with a happy voice. When the participants were asked to identify the expression of a face while ignoring a simultaneously heard voice, their choices were nevertheless influenced by the tone of the voice. Conversely, when asked to identify the tone of a voice while ignoring a simultaneously presented face, the participants were influenced by the expression in the face. The study by de Gelder & Vroomen (2000) clearly shows mandatory integration of auditory and visual signals rather than post-perceptual decision under cognitive control. Similar results were obtained when using dynamic, instead of static, facial expressions combined with voice (Collignon *et al.*, 2008). Other studies have shown that the presence of facial expression combined with vocalization usually enhances the observers' ability to identify emotional expression (Massaro & Egan, 1996; Dolan *et al.*, 2001; Kreifelts *et al.*, 2007). Since the majority of these integration processes occur mandatorily and irrespective of attention (de Gelder & Vroomen, 2000; Vroomen & de Gelder, 2000; Ethofer *et al.*, 2006), one might assume that the audio-visual integration of nonverbal affective information is an automatic process. This assumption gains further support in the results of electrophysiological and fMRI experiments that provide evidence for audio-visual integration during an early perceptual stage (de Gelder *et al.*, 1999; Pourtois *et al.*, 2005; Kreifelts *et al.*, 2007).

Only a small number of studies have examined how observers integrate signals from emotional body movement and voice, and re-

sults so far follow a similar pattern to studies of emotional faces and voices. Two studies have demonstrated that perception of emotions expressed in static body postures (Van den Stock *et al.*, 2007) and whole body movement (Van den Stock *et al.*, 2008) was influenced by affective information in voice. Van den Stock *et al.* (2008) presented dynamic whole body expressions of emotion matched with nonverbal auditory information consisting of human vocalizations and also animal sounds. They instructed participants to attend to the action displayed by the body and to categorize the expressed emotion. The results indicate that recognition of body language was biased towards the emotion expressed by the simultaneously presented auditory information. In a separate study, Stienen *et al.* (2011) presented participants with static postures of emotional body expression combined with a happy or fearful voice. They showed that when bodily expressions were presented outside visual awareness, it still influenced prosody perception.

It is clear that body movement and voice are closely connected and for two people to be engaged in discourse is a very natural situation and one that we encounter on a daily basis. Often, social synchrony and patterns of entrainment can be readily perceived in social interactions (McClave, 2000). Jessen *et al.* (2012) have also shown that ecologically valid complex stimuli such as joined body and vocal expressions are effectively integrated very early in perceptual processing. Nevertheless, there have been literally no studies examining integration of body movement and voice when observers watch two people interacting. For example, in the Clarke *et al.* (2005) study mentioned previously, the voice dialogue was not used together with point-light captures. Clarke *et al.* (2005) explained that this was to avoid any form of auditory-based attribution that the observer would make on the basis of speech and its semantic content. Nevertheless, it leads to the loss in content of the action where the voice dialogue was integral to the interaction between the actors. On the other hand, Van den Stock *et al.* (2007) and Stienen *et al.* (2011) used combined body movement and vocal expression in their experiments, but they missed the following key aspects important in the study of perception of emotions in social interactions:

- they used a single actor instead of dyadic interactions;
- they used static displays of body postures instead of dynamic displays of whole body movement;
- they used short vocalizations with only prosody information instead of short, but fully meaningful and intelligible dialogues;

- they used whole-body displays instead of using motion-specific and cue-controlled point-light displays.

These four points highlight the key elements that separate the work in this thesis from the methodological approach of Clarke *et al.* (2005); Van den Stock *et al.* (2007, 2008); Stienen *et al.* (2011) and Manera *et al.* (2011). Instead, this thesis focuses on developing, validating and using short dyadic point-light displays with simple voice dialogue. Such an approach enables us to test and apply a stimulus set for further research on the integration of emotional and social signals from body movement and voice.

1.6 SUMMARY OF THE GOALS OF THIS THESIS

Our sensory apparatus allows us to register the body movements and voice of other conspecifics to make sense of our social environment. Emotions are also fundamental aspects of being human. Several studies in the past have investigated the perception of emotions from body movement (by employing the point-light display technique) and voice separately (e.g. Dittrich *et al.*, 1996; Pollick *et al.*, 2001; Scherer, 2003; Atkinson *et al.*, 2004; Clarke *et al.*, 2005; Bryant & Barrett, 2007), but they have not considered multisensory aspects in the context of multiagent social interaction. This thesis investigates a range of independent themes related to the perception of emotions from body movement and voice in the multiagent context. The goal of this project is to better understand how we integrate emotional and social signals in the different sensory conditions when we watch and hear other people interacting. A further aim is to develop and validate a unique stimulus set that can be used to investigate how perception of emotions is affected by the multiagent context and how such context influences the integration of signals from body movement and voice.

These questions are addressed in four empirical chapters. In Chapter 2, we describe the methods and process of creating a new stimulus set of dyadic point-light interactions combined with voice dialogues. We highlight the motivation behind creating the stimulus set and describe details of the complex process of capturing and preparing the stimulus set for experimental research. Next, we establish whether observers can recognize the intended emotions when they are presented with the stimulus set developed in Chapter 2. This process and the results of the experimental validation of the stimulus set are described in Chapter 3. At this stage, we conduct a series of experiments presenting participants with visual point-light displays, auditory voice dialogues or combinations of both visual and auditory displays.

In Chapter 4, we examine how people identify emotional interactions from those displays under impoverished visual and auditory conditions. For the experiments described in Chapter 4, we use a subset of our validated stimulus set to investigate how presentation of distorted visual or auditory displays influence emotional perception of interactions. The conditions we test for visual displays are: scrambling the spatial location of point-light displays, presenting displays upside-down and presenting displays from different viewpoints. For auditory displays, we use brown noise and low-pass filtering as distortion methods. All these stimuli distortion techniques are commonly used by the research community to degrade the quality of point-light displays and auditory clips (refer to Ahlström *et al.*, 1997 for a review on point-light displays methods, and Knoll *et al.*, 2009 for a review on voice methods). The experiments and methods described in Chapter 4 build on the principle that the natural social environment is inherently noisy (Risko *et al.*, 2012) due to varying environmental conditions (e.g. weather, amount of light, distance from agents, location from which we observe agents). For example, we can observe people in poor lighting conditions or we can hear them during heavy rain. These conditions inevitably distort the quality of sensory information we receive which can respectively affect our judgement of a perceived situation. Degrading the quality of visual and auditory stimuli serves to replicate such natural environmental conditions and to understand how degraded conditions can affect perception of emotions when watching interactions between two people.

The study of either body movement or voice can provide some understanding of emotional perception, but to get a clearer picture we need to understand how these signals are integrated together. Integration between faces and voices has been well studied, showing bidirectional links between vision and audition; however, integration between body movement and voice has received less attention. In Chapter 5, we examine how participants integrate emotional signals from body movement and voice when they watch interactions between two actors. In a series of experiments we use a paradigm frequently employed in multisensory integration studies on faces and voices (e.g. de Gelder & Vroomen, 2000; Collignon *et al.*, 2008). Specifically, we introduce conditions where the emotional valence of the visual component (e.g. happy body movement) is different from the valence of the auditory component (e.g. angry voice). These mismatching displays help us to understand which sensory cues participants use more of to make judgements of emotions. Similar to Collignon *et al.* (2008), we also introduce conditions where participants are asked to specif-

ically focus on either vision or voice and ignore the other modality. This further helps us to understand whether a mismatch between emotional signals affects perception of emotion when watching interactions even if participants are asked to ignore one of those signals. In the final study described in Chapter 5, we use a paradigm of optimal cue integration similar to Ernst & Banks (2002) to investigate whether two cues can reduce sensory uncertainty for high level factors such as perceived emotions. We conclude with Chapter 6 which summarizes the most important results from the previous chapters and discusses them in the context of the existing research. In Chapter 6, we also highlight limitations and potential applications and further direction in research using the results and the stimulus set developed in this thesis.

2

CREATION OF A STIMULUS SET FOR THE STUDY OF MULTISENSORY SOCIAL INTERACTIONS.

2.1 INTRODUCTION

2.1.1 *Chapter summary*

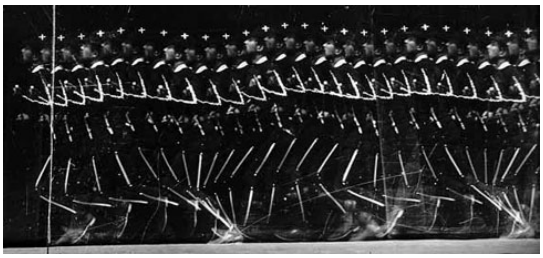
This chapter describes the process of creating a set of stimuli for the study of multisensory social interactions. A short overview of motion capture systems is given with an outline of point-light displays as a method to study perception of motion. There follows a description of existing stimulus sets for the study of the perception of emotion from body movement and the voice. The advantages and limitations of existing stimulus sets are provided with a discussion of the main reasons why a new set was created. Subsequently, we move to the second part of the chapter which describes the methodology behind the creation of stimulus sets for the study of social interactions. We highlight the difficulties of creating a realistic, multisensory and affective stimulus set that involves social interactions. We then move to a description of the stimulus creation process which includes actor recruitment, capturing movement and voice data, post-processing of the captured data and preparation of the final stimulus set for validation. We conclude this chapter with a brief summary and discussion of the advantages and disadvantages of the created stimulus set.

2.1.2 *Brief overview of motion capture techniques used to create point-light displays.*

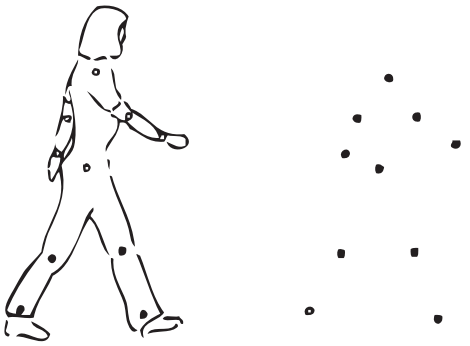
The fascination with human motion can be traced back to ancient Greece when Aristotle (-383 to -322) published *On the Parts of Animals*, but only in modern times have we gained the technological capability to capture details of human movement. In the 19th century, Marey (1898) used chronophotography[1] to capture rapid sequences of human and animal movement, as seen in Figure 4a. Marey (1898)

1 Chronophotography is an antique photographic technique from the Victorian era which captures movement in several frames of print. These prints can be subsequently arranged either like animation cels or layered in a single frame. It is a predecessor to cinematography and moving film, involving a series of different cameras, originally created and used for the scientific study of movement (MacDonnell, 1972).

recorded actors dressed in black robes with limbs marked by white lines. However, his solution only gave a two-dimensional (2D) representation of human movement and it was an extremely labour intensive process (Dekeyser *et al.*, 2002). Nevertheless, Marey's (1898) capture solution using chromophotography may be considered to be the precursor to the modern methods used by vision scientists to study the perception of human movement - point-light displays.



(a) Chronophotograph of running man - from Marey (1898).



(b) Point-light walker - from Johansson (1973).

Figure 4: Example of chronophotography and schematic drawing of point-light walker.

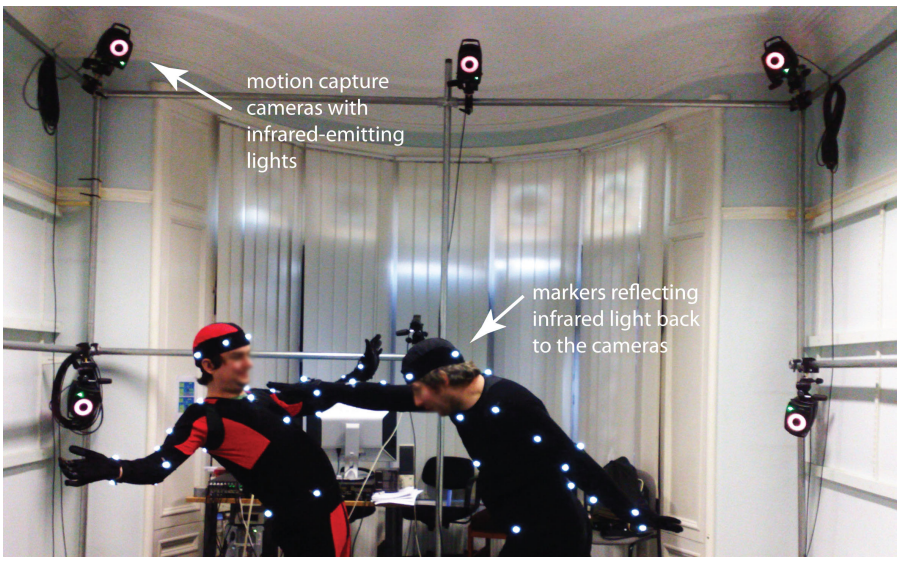
Point-light displays came to light when Johansson (1973) pioneered studies into the use of image sequences for human motion analysis. He used a video camera in a darkened room to record actors wearing a black suit with light bulbs attached to specific joints on the body. The recordings produced very natural point-light animations after their contrast was tuned to maximum and their brightness to minimum (Figure 4b). This recording technique proved to be very successful amongst vision scientists (Cutting & Kozlowski, 1977; Barclay *et al.*, 1978) but it continued to suffer from limitations similar to Marey's (1898) technique: it was only possible to formulate a 2D representation of movement and points were frequently occluded by actors during movement. Cutting (1978) introduced a further advancement for creating point-light displays. He described a programming

algorithm for simulating a point-light walker on a computer display. The Cutting (1978) algorithm became very popular because it allowed manipulation of different features of point-lights, such as contrast (Mather *et al.*, 1992), temporal coordination (Bertenthal & Pinto, 1994) and position of point-lights (Cutting, 1981; Verfaillie, 1993). The Cutting (1978) algorithm also allowed for the setting up of experiments very quickly without the need to record actors. However, the main limitation of algorithm-based simulation of movement was the lack of fine-features of natural movement and the need to set up a separate algorithm for each independent action (Dekeyser *et al.*, 2002).

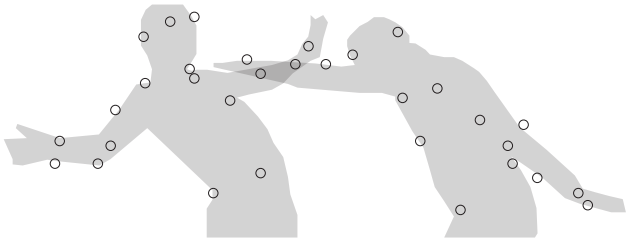
It was the technological advancement in motion capture systems that allowed vision scientists to truly capture a three-dimensional (3D) representation of human movement. There are several motion capture techniques that have been developed over the last 20 years but most can be categorized as either optical or non-optical systems (Moeslund & Granum, 2001). Optical systems use video processors connected to cameras that capture the signals emitted or reflected by the markers on the actor's body (Moeslund & Granum, 2001; Klette & Tee, 2008). Non-optical systems are not visually based. Instead, they use other methods of transmitting movement information such as wireless inertial sensors (Klette & Tee, 2008) or mechanical exoskeletons worn by actors (Vlasic *et al.*, 2007). In general, motion capture systems have a number of advantages compared to the video recording and algorithm-based synthesis. Motion capture allows us to record the position of markers as a set of 3D coordinates rather than 2D video sequences. Capturing the position of the markers enables the manipulation of the captured motion, making it possible to display the same action from any viewpoint, or distort other parameters of recorded movement such as velocity or size (Dekeyser *et al.*, 2002). This flexibility in the data format of 3D motion recording allows one to use the same captures to create a wide range of computer animated characters - either point-lights or solid body displays.

A motion capture system only records the markers and ignores all other layers of information present in the recording space, such as clothing, facial expression and surrounding objects. Moreover, this system allows the tracking of markers with a much higher frame rate compared to conventional video cameras, enabling greater detail in motion capture. The main disadvantages of motion capture compared to video capture are the high cost of the specialized equipment, the considerable effort associated with the recording and the inability to use motion capture systems in natural environments (Dekeyser *et al.*, 2002; Klette & Tee, 2008). We used the Vicon optical motion

capture system (Vicon, 2010) which utilizes reflected infrared light to passively locate a position of markers attached to actors, as seen in Figure 5.



(a) Vicon infrared cameras and actors wearing markers traced by the cameras.



(b) The system ignores surrounding environment, clothing and skin, capturing only the position of markers (black circles).

Figure 5: Scene from optical motion capture recording

2.1.3 *Existing stimulus sets for the study of perception of emotions from body movement and voice*

Among the numerous stimulus sets that study the perception of emotions, there is a clear lack of sets that employ audio-visual stimuli in the context of social interactions. There are a few speech-oriented stimulus sets that use naturalistic, interactive discourse and a varied range of emotional interactions, some with video recordings. Douglas-Cowie *et al.* (2003) created the Belfast natural database composed of 125 natural clips taken from television shows. Scherer & Ceschi (1997) cconducted the Geneva Airport Lost Luggage study with 109 natural, unobtrusive videotapings of passengers at a lost luggage counter,

followed by interviews with the passengers. There is also a Reading-Leeds database with five hours of material recorded from natural, unscripted interviews taken from radio and television shows, in which speakers have been induced by interviewers to relive emotionally intense experiences (Roach *et al.*, 2009). However, the drawback of naturalistic stimuli is the lack of control over speech content and the limited possibility of statistical analysis of naturalistic stimuli (Douglas-Cowie *et al.*, 2003; Ververidis & Kotropoulos, 2006). Additional factors, such as differences in the language spoken, the length of clips, and the vague specification of emotional expressions adds to the problems of using these naturalistic stimuli in the studies on perception of emotional social interactions.

In a similar manner, only a limited number of stimulus sets utilized point-light displays and conversation in the context of emotional social interactions. The Manera *et al.* (2011) motion capture database included 20 communicative interactions such as sharing, ordering, giving information, helping and offering, but without voice capture and with no emotional component involved. Still, it is currently the only validated stimulus set to study communicative social interactions using a point-light display approach. However, the focus of Manera *et al.* (2011) was to study the perception of visual communicative interactions, rather than audio-visual emotional interactions. In addition, Manera *et al.* (2011) used only a small number of actors to create the stimuli (one male and one female). Douglas-Cowie *et al.* (2003) suggested that a larger number of actors would increase the variety of emotional expression that is frequently specific to the individual differences between actors. In the case of the Manera *et al.* (2011) stimulus set, actions from only a few actors served well to explore simplified and emotionless communicative actions, but a reduced number of actors would be a limitation in examining emotional interactions.

Busso (2008) created a motion capture database of actors' faces and hands combined with conversational speech, recorded in improvised interactive scenarios with the emotional interactions being happiness, anger, sadness, frustration and an emotional neutral state. The Busso (2008) database was recorded during conversations but the stimulus set itself does not show the interaction between actors, only actors' heads and hands, which makes it unsuitable for examining perception in a multiagent context. The Clarke *et al.* (2005) study is one of the only attempts to use point-light display interactions to examine the perception of anger, love, sadness, fear and joy. Clarke *et al.* (2005) used emotional point-light interactions that were captured during scripted dialogue but they did not include auditory dialogue with

the visual displays and their stimulus set has not been made publicly available. None of the other existing stimulus sets truly combine body movement and voice in the context of emotional social interactions.

2.1.4 *Motivation and challenges behind creating a stimulus set with emotional social interactions*

Considering the lack of stimuli for the study of emotional social interactions using body movement and voice, the first major goal of the thesis was to create such a stimulus set. The task of creating an emotional interactions dataset has been difficult, taking into account the complexity of human movement and interactions. Ideally, we would capture natural interactions by placing motion capture cameras and microphones on the streets, in pubs and in public places to capture real world interactions in their natural environment. This was clearly not possible due to both technical and ethical issues. Instead, we focused on a more practical and realistic application using an optical motion and voice capture system and inviting pairs of actors to the lab.

One challenge was capturing actors with different levels of acting experience. We captured both experienced and inexperienced actors to address some of the ambiguity in the existing studies regarding actors' experience. Ma *et al.* (2006) and Rose & Clarke (2009) argued that experienced actors tend to systematically exaggerate emotional expressions, a trait which emerges from their theatrical training. Roether *et al.* (2009) found no differences between experienced and inexperienced actors in terms of acting quality. Still, Ma *et al.* (2006) highlighted that exaggerated behaviour could be a part of natural expression and it is sometimes difficult to draw the line as to where genuine expression turns into exaggeration. However, Busso (2008) argued that experienced actors are typically better than inexperienced actors during scripted scenarios. We decided to use both experienced and inexperienced actors to examine whether there would be any differences for the observers to identify emotions depending on the experience of the participating actors.

A further challenge was posed by the complexity of emotional social interactions. It is difficult to obtain realistic emotional interactions for the entire spectrum of emotions using simulated actions. This was one of the reasons we decided to capture only happy and angry interactions with different levels of intensity. In normal daily life, people express emotions with various intensity and we wanted to take this variability into consideration. We also wanted to obtain a large

variance of interactions within the happy and angry emotional expressions rather than having a broad scope of different emotions. In terms of reference to the structure of representation of affect, anger and happiness can be anchored by two bipolar but independent dimensions of experience (pleasantness and activation) which maps well in the context of a circumplex model of emotion (Russell, 1980). In the context of this model, both anger and happiness represent the same level of arousal but with opposite valence, which makes those emotions easier and more valid to test as a stimuli (Remmington *et al.*, 2000; Pollick *et al.*, 2001). Furthermore, anger and happiness are the most frequently reported emotions when people are asked to introspect about their experienced affects (Scherer & Tannenbaum, 1986). Both of these emotions represent emotional states or moods that might last for an extended period of time (Ma *et al.*, 2006). A number of studies found that actors find angry and happy emotional expressions easy to convey in various scenarios and observers can easily recognize such expressions, although anger is usually easier to identify than happiness (Pollick *et al.*, 2001, 2002; Ma *et al.*, 2004). We decided to avoid reactive emotions such as surprise or disgust because they are associated with specific movements and are difficult to perform (Konijn, 2000). Finally, anger and happiness are two of the most studied emotions in the broad literature of examining perception of faces, voices and body movement. This broad literature provides a good starting baseline to explore the topic further in the direction of perception of body movement and voice in angry and happy social interactions.

The remaining paragraphs of this chapter describe the methods used to set up, calibrate and capture interactions, the scripts and scenarios used for the captures, and the procedure for post-processing the captured data with the creation of stimuli.

2.2 METHODS

2.2.1 *Actor selection*

A group of 20 actors was selected and combined into ten pairs - five experienced and five non-experienced. The mean duration of acting experience for experienced actors was 9.68 years, ranging from 5 to 25 years. All actors were English-speaking, UK-born males, with a mean age of 26.12 years, ranging from 17 to 43 years. Pairs of actors were simultaneously participating in each session, they knew each other well and were paid for their time. Before each session actors were briefed on the purpose of the study and signed a consent form.

A graphical overview of all stages of stimuli preparation, capture and post-processing are shown on Fig. 6

2.2.2 *Motion and voice capture setup and calibration*

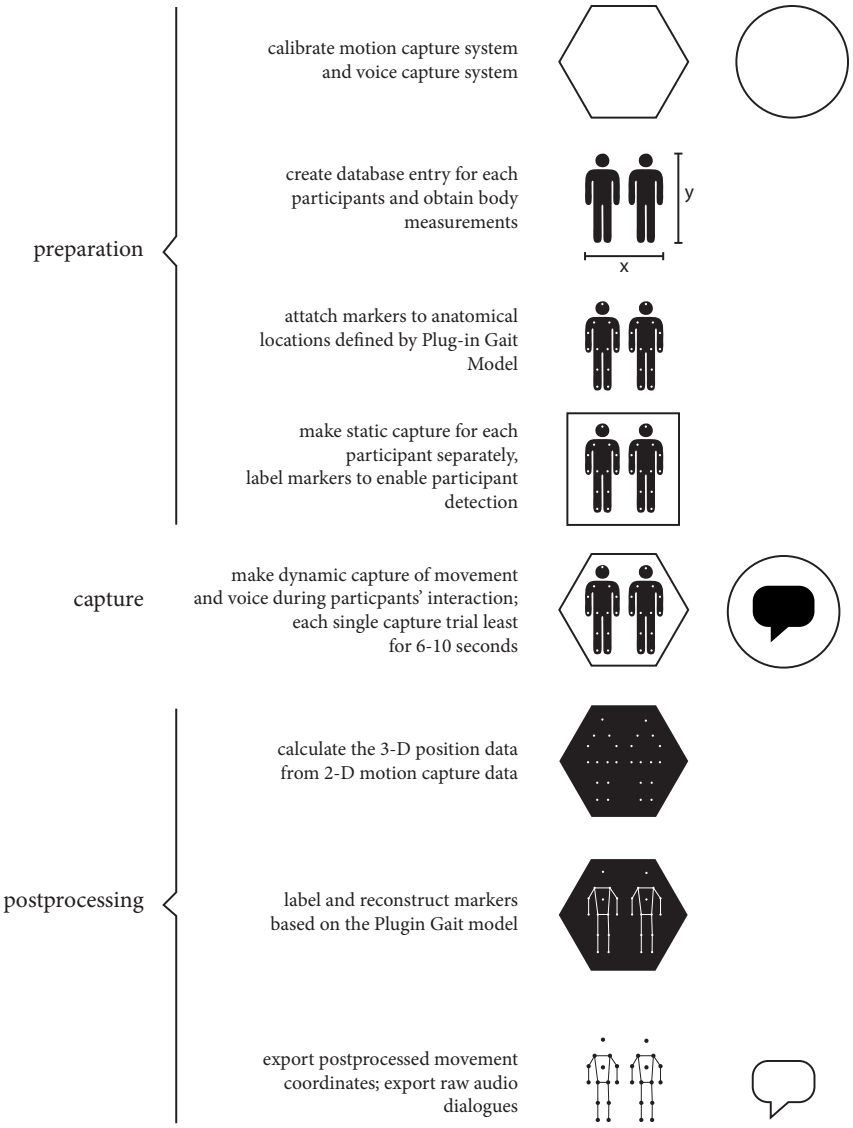


Figure 6: Graphical overview of stimuli creation stages of preparation, capture and post-processing.

2.2.2.1 *Technical specification of the motion and voice capture system*

Motion capture took place at the University of Glasgow in the School of Psychology using 12 Vicon MXF40 cameras (Vicon, 2010) that offer online monitoring of 3D motion signals. At all times, the system was recording at a rate of 120 frames per second (fps). The audio capture was done using a custom-upgraded Vicon Analogue Card

(Vicon, 2010) connected to AKG D7S Supercardioid Dynamic Microphone, and it was recording with 44.1kHz and 24-bit sampling rate. The choice of a dynamic microphone was justified by the absence of any sound proofing treatment in the motion capture room. The AKG D7S has a frequency response optimised for vocal use and additional features to reduce different types of noise: mechanical/pneumatic transducer shock mount that reduces handling and cable noise, and a high pass filter and humbucking coil to minimise low-frequency noise. During the recording, the audio capture was fully synchronized with the motion capture via the Vicon Analogue Card. The entire capture setup, including floor measurements, the location of cameras, microphone and actors, is illustrated on Figure 7. All Vicon MXF40 (Vicon, 2010) cameras and Vicon Analogue Card (Vicon, 2010) with AKG microphone were connected to two Vicon MX Ultranet HD units (Vicon, 2010) that both powered up the cameras and preprocessed some of the captured data. Ultranet HD units (Vicon, 2010) were connected via ethernet to a single PC computer. The PC was specifically configured to handle large data streams from the capture system and it was pre-installed with Vicon Nexus 1.3, Vicon Bodybuilder (Vicon, 2012) and MATLAB 2010 (Mathworks, 2010) software. Vicon Nexus 1.3 (Vicon, 2010) was used for most of the capture operations including calibration, capturing, storage, and post-processing of raw capture data.

2.2.2.2 *Preparing actors and capture sessions*

After calibration of the motion capture system, each capture session started with taking actors' measurements (Table A.1 in Appendix) and placing 39 retroreflective, 14mm, spherical markers on specific anatomic locations on their bodies. Those anatomical locations were defined by the Plug-in Gait Model (Figure 8) and are described in Table A.2 in the Appendix. Plug-in Gait is based on the widely accepted Newington-Helen Hayes gait model and it uses a defined marker set and a set of subject measurements to create outputs of the joint kinematics and kinetics for a gait analysis patient (Kadaba *et al.*, 1990; Davis *et al.*, 1991).

During the capture session actors were positioned, one facing the other, at a distance specified by a marked position on the floor, approximately 1.3 meters (Figure 9a). This interpersonal distance varied between 1 - 1.6 meters and it flexibly changed during the capture trials, depending on how much actors moved when interacting. However, at the beginning of each single capture trial actors were asked to come back to the start position marked on the floor. The overall

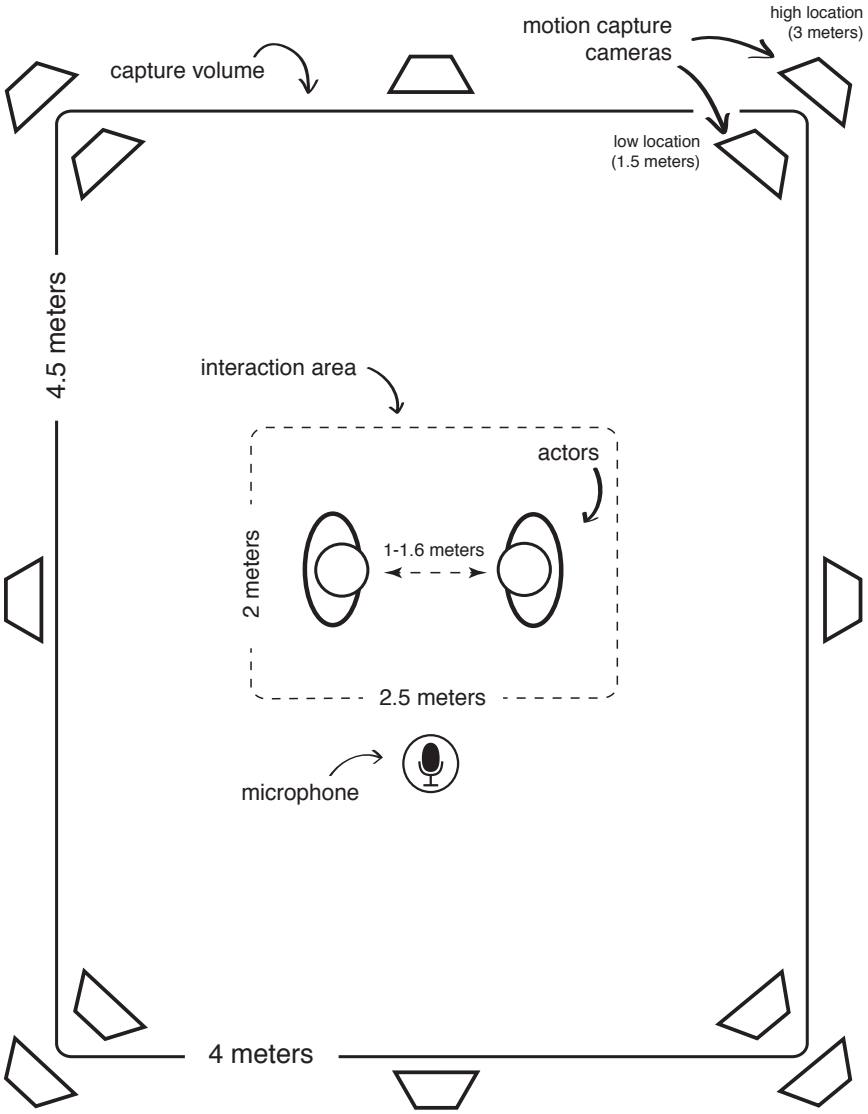


Figure 7: Motion capture room - cameras and microphone setup and capture areas (schematic view from the top).

space of interaction was limited to around 3 x 2 meters, but since the participants were within the comfortable personal space as defined by Hall (1966), we expected that the influence of proxemics[2] did not affect their natural interaction.

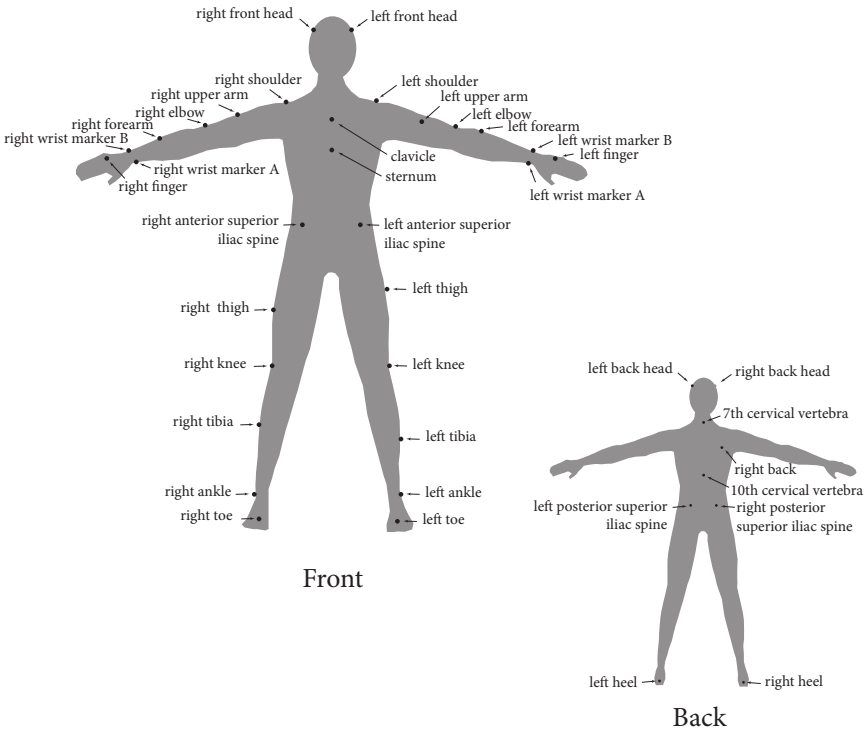


Figure 8: Marker placement for the Plug-in Gait Model (see Figure A.2 in the Appendix for exact anatomical locations of markers).

We captured three emotional interactions: angry, happy and neutral. Angry and happy interactions were captured at three different intensity levels: low, medium and high. To help actors convey angry and happy interactions at different levels of intensity, they were given short and simple scenarios of the emotional situations and asked to imagine themselves in those situations. Table 1 describes the exact scenarios given to actors. The order of given scenarios and order of emotional interactions conveyed was randomized for each pair. Actors were also instructed to recall their own past situations associated with the relevant emotional scenario to help them induce the emotion. The hypothetical scenarios were based on simple common

2 The notion of personal space was introduced by Hall (1966), who created the concept of proxemics. Proxemics can be defined as the interrelated observations and theories of man's use of space as a specialized elaboration of culture (Hall, 1966). Hall (1966) describes the subjective dimensions that surround each person and the physical distances they try to keep from other people, according to subtle cultural rules.

situations (Scherer, 1986). Neutral served as a control condition, and actors were asked to interact in a neutral, emotion-less manner. Actors were given relative freedom in expressing the emotions during interactions (Rose & Clarke, 2009). They were encouraged to act naturally, but they were instructed to avoid touching each other and we were careful to give them only verbal instructions rather than performing actions ourselves (Clarke *et al.*, 2005; Ma *et al.*, 2006; Roether *et al.*, 2009). People typically use touch to share their feelings with others, and to enhance the meaning of other forms of verbal and nonverbal communication (Gallace & Spence, 2010). Touch also appears very early in human development and naturally becomes on its own a powerful indicator of affect (Harlow, 1958).

In terms of the content of verbal exchange, actors were asked to interact exchanging one of two dialogues in each single trial, as seen on table 2. We purposely chose two formats of dialogue, one being inquiry (question and answer) and another being deliberation (two affirmative sentences), as specified in Krabbe & Walton (1995). We wanted to check whether those different formats of dialogue influenced the identification of emotional interaction between the actors by the observers (Craggs & Wood, 2003). We also picked relatively neutral words to compose dialogues, so that dialogues were easy to articulate in either a happy or angry emotional manner.

A single capture trial lasted no longer then 6-10 seconds. In each trial, the recording started around 1 second before actors were given a signal to begin the interaction. To signal the start of each capture trial, a 1-second long, digital square wave sound was played (a simple beep). Recording stopped around 2-3 seconds after actors stopped their interaction. For each pair of actors we completed 10 practice trials before the capture trials. Practice trials were conducted to give actors more time to prepare, adjust to their roles and for us to check if the motion and voice capture system had been calibrated correctly. Immediately after the practice trials we initiated the capture trials during which we collected the material used for creating the stimulus set. For each pair we obtained 84 capture trials. This consisted of 2 emotions (happy, angry) x 3 intensities (low, medium, high) x 2 dialogue versions (question/answer, affirmative/affirmative) x 2 actors order x 3 repetitions + 12 neutral conditions (2 dialogue versions * 6 repetitions of each action). This resulted in a total of 840 film clips for all 10 couples, but we had to remove one couple from further processing due to a high noise level in the motion and voice capture data. There were another 100 data trials from 10 practice captures for

Emotion	Intensity	Scenario
Angry	Low	You have just discovered that you've been charged extra for one product that was clearly on promotional offer.
		You have been in a rush for an important meeting, and sudden rain has completely soaked you.
	Medium	Someone just stepped on your toe.
		Someone cut in line, when you've been waiting for over 20 minutes.
	High	You have just discovered that your wallet has been stolen.
		Your parents or partner just got into a big argument with you about something silly.
Happy	Low	You wake up on a Saturday after a number of wintry-cold rainy days, and the temperature is around 20 degrees.
		You have just finished this report you've been working on for two weeks.
	Medium	You unexpectedly run into someone you like very much and haven't seen in a long time.
		You got an amazing bargain in the local retail store sale.
	High	You buy a lottery ticket and you win £10,000 instantly.
		Its your birthday and your friends got you unexpected and awesome gift.
Neutral	NA	You think about the situation when you talked about something neutral and casual, like your typical day at work or your typical journey in a bus.

Table 1: Scenarios given to actors during emotional and neutral interactions.

	Dialogue A - Inquiry	Dialogue B - Deliberation
Actor 1	Where have you been?	I want to meet with John.
Actor 2	I have just met with John.	I will speak to him tomorrow.

Table 2: Two dialogue versions used during the capture trials.

each couple, but these practice captures were excluded from further post-processing.

2.2.2.3 *Post-processing procedure*

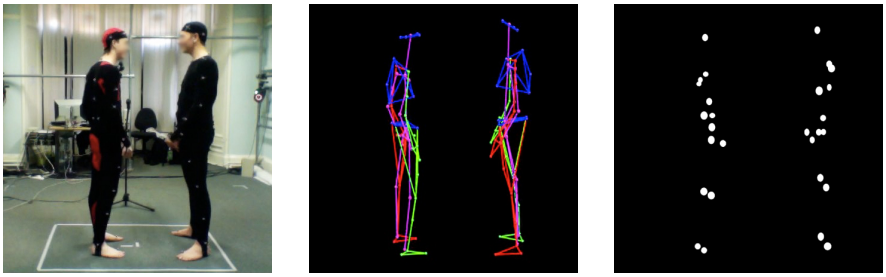
There were five main stages of post-processing:

- calculating the 3-D position data from 2-D camera data,
- labelling of the reconstructed markers based on the Plug-in Gait model,
- interpolating missing data points,
- exporting raw coordinates and creating point-light displays in MATLAB,
- exporting raw audio dialogues to processing them in Adobe Audition.

The first three operations were executed automatically in Vicon Nexus 1.3 (Vicon, 2010). Creating final point-light displays required a few additional steps. From the trajectories of the 39 original markers, we computed the location of 'virtual' markers positioned at major joints of the body. The 15 virtual markers used for all the subsequent computations were located at the joints of the ankles, the knees, the hips, the wrists, the elbows, the shoulders, at the centre of the pelvis, on the sternum, and in the centre of the head (Figure 9b). Commercially available software Vicon BodyBuilder (Vicon, 2010) for biomechanical modelling was used to achieve the respective computations. A similar approach has been used by Dekeyser *et al.* (2002), Troje (2002) and Ma *et al.* (2006). The advantage of this procedure was that it was a quick and automated way of creating the virtual joint centres for both actors without the need of manual adjustments (Dekeyser *et al.*, 2002; Ma *et al.*, 2006).

After attaching virtual markers, the 3D (x, y, z) position coordinates for those markers were exported from Vicon Nexus 1.3 (Vicon, 2010) as a tab-delimited text file. Those coordinate files were formatted in a way that position co-ordinates were represented in columns, while each frame of data record was represented in rows. Co-ordinate files were imported into Matlab (Mathworks, 2010) and an algorithm was applied to generate the final point-light displays. The algorithm was based on the one used by Pollick *et al.* (2001). The algorithm converted 15 virtual markers from each actor into point-light displays, generated as white dots on a black background from the side view, as seen in Figure 9c. The algorithm exported point-light displays in

the Audio Video Interleave (AVI) format, with a frame size of 800 by 600 pixels. The frame rate of exported displays was reduced from the original 120 fps to 60 fps. Both Matlab (Mathworks, 2010), and Adobe Premiere 1.5 (Adobe Systems, 2004) were used for creating final displays, only allowing editing of the movie up to 60 fps, and therefore compression from 120 fps was necessary.



(a) Photo view from capture trial. (b) Post processed view from Nexus. (c) Dyadic point-light displays view.

Figure 9: Post processing pipeline.

The audio dialogues recorded with the Vicon Analogue Card were all saved by Vicon Nexus 1.3 (Vicon, 2010) in the Audio Interchange File Format (AIFF), and each audio dialogue was automatically linked with the corresponding capture trial. Adobe Audition 3 (Adobe Systems, 2007) was used to post process the dialogues. Every audio dialogue was first amplified by 10dB and then a noise reduction was applied. Following this all audio dialogues were normalised to create a consistent level of amplitude, and to obtain the average volume of around 60dB. Finally, each audio dialogue was exported as a Waveform Audio File Format (WAV) file with a resolution of 44.1kHz and 24-bit sampling rate.

2.2.2.4 *Creation of final stimuli*

Adobe Premiere 1.5 (Adobe Systems, 2004) was used to create a final stimulus set. The AVI point-light displays produced by MATLAB 2010 (Mathworks, 2010) were imported to Adobe Premier 1.5 (Adobe Systems, 2004) together with the corresponding WAV dialogues post processed with Adobe Audition 3 (Adobe Systems, 2007). Initially, each point-light display was combined with its corresponding WAV dialogue. The start point was defined immediately after the sound signaling the start of the recording for actors (1-second long, square-wave buzzer signal described in chapter 2.2.2.2). The end point was specified by the relative end of the dialogue within 1 second after the dialogue finished. The length of the final, truncated display varied

between 2.5 and 3.5 seconds. All displays with truncated start/end points were exported to AVI format in three versions: auditory-only (dialogues), visual-only (point-light displays) and audio-visual (dialogues combined with point-light displays. Figure 10 shows an example sequence of point-light displays with dialogue. The final, non-validated stimulus set was composed of 252 unique displays that consisted 9 actor couples x 2 emotions (happy and angry) x 3 intensities (low, medium, high) x 2 dialogue versions (inquiry, deliberation) x 2 repetitions plus 36 neutral displays. However, each display was created in three modality formats: visual point-light displays, auditory dialogues and a combination of point-light displays and dialogues. Therefore the final count of all displays in stimulus set with three modality formats was 756.

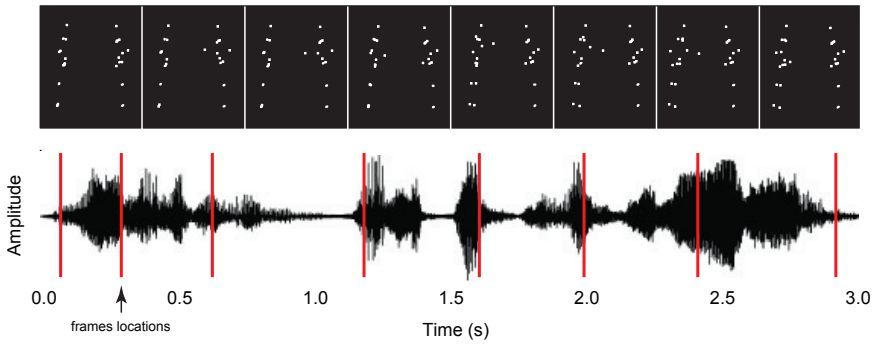


Figure 10: Example time series of dyadic point-light displays and dialogue showing eight frames with interaction between two point-light actors, together with oscillogram of dialogue voice amplitude

2.2.3 *High-level description of actors' portrayal of emotions*

With a large number of captures obtained for our stimuli set, there was a significant variance in how actors conveyed and portrayed the intended emotions. At the same time, it was intuitively clear that some similarities existed in the characteristics of body movement and voice when actors engaged in emotional dialogues at different levels of intensity. Two common approaches are used to analyze such similar characteristics of movement and voice - either a low-level approach (i.e. calculate kinematics of movement or prosodic features of voice) or a high-level approach (i.e. qualitatively describe the movement and voice). In our scenario, the use of a low-level approach was impractical, mainly due to the complexity of our stimuli set that incorporated mutisensory dyadic interactions. We therefore decided to provide a brief qualitative and descriptive highlight of movement and

voice by examining all 252 unique displays of interactions. As a baseline to describe the features of body movement, we used Kleinsmith & Bianchi-Berthouze (2013) review and for a voice description baseline we used studies by Scherer (2003) and Sauter *et al.* (2010). While the variance of such features was very high, we provide some examples in Table 3. It is important to note that the majority of features match a number of examples provided in the papers of Kleinsmith & Bianchi-Berthouze (2013) and Scherer (2003) although intensity differences and the dyadic context of interaction provided an interesting level of differentiation. In angry interactions, the best existing correspondence was Wallbott (1998) description of "cold anger" for low and medium intensity (lateralized hand/arm movement, slow and tense leaning of upper body), and "hot anger" for high intensity (shoulders lifted, arms stretched out frontal, fast arm movement, fast movement of upper body, rapid step forward). For happy interactions, low intensity corresponded to what was described as "content" by Gross *et al.* (2010) (expanded limbs and torso, muscles relaxed), medium intensity as "joy" or "elated joy" by Gross *et al.* (2010) & Wallbott (1998) (straight trunk and legs, shoulders lifted, head bent backwards), and high intensity as "sympathy" or "triumphant" by Kleinsmith *et al.* (2011) & Meijer (1989) (arms raised and extended laterally, stepping forward, arms opened frontally). Regarding voice features, angry interactions correspond to Scherer (2003) descriptions with high intensity, high frequency and high rate of speech articulation, happy with high intensity but slower articulation rate, and neutral with passive voice, with low level of contours, low volume and low level of variability.

2.3 DISCUSSION

We have described our method for the production of the first data set that can be used for the study of audio-visual integration from emotional social interactions. Using a passive optical motion capture system, synchronized with audio capture, we recorded 840 interactions between ten different pairs of actors. Captured movement and conversations were converted into formats useful for animation as point-light displays combined with voices. The final stimulus set consists of 256 unique clips that present happy, angry and neutral emotional interactions with low, medium and high levels of emotional intensity.

There are three main features that make our stimulus set particularly suitable for the study of audio-visual integration in the social

Emotion	Intensity	Movement features	Voice features
Angry	Low	Small upper-body/head leaning	Slow speech rate with no accenting
		Small pointing or shaking gesture with one or two arms	Fast speech rate with no accenting
		Small upper body backward bend	Low volume of speech
	Medium	Tense twist of body position from front facing to side facing	Fast speech rate with accenting
		Position freeze, shoulder jacking up	Higher volume of speech
		Small upper-body/head leaning	
	High	Shoulders lifted, arms stretched out frontal	Strong accenting with pasues
		Fast movement of upper body, rapid step forward	Loud speech
		fast upward pointing or shaking gesture with one or both arms	
Happy	Low	Expanded limbs and torso	Steady speech rate with accenting
		Small upward pointing or shaking gesture with one arm	Low volume of speech
		Relaxed small scale hand movements	
	Medium	Straight trunk & legs	Steady speech rate with accenting
		Shoulders lifted, head bend backwards	Normal volume of speech
		Stepping from leg to leg in place	
	High	Arms raised and extended laterally	Fast speech rate with accenting
		Repetetive, upward, "jumpy" movement on toes	Normal volume or loud speech
		Relaxed hip movement backwards and forward	
Neutral	NA	Low level of movement or relaxed and slow movements of hands and arms, steady posture	Low volume, passive, steady speech with no accenting

Table 3: High-level qualitative examples of specific body movement and voice features that were most commonly observed amongst actors as a mean to act angry, happy and neutral emotions on low, medium and high level of intensity.

context. First, we captured both movement and voice in a synchronised manner and therefore provide the first data set to study audio-visual emotional interactions. stimulus sets with point-light dyadic interaction have been created before (Clarke *et al.*, 2005; Manera *et al.*, 2011) but none have combined point-light displays with dialogue. The stimulus set has been designed to study point-light display and voice as combined stimuli or separate stimuli if such an approach is required. The stimuli also consist of entire body movements in contrast to other existing stimulus sets which include only sections of body (e.g. only faces and hands in Busso, 2008).

Second, we simplified the design of the emotional component of the stimulus set by using only happy and angry emotional interactions. Existing stimulus sets include a broader spectrum of emotional expressions (e.g. citealpBusso2008a) but many of those expressions were difficult to validate considering the ambiguous and reactive nature of some emotions, such as fear or disgust (Ma *et al.*, 2006; Roether *et al.*, 2009). Differences between the perception of happy and angry interactions have been widely reported in the neuroimaging and multisensory literature (Massaro & Egan, 1996; Fox *et al.*, 2000; Johnstone *et al.*, 2006; Ikeda & Watanabe, 2009; Stienen *et al.*, 2011). From the practical perspective, it was also easier for our actors to perform happy and angry interactions and to create scenarios to help them act these emotions. This way, we avoided creating unrealistic stimuli and provided a better variety within the emotional interactions. To increase variability within the emotional expressions of anger and happiness, we captured interactions at three different levels of intensity. During the capture of the stimulus set, we used realistic scenarios and role-plays to make the set more ecologically valid and to help the actors to engage in the scene in a more realistic way (Risko *et al.*, 2012). Each single interaction consists of the action of one actor and the response of the other and each interaction lasts for about three seconds. We also used a mix of experienced and inexperienced actors during the recording to increase the variety that can arise from any acting strategies that people use to express emotions (Ma *et al.*, 2006).

Third, we can easily customize different parameters of point-lights because of the universal format in which it has been created. Actors' motions can be analyzed, extracted and manipulated. Parameters such as orientation, speed or size of point-lights can easily be changed. Audio dialogue has also been normalized and extracted in a widely available format so it can be easily manipulated and analyzed for any speech-related parameters. The compatibility of formats available for the stimulus set gives researchers the option to drive 3D

animation in case they want to look at the effects of forms other than point-light displays. Overall, the stimulus set developed in this thesis presents on its own a simple, customizable and compact tool to explore both unimodal and mutlimodal aspects of emotional social interactions.

3

VALIDATING STIMULI TO STUDY THE PERCEPTION OF EMOTIONAL SOCIAL INTERACTIONS

3.1 INTRODUCTION

Chapter 2 outlined the procedure for creating a stimulus set for the study of emotional social interactions. This chapter describes the experiment conducted to validate the stimulus set. The main goal was to examine how accurately the emotional interactions were identified by observers when presented with the displays as point-lights (visual group), voice dialogues (auditory group) or a combination of point-lights and dialogues (audio-visual group). The reason for using a between-subject design was to avoid audio-visual facilitation, or carry-over effects, that could impact emotional identification when visual, auditory and audio-visual displays are presented together in one set. Audio-visual facilitation has been shown in studies using emotional faces and voices, when audio-visual conditions enhanced perceived emotion with respect to auditory-only and visual-only conditions (Vines *et al.*, 2006; Collignon *et al.*, 2008). We also wanted to restrict presentation of every display to a single occasion to avoid the practice effects that can occur when participants see a repetition of a specific display (Heiman, 2002). We will refer to the experiments described in this chapter as 'validation experiments'.

A number of existing studies show that observers can identify a range of emotions from the point-light displays of single actors (Dittrich *et al.*, 1996; Pollick *et al.*, 2001; Atkinson *et al.*, 2004) or the interactions between two actors (Dittrich, 1993; Clarke *et al.*, 2005; Lorey *et al.*, 2012). In the case of identifying emotions from voice dialogue, a review of approximately 30 studies conducted up to the early 1980s also revealed that observers can accurately identify emotions from voice (Scherer, 1989), and recent studies confirmed these results (Scherer, 2003). However, there has been no research to assess the ability of observers to identify emotion from interpersonal actions from body movement AND voice. Spoken dialogue and body movement are closely interconnected and for two people to be engaged in discourse is a very natural situation encountered on a daily basis (Bull, 1990; McClave, 2000; Clarke *et al.*, 2005).

As well as examining how easily participants recognize happy and angry emotional interactions from point-light displays and dialogues, a number of supplementary questions were investigated. We wanted to establish whether emotional intensity has any effect on the accuracy of identifying happy and angry displays. Intuitively, less emotionally intense displays would be expected to be more ambiguous for the observers and evoke less accurate judgements. In studies using single actors, both body movement and voice have been reported to provide important information about emotional intensity (Harrigan, 2005; Sevdalis & Keller, 2011; Dael *et al.*, 2012). Our goal was to examine how varying emotional intensity influences both accuracy of, and confidence in, emotional judgements in the context of social interactions.

We also wanted to determine if there was a difference in how accurately displays by experienced actors were judged when compared to displays by inexperienced actors. As mentioned in Section 2.2.1 of Chapter 2, there were conflicting findings regarding the professional experience of actors used in creating emotional stimulus sets (Ma *et al.*, 2006; Busso, 2008; Roether *et al.*, 2009). We wanted to better understand if similar differences in perception of experienced and inexperienced actors are indeed present in our stimulus set. We also investigated whether there was any effect on the accuracy of emotional judgements due to the dialogue type (i.e. enquiry versus deliberation - see Table 2 in Chapter 2) on the accuracy of emotional judgements.

Finally, we also included neutral displays in all three experimental groups. As we discussed in Section 2.2.3 of Chapter 2, the neutral portrayal of actions was relatively passive and very low in body movement with unemotional, low-frequency, contour-less speech. Hypothetically, we expected that all neutral displays should receive a relatively equal number of happy and angry judgements unless participants were biased towards one of the emotions, or there were some specific features in the movement and voice of the neutral displays that resembled features in happy and angry interactions. While we included neutral displays, we did not explicitly ask participants to identify neutral interactions. Neutral displays served as a form of "control condition" so participants only made judgements of happiness and anger.

In all three experimental groups (visual, auditory, audio-visual), we introduced a short debriefing questionnaire. Among other questions, participants were asked how many actors were used to create the displays they saw or heard during the experiment. This question directly required participants to think about the movies they saw and to re-

call how many different actors they witnessed, giving us some hints regarding the detection of identity of the actors. A number of studies have shown that people can recognize the identity of a person from the point-light displays of a single actor (Cutting & Kozlowski, 1977; Hill & Pollick, 2000; Ma *et al.*, 2006) and their voice (Belin *et al.*, 2004; Campanella & Belin, 2007). We wanted to examine how the perception of identity compares when observing body movement and hearing voice when participants view social interactions. We also asked participants various questions about their perception of their own performance in the task and to describe the strategies they used when making judgements of emotions as well as any other comments they had.

This chapter describes the methods and results for all three experimental groups (visual, auditory, audio-visual). The description of the validation experiment begins by highlighting methods of conducting these three between-subject experiments as they share many common points in design, procedure and stimuli. The main results for each group are described separately with analysis of variance and post hoc Tukey analysis being conducted on the number of correct judgements and confidence ratings. An analysis of supplementary results follows including analysis of factors such as the actors' experience, dialogue effect, and questionnaire results. Finally, the results are discussed in the context of the validity of the stimulus set for further investigation on the perception of emotions from body movement and voice in social interactions.

3.2 METHODS

3.2.1 *Participants*

We separately recruited a total of 43 participants for three independent groups:

- visual group (15 participants, 7 of them male, with a mean age of 25.8 years, ranging from 17 to 45 years),
- auditory group (13 participants, 6 of them male, with a mean age of 20.5 years, ranging from 17 to 26 years),
- audio-visual group (15 participants, 8 of them male, with a mean age of 22.5 years, ranging from 18 to 37 years).

All participants were English-speaking, UK-born, and they all reported normal hearing and normal or corrected-to-normal vision. All the participants were naive to the purpose of the study and they lacked any

prior experience with point-light display movies or images. The study received ethical approval from the University of Glasgow's Faculty of Information and Mathematical Sciences Ethics Review Board. Every participant signed a consent form.

3.2.2 *Stimuli, Design and Procedure*

We used the stimulus set described in chapter 2, composed of happy, angry and neutral dyadic interactions presented as point-light displays (visual group), voice dialogues (auditory group) and a combination of point-light displays and voice dialogues (audio-visual group). In summary, the stimulus set was composed of 252 unique displays that consisted of 9 actor couples, 2 emotions (happy and angry), 3 intensities (low, medium, high), 2 dialogue versions (inquiry, deliberation), 2 display versions plus 36 neutral displays (i.e. 9 actor couples, 2 dialogue versions, 2 display versions).

The task was exactly the same for all three groups. After being presented with the display, participants were given two questions. First, participants were asked to identify whether interaction was happy or angry. They did so by choosing red **H** for happy or red **A** for angry on the keyboard. Immediately after a response, the second screen was presented. In this second screen participants were asked how confident they were about their choice of emotion on a rating scale from *1* to *9*, where *1* referred to *very low confidence* and *9* referred to *very high*. Each display was presented only once and the order of all displays was randomized. We used Neurobehavioral Presentation 13.1 software (Neurobehavioral Systems, 2008) to present the displays and collect the responses[1]. At the end of experiment, each participant was asked to complete a short questionnaire with the following items:

- how many trials do you think you got right (give percentage)?
- how many actors do you think we used to create the displays you have seen/heard (give single number)?
- did you use any specific strategies to do the task (use no more then five keywords to describe it)?
- do you have any other observations you would like to mention?

For the majority of data analysis we used repeated measure analysis of variance (ANOVA). We also calculated generalized *eta squared*

1 Due to technical difficulties (i.e. undetected bug in the Neurobehavioral Presentation code) we failed to log reaction times for all three experimental groups.

(η^2_G) measures of effect size. η^2_G is preferred to *eta squared* and partial *eta squared* because it provides comparability across between-subjects and within-subjects designs, and it can easily be computed from information provided by standard statistical packages (Olejnik & Algina, 2003; Bakeman, 2005). All p-values from posthoc Tukey analysis are presented after adjustment for multiple comparison.

3.3 RESULTS FOR VISUAL GROUP

Identification accuracy increased with intensity for angry but not happy visual displays

The mean number of correct responses were analysed by carrying out a within-subjects ANOVA with 'emotion' (happy and angry) and 'intensity' (low, medium, high) as within factors. There was a significant main effect of factor 'emotion' ($F(1,14) = 17.91$, $p < 0.001$, $\eta^2_G = 0.37$), and factor 'intensity' ($F(2,28) = 41.87$, $p < 0.001$, $\eta^2_G = 0.27$) was also found significant, indicating that participants' accuracy of judgements differed between emotional displays and levels of intensity. A significant interaction between factors 'emotion' and 'intensity' was also found ($F(2,28) = 39.89$, $p < 0.001$, $\eta^2_G = 0.43$) and Figure 11 shows that there was a difference in how accurately happy and angry displays were judged on different levels of intensity. Indeed, Posthoc Tukey analysis showed that accuracy of judgements for angry displays increased significantly between each level of intensity ($p < 0.001$), as seen on Figure 11. But there was no difference in accuracy of judgements between any level of intensity for happy displays (low and medium: $p = 0.40$, low and high: $p = 0.27$, medium and high: $p = 1.00$). When comparing happy and angry displays on different levels of intensity, there was no significant difference in accuracy of judgements between high ($p = 0.08$) intensity happy and angry displays, however low and medium intensity happy displays were judged more accurately than corresponding low and medium intensity angry displays ($p < 0.01$).

Low-intensity angry displays judged below level of chance in visual group

Figure 11 shows that low-intensity angry displays were judged below the level of chance (0.5). We hypothesized that this effect was either due to differences in actors' depictions of emotions or due to differences in some features of movement in low intensity angry displays that were more ambiguous to the observers. When comparing mean accuracy of emotion judgements for low, medium and high intensity displays with specific actor couples[2], it was clear that some couples were given more accurate judgements than others, and this effect was more profound for angry then for happy displays, as seen

2 Accuracy of judgements by actor is showed for auditory group (Figure A.1 and A.2) and audio-visual group (Figure A.3 and A.4) in the Appendix.

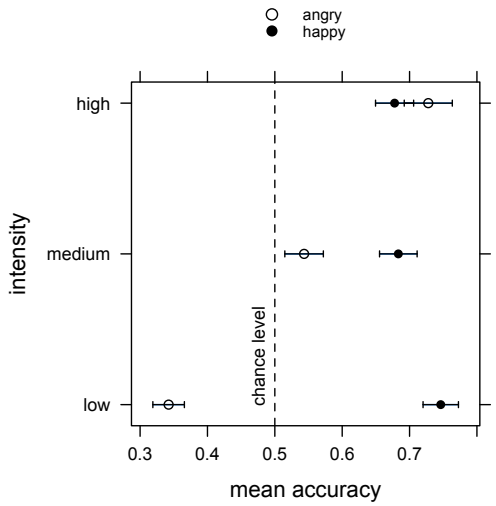
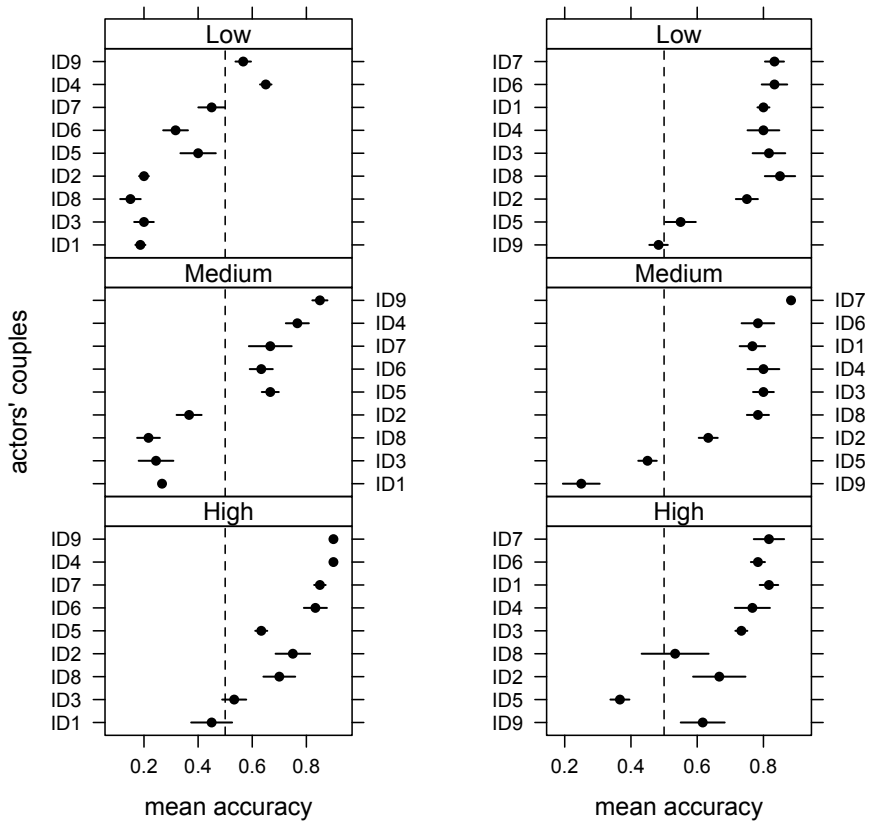


Figure 11: Mean accuracy of emotion judgments for happy and angry displays at low, medium and high intensity in the visual experiment. The error bars represent one standard error of the mean.

on Figure 12. Therefore, one potential explanation for the low accuracy of judgement for low intensity angry displays was the variance in the number of correct judgements between different actor couples contributing to an overall low number of accurate judgements for low intensity angry displays (Figure 12). Another explanation was that some features of movement in displays with low intensity anger were too similar to corresponding features in happy displays, and this qualitative point is further elaborated upon in the Discussion section of this chapter.

The mean confidence ratings were analysed by carrying out a within-subjects ANOVA with 'emotion' (happy and angry) and 'intensity' (low, medium, high) as within factors. The within factor 'intensity' was found to be significant ($F(2,28) = 25.26$, $p < 0.001$, $\eta^2_G = 0.11$), indicating that participants' confidence ratings differed between levels of intensity (Figure 13). There was also a significant interaction between factors 'emotion' and 'intensity' ($F(2,28) = 6.21$, $p \leqslant 0.01$, $\eta^2_G = 0.01$) indicating that confidence ratings for some emotional displays were more affected by the change in the level of intensity. Indeed, posthoc Tukey analysis revealed that confidence ratings increased significantly between low and high intensity angry displays ($p \leqslant 0.04$), and low and high intensity happy displays ($p \leqslant 0.05$). There was no significant effect of factor 'emotion' ($F(1,14) = 3.20$, $p = 0.10$, $\eta^2_G = 0$) and Figure 13 shows that participants gave similar confidence ratings for both happy and angry displays.



(a) Angry displays.

(b) Happy displays.

Figure 12: Mean accuracy of emotion judgments for (a) angry and (b) happy displays at low, medium and high intensity levels with specific actors' couples (ID1-ID9) in visual experiment. The error bars represent one standard error of the mean and the dashed line represents the level of chance (0.5).

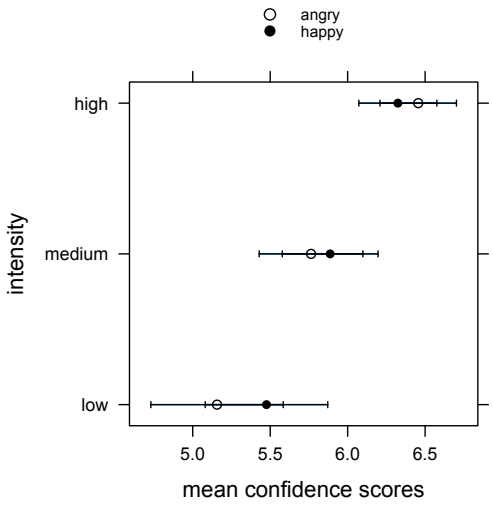


Figure 13: Mean confidence ratings for happy and angry displays at low, medium and high intensity in visual experiment. The error bars represent one standard error of the mean.

3.4 RESULTS FOR AUDITORY GROUP

Identification accuracy increased with intensity for angry but not happy auditory displays

The mean number of correct responses were analysed by carrying out a within-subjects ANOVA with 'emotion' (happy and angry) and 'intensity' (low, medium, high) as within factors. There was no significant main effect of factor 'emotion' ($F(1,12) = 4.26$, $p = 0.06$, $\eta^2_G = 0.12$), indicating that happy and angry displays were judged with similar accuracy (Figure 14). There was a significant main effect of 'intensity' ($F(2,24) = 28.71$, $p < 0.001$, $\eta^2_G = 0.22$) indicating that participants' accuracy of judgements differed between emotional displays and levels of intensity (Figure 14). A significant interaction between the factors 'emotion' and 'intensity' was also found ($F(2,24) = 11.18$, $p < 0.001$, $\eta^2_G = 0.23$) and Figure 14 shows that accuracy judgements for angry displays were more affected by the change in the level of intensity then for happy displays. Indeed, posthoc Tukey analysis revealed that accuracy of judgements for angry displays increased significantly between each level of intensity ($p < 0.05$), but there was no difference in accuracy of judgements between any level of intensity for happy displays (low and medium: $p = 0.41$, low and high: $p = 1$, medium and high: $p = 0.30$). Comparing happy and angry displays on different levels of intensity, there was no significant difference in accuracy of judgements between high intensity happy and angry displays ($p = 0.41$). However, low and medium intensity happy displays were judged more accurately then low and medium intensity angry displays ($p < 0.05$).

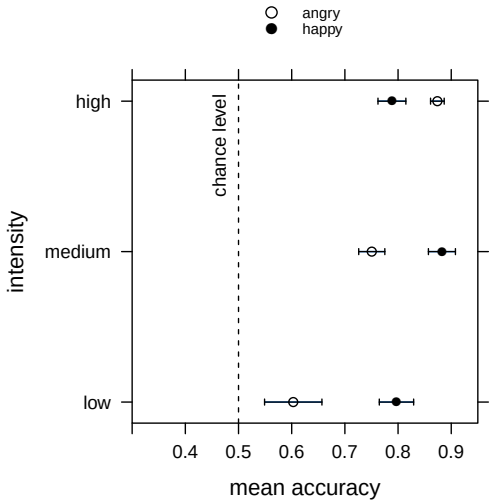


Figure 14: Mean accuracy of emotion judgments for happy and angry displays at low, medium and high intensity in auditory experiment. The error bars represent one standard error of the mean.

The mean confidence ratings were analysed by carrying out a within-subjects ANOVA with 'emotion' (happy and angry) and 'intensity' (low, medium, high) as within factors. No main effect of factor 'emotion' ($F(1,12) = 1.48$, $p = 0.25$, $\eta^2_G = 0.01$) was found, and Figure 15 shows that participants rated happy and angry displays with similar confidence. The within factor 'intensity' was found to be significant ($F(2,24) = 67.63$, $p < 0.001$, $\eta^2_G = 0.46$), indicating that participants' confidence ratings differed between levels of intensity (Figure 15). There was also a significant interaction between factors 'emotion' and 'intensity' ($F(2,24) = 14.79$, $p < 0.001$, $\eta^2_G = 0.09$) indicating that confidence ratings for some emotional displays were more affected by the change in the level of intensity. Indeed, posthoc Tukey analysis revealed that confidence ratings for angry displays increased significantly between each intensity level ($p < 0.05$). Confidence ratings also increased significantly between low and high intensity happy displays ($p < 0.05$), but there was no difference between low and medium ($p = 0.19$), and medium and high intensity happy displays ($p = 0.77$). There was no significant difference in confidence ratings when comparing corresponding intensities of happy and angry displays (angry low and happy low: $p = 0.22$, angry medium and happy medium: $p = 0.93$, angry high and happy high: $p = 0.66$).

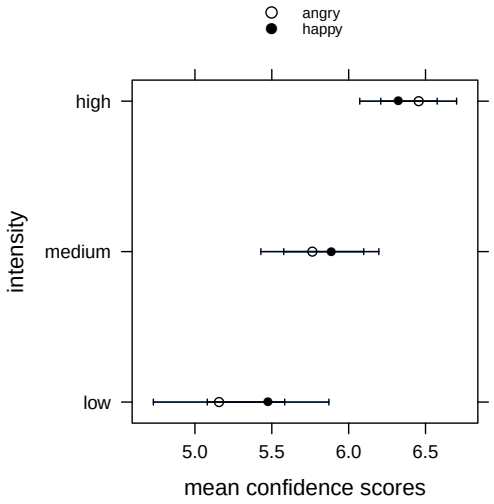


Figure 15: Mean confidence ratings for happy and angry displays at low, medium and high intensity in auditory experiment. The error bars represent one standard error of the mean.

3.5 RESULTS FOR AUDIO-VISUAL GROUP

Identification accuracy increased with intensity for angry but not happy audio-visual displays

The mean number of correct responses were analysed by carrying out a within-subjects ANOVA with 'emotion' (happy and angry) and 'intensity' (low, medium, high) as within factors. There was a significant main effect of factor 'emotion' ($F(1,14) = 6.74$, $p < 0.05$, $\eta^2_G = 0.20$), and factor 'intensity' ($F(2,28) = 57.24$, $p < 0.001$, $\eta^2_G = 0.30$) indicating that participants' accuracy of judgements differed between emotional displays and levels of intensity (Figure 16). A significant interaction between factors 'emotion' and 'intensity' was also found ($F(2,28) = 7.24$, $p < 0.001$, $\eta^2_G = 0.12$), and Figure 16 suggest that accuracy judgements for angry displays were more affected by the change in the level of intensity then for happy displays. Indeed, posthoc Tukey analysis showed that accuracy of judgements for angry displays increased significantly between each level of intensity ($p < 0.05$), but there was no difference in accuracy of judgements between any level of intensity for happy displays (low and medium: $p = 0.71$, low and high: $p = 0.47$, medium and high: $p = 1$). There was no difference in accuracy of emotion judgements between medium intensity happy and angry displays ($p = 0.08$) or high intensity happy and angry displays ($p = 1.00$), but low intensity happy displays were judged more accurately comparing to low intensity angry displays ($p < 0.001$).

The mean confidence ratings were analysed by carrying out a within-subjects ANOVA with 'emotion' (happy and angry) and 'intensity' (low, medium, high) as within factors. There was a significant main

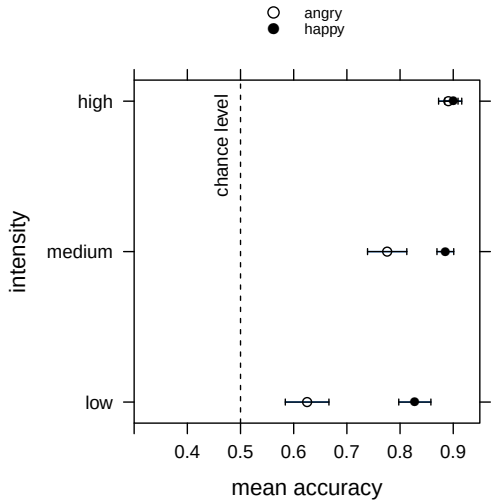


Figure 16: Mean accuracy of emotion judgments for happy and angry displays at low, medium and high intensity in audio-visual experiment. The error bars represent one standard error of the mean.

effect of factor 'emotion' ($F(1,14) = 14.81$, $p < 0.05$, $\eta^2_G = 0.07$) and Figure 17 suggests that happy displays were rated with higher confidence then angry displays. There was also a significant main effect of factor 'intensity' ($F(2,28) = 101.93$, $p < 0.001$, $\eta^2_G = 0.38$), indicating that participants' confidence ratings differed between levels of intensity (Figure 17). A significant interaction between factors 'emotion' and 'intensity' was also found ($F(2,28) = 3.40$, $p = 0.05$, $\eta^2_G = 0.01$) indicating that happy and angry displays were affected differently by the change in the level of intensity (Figure 17). Indeed, posthoc Tukey showed that confidence ratings increased for angry displays between low and high ($p < 0.001$), and medium and high intensity ($p = 0.04$), but not between low and medium intensity ($p = 0.15$). Confidence ratings also increased for happy displays between low and high intensity ($p < 0.001$), but not between low and medium ($p = 0.08$), and medium and high intensity ($p = 0.40$). There was no significant difference in confidence ratings comparing corresponding intensities in happy and angry displays (angry low and happy low: $p = 0.59$, angry medium and happy medium: $p = 0.43$, angry high and happy high: $p = 0.97$).

3.6 COMPARISON OF RESULTS BETWEEN ALL EXPERIMENTAL GROUPS

Audio-visual and auditory groups showed better accuracy of emotion identification and higher confidence in their judgements than visual group

We wanted to further investigate whether there were differences in accuracy of emotion judgements between visual, auditory and audio-visual groups. The mean number of correct responses were analysed

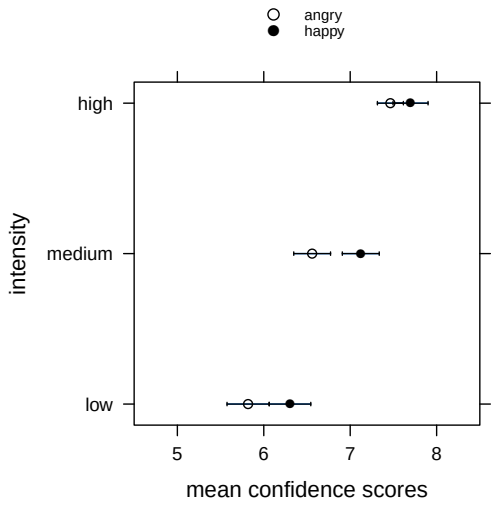


Figure 17: Mean confidence ratings for happy and angry displays at low, medium and high intensity in audio-visual experiment. The error bars represent one standard error of the mean.

by carrying out a mixed design ANOVA with 'emotion' (happy and angry) and 'intensity' (low, medium, high) as within factor, and 'group' (visual, auditory, audio-visual) as between factor. There was a significant main effect of between-factor 'group' ($F(2,40) = 93.01$, $p < 0.001$, $\eta^2_G = 0.40$), and it's clear from Figure 18 that there were differences between groups in the overall number of accurate judgements. There was a significant main effect of within factors 'emotion' ($F(1,40) = 17.86$, $p < 0.001$, $\eta^2_G = 0.17$) and 'intensity' ($F(2,80) = 43.84$, $p < 0.001$, $\eta^2_G = 0.11$), which confirmed earlier results from all experimental groups that happy displays were judged more accurately than angry displays and that there were differences in accuracy of judgements between levels of intensity (Figure 18).

We also found interactions between factors 'group' and 'intensity' ($F(4,80) = 2.61$, $p \leqslant 0.04$, $\eta^2_G = 0.01$), 'emotion' and 'intensity' ($F(2,80) = 36.89$, $p < 0.001$, $\eta^2_G = 0.21$), and 'group', 'emotion', and 'intensity' ($F(4,80) = 3.85$, $p \leqslant 0.01$, $\eta^2_G = 0.05$), but no significant interaction between factors 'group' and 'emotion' was found ($F(2,40) = 1.16$, $p = 0.32$, $\eta^2_G = 0.03$). Those interactions indicated that accuracy judgements for angry and happy displays were differently affected by the change in the level of intensity between visual, auditory and audio-visual groups (Figure 18). In summary, posthoc Tukey analysis revealed that happy and angry displays on all levels of intensity in the visual group were judged less accurately comparing to corresponding displays in auditory and audio-visual groups ($p < 0.05$), as seen on Figure 18. We also found that high intensity happy and angry

displays in the audio-visual group were judged more accurately comparing to high intensity happy, and angry displays in auditory group ($p < 0.05$; Figure 18).

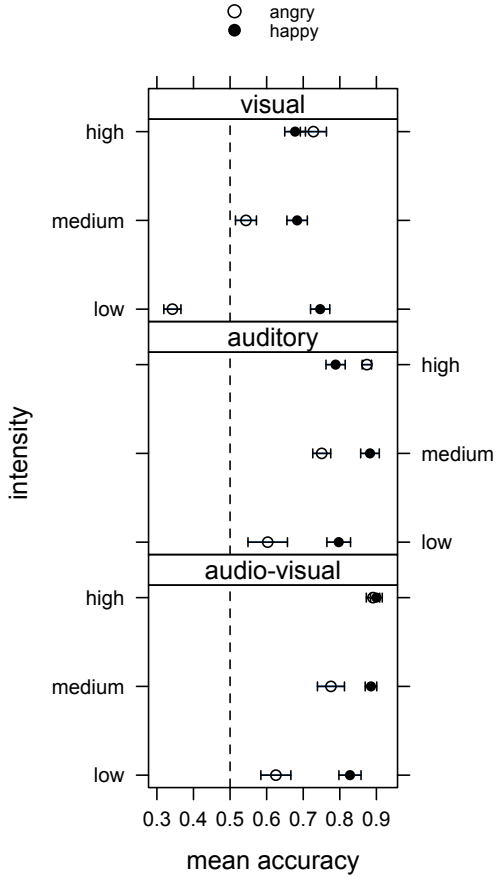


Figure 18: Mean accuracy of emotion judgments for happy and angry displays at low, medium and high intensity in visual, auditory and audio-visual experiment. The error bars represent one standard error of the mean, and the dashed line indicates the level of chance (0.5).

We also compared the differences in confidence ratings between visual, auditory and audio-visual groups. The mean confidence ratings were analysed by carrying out a mixed design ANOVA with 'emotion' (happy and angry) and 'intensity' (low, medium, high) as within factor, and 'group' (visual, auditory, audio-visual) as between factor. There was a significant main effect of between-factor 'group' ($F(2,40) = 4.42$, $p \leqslant 0.02$, $\eta^2_G = 0.16$) indicating that there were differences between groups in the overall confidence ratings (Figure 19). There was a significant main effect of within factor 'intensity' ($F(2,80) = 34.51$, $p < 0.001$, $\eta^2_G = 0.07$), which confirmed earlier results from all experimental groups that there were differences in accuracy of

judgements between levels of intensity (Figure 19). No significant effect of factor 'emotion' was found ($F(1,40) = 1.07$, $p \leqslant 0.31$, $\eta^2_G = 0$) which confirmed earlier results that confidence ratings did not differ between happy and angry displays (Figure 19).

We also found no significant interaction between factors 'group' and 'emotion' ($F(2,40) = 2.94$, $p = 0.06$, $\eta^2_G = 0.01$) and factors 'group' and 'intensity' ($F(4,80) = 2.25$, $p = 0.07$, $\eta^2_G = 0.01$) indicating that the effects of emotion and intensity do not depend on the group. Finally, we also found two significant interaction: between factors 'emotion' and 'intensity' ($F(2,80) = 4.61$, $p \leqslant 0.01$, $\eta^2_G = 0$) and between factors 'group', 'emotion', and 'intensity' ($F(4,80) = 3.94$, $p \leqslant 0.01$, $\eta^2_G = 0.01$). It is clear from Figure 19 that the audio-visual group gave the overall highest confidence ratings compared to auditory and visual groups, but this difference varied depending on emotional valence and emotional intensity of displays. Overall, confidence ratings were higher in the audio-visual group rather then the visual group when participants viewed both happy and angry displays on all levels of intensity - low ($p \leqslant 0.05$), medium ($p \leqslant 0.01$) and high ($p < 0.001$). Confidence ratings were also higher in the audio-visual group rather then the auditory group but only when participants viewed happy displays at medium ($p \leqslant 0.05$) and high ($p \leqslant 0.03$) intensity level (Figure 19). Finally, confidence ratings were higher in the auditory rather then the visual group but only when participants viewed happy and angry displays at medium ($p \leqslant 0.05$) and high ($p \leqslant 0.05$) intensity levels (Figure 19).

3.7 SUPPLEMENTARY RESULTS FOR ALL EXPERIMENTAL GROUPS

3.7.1 *Neutral displays*

In all three experimental groups we wanted to establish whether the proportion of angry judgements was different from the proportion of happy judgements when participants viewed neutral displays. In the visual group, a binomial test showed that the proportion of happy judgements was significantly higher ($p < 0.05$) than the proportion of angry judgements, but only marginally, as seen in Figure 20. However, the binomial test showed there was no significant difference in the proportion of happy and angry judgements for neutral displays in auditory ($p = 0.17$) and audio-visual ($p = 0.72$) groups (Figure 20).

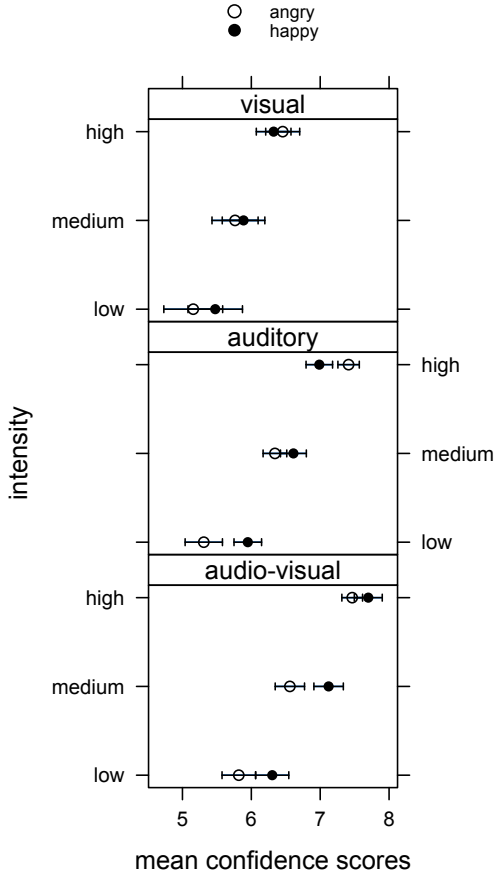


Figure 19: Mean confidence ratings for happy and angry displays at low, medium and high intensity in visual, auditory and audio-visual experiment. The error bars represent one standard error of the mean.

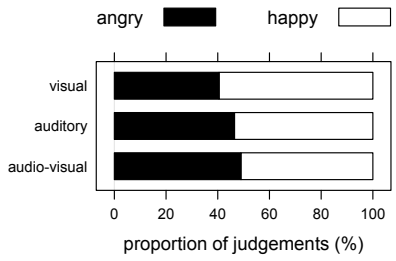


Figure 20: Proportion of angry and happy judgements for neutral displays in visual, auditory and audio-visual experimental groups.

3.7.2 *Actors experience*

Emotions portrayed by experienced actors were identified less accurately than non-experienced actors, but only in visual group - no difference in auditory and audio-visual groups

In Section 2.2.2.2 of Chapter 2 we explained that we use two groups of actors to record our stimulus set: experienced (at least five years in some acting training) or non-experienced (no experience in acting). We wanted to examine whether actors' experience had impact on how accurate participants were able to identify emotional interaction portrayed by those actors. The mean number of correct responses were analysed by carrying out a mixed design ANOVA with 'actors' experience' (experienced, non-experienced) as within factor, and 'group' (visual, auditory, audio-visual) as between factor. We found a significant effect of between factor 'group' ($F(2,42) = 62.11$, $p < 0.001$, $\eta^2_G = 0.69$) which was expected based on the previously described results (Section 3.6) that there were differences in the accuracy of correct judgements between visual, auditory and audio-visual experimental groups. Importantly, we found a significant effect of factor 'experience' ($F(1,42) = 7.29$, $p \leqslant 0.01$, $\eta^2_G = 0.04$) indicating that there were differences in how accurately displays with experienced and non-experienced actors were judged by participants (Figure 21). We also found interactions between factors 'experience' and 'group' ($F(2,42) = 14.45$, $p < 0.001$, $\eta^2_G = 0.15$) indicating that the accuracy of judgements for displays with experienced and non-experienced actors varied between experimental groups. Indeed, posthoc Tukey analysis revealed that interactions portrayed by experienced actors were identified less accurately compared to interactions portrayed by non-experienced actors, but only in the visual group ($p < 0.001$). There were no differences in response accuracy for displays with experienced and non-experienced actors in auditory ($p = 1$) or audio-visual ($p = 0.97$) groups.

3.7.3 *Dialogue type*

No significant effect of dialogue type on emotion judgements

As mentioned in Section 2.2.2.2 of Chapter 2 there were two types of auditory dialogues we used when recording the stimulus set with actors: enquiry (when one actor asked a question, and another replied) and deliberation (when one actor articulated a statement, and another responded with a related statement)[3]. We wanted to examine whether there was any difference in the response accuracy depending on the type of dialogue that participants heard in all three experimental groups. Obviously, there was no auditory dialogue in the visual group, but we nevertheless included this group as a control group.

3 See Table 2 in Chapter 2 for detailed content of dialogues.

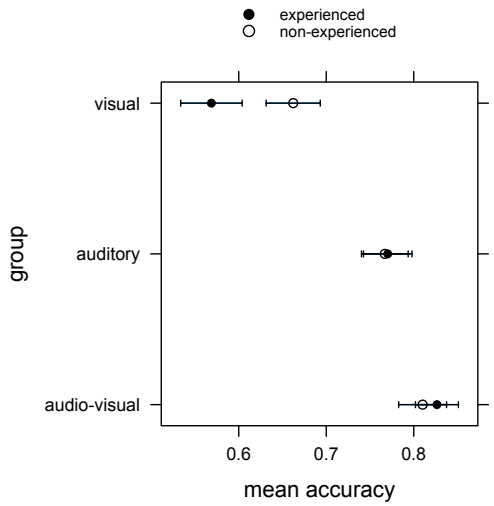


Figure 21: Mean accuracy of emotion judgments for displays with experienced and non-experienced actors in visual, audio-visual and auditory experimental groups. The error bars represent one standard error of the mean.

The mean number of correct responses were analysed by carrying out a mixed design ANOVA with 'dialogue' (enquiry, deliberation) as within factor, and 'group' (visual, auditory, audio-visual) as between factor. We found no significant effect of factor 'dialogue' ($F(1,42) = 0.01$, $p = 0.92$, $\eta^2_G = 0$) and no interaction between factors 'dialogue' and 'group' ($F(2,42) = 0.54$, $p = 0.58$, $\eta^2_G = 0.01$) indicating that dialogue type did not influence the accuracy of participants' judgements in any experimental group (Figure 22). We found a significant effect of between factor 'group' ($F(2,42) = 0.54$, $p = 0.58$, $\eta^2_G = 0.01$) which confirmed previously described results (Section 3.6) that there were differences in the accuracy of correct judgements between visual, auditory and audio-visual experimental groups.

3.7.4 *Questionnaire results*

During debriefing questionnaires we asked participants two questions: How many trials do you think you got correct (give percentage)? And: How many actors do you think we used to create the displays you have seen/heard (give single number)? To compare answers to those questions between the three experimental groups, we used a one-way between-subject ANOVA. In the case of the question about guessing the number of correct responses, we found that there were differences in visual, auditory and audio-visual groups in how many trials participants thought they got correct ($F(2,40) =$

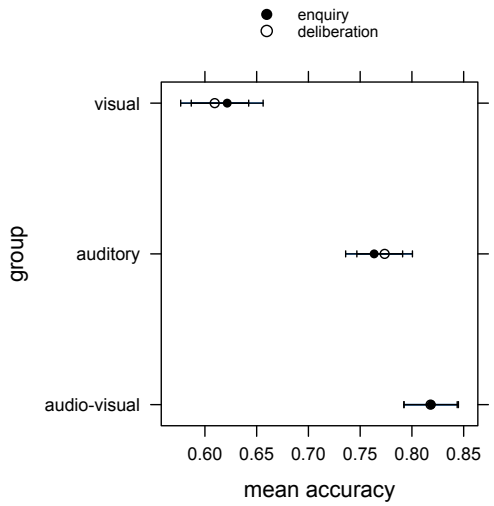
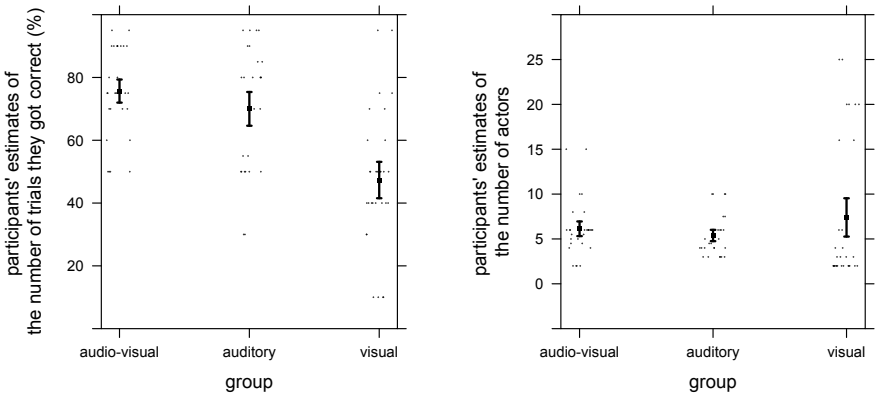


Figure 22: Mean accuracy of emotion judgments for two types of dialogue (enquiry and deliberation) in visual, audio-visual and auditory experimental groups. The error bars represent one standard error of the mean.

8.85, $p < 0.001$, $\eta^2_G = 0.31$). Posthoc Tukey analysis revealed that participants were less certain about the number of correct answers they gave in the visual group, compared to auditory ($p \leqslant 0.01$) and audio-visual ($p < 0.001$) groups (Figure 23a). There was no difference in how many trials participants thought they got correct between auditory and audio-visual groups ($p = 0.72$). Indeed, Figure 23a clearly shows that the median number of responses participants thought they got right was much lower for the visual group, compared to the auditory and audio-visual groups.

A separate one-way between-subject ANOVA showed that there was no difference between visual, auditory and audio-visual groups on how many actors they thought were used to create displays ($F(2,40) = 0.50$, $p = 0.61$, $\eta^2_G = 0.02$). While participants in the visual group reported seeing a slightly higher number of actors on average (7.4), compared to participants in the auditory (5.58) and audio-visual (6.13) groups, there was a large variance in the participants answers within visual group (Figure 23b). Figure 23b clearly shows that the variance of answers in the visual group was very high and medians in all three groups are very similar, which supports the ANOVA results.

We also calculated the frequency of different words that participants used to describe their strategies when doing the experimental task (Table 4). In the visual group almost every participant used the word *speed*, followed by terms for specific body areas: *arm*, *head* and *hands*. In the auditory group, the words *tone* and *volume* were most fre-



(a) How many trials do you think you got correct?

(b) How many actors do you think we used to create the displays you have seen/-heard?

Figure 23: Summary of participants responses to (a) question 1, and (b) question 2 in the questionnaire in visual, auditory and audio-visual experimental groups. Error bars represent standard error of the mean and points represent individual responses.

quently used to describe strategies, while in the audio-visual group - the words *tone* and *movement*.

3.8 DISCUSSION

The experiment described in this chapter was designed to validate the stimulus set created for the study of emotional social interactions. The stimulus set consisted of 252 dyadic point-light displays and voice dialogues performed by nine different actor pairs in happy, angry and neutral manner. We investigated whether observers could identify emotions from the point-light displays and voice dialogues of those actor pairs engaged in a 3-second long social interaction. We used a between-subject design and participants were assigned to one of three separate experimental groups: visual (point-light displays only), auditory (dialogue only) and audio-visual (point-light displays combined with dialogue).

Overall, happy interactions were identified more accurately than angry interactions. A few studies have found similar results; for example, Dittrich *et al.* (1996) showed that happy displays of point-light dancers were identified more accurately compared to angry displays. Belin *et al.* (2008) created and experimentally validated a dataset of nonverbal affect bursts showing that vocal expressions of happiness were better recognized than anger, while Sauter *et al.* (2010)

times used	words	group
10	speed	visual
5	arm	
4	proximity	
4	head	
3	hands	
2	stance	
2	instinct	
2	feet	
1	exaggeration	
8	tone	auditory
6	volume	
3	imagine vis.	
3	speed	
3	amplitude	
2	pronunciation	
1	inflection	
1	stress	
1	force	
1	pitch	
1	accent	
9	tone	audio-visual
9	movement	
4	volume	
2	pitch	
1	arm	
1	hands	
1	proximity	

Table 4: Number of times specific words were used in the questionnaire to describe participants' strategies to do the experimental task, within visual, auditory and audio-visual groups.

showed no differences between nonverbal vocal expression of happiness and anger. However, a majority of studies have found the opposite: observers were better at identifying angry rather than happy emotional expressions when listening to voices (Scherer, 1986), viewing faces (Massaro & Egan, 1996; Fox *et al.*, 2000; Knyazev *et al.*, 2009), watching the actions of a single actor (Pollick *et al.*, 2001; Ma *et al.*, 2004) or watching interactions between two actors (Clarke *et al.*, 2005). For example, Clarke *et al.* (2005) found that angry interactions were 17% more accurately identified compared to happy interactions. Indeed, a number of studies argue that detection of anger serves as an evolutionary indicator of threat (Pichon *et al.*, 2008), and specific brain areas such as the amygdala are tuned to detect angry actions from body movement (de Gelder, 2006) and voices (Johnstone *et al.*, 2006). In short, our result showing that happy interactions were identified more accurately than angry interactions is contradictory to the broader literature. However, it is important to consider our results in the context of the additional factor we introduced - emotional intensity. In all experimental groups, accuracy increased with higher levels of intensity for angry displays, but intensity did not influence accuracy of judgements for happy displays. One explanation for this is that some parameters in low and medium intensity angry displays were more ambiguous to the observers. For example, such features as high velocity or acceleration of body movement, and high intensity or pace of voice were found to be related to angry expressions (Scherer, 2003; Ma *et al.*, 2006). Low and medium intensity angry displays lacked some of these cues and such displays were misidentified as happy because observers detected specific cues that were irrelevant to angry expressions. The effect of emotional intensity was further evident in the participants' confidence in their emotion judgements. In all groups, confidence increased with higher intensity displays, which is clearly visible in Figure 19.

Intensity was also an important factor when we compared the identification performance between the three experimental groups. This comparison clearly indicated that the visual group gave the least accurate emotion judgements compared to the auditory and audio-visual groups (see Figure 18). There were no differences between the auditory and audio-visual group for displays at low and medium emotional intensity levels. However, the high intensity displays in the audio-visual group were judged more accurately than the high intensity displays in the auditory group. This indicated that, to some extent, emotional intensity influenced how participants integrated emotional signals when they were presented with both modalities com-

pared to unimodal conditions. This finding was further supported by the higher confidence ratings in the audio-visual group rather than the auditory group, although this was only the case for happy displays at medium and high levels of emotional intensity. Still, such a result may indicate some level of audio-visual facilitation, as found in the studies with faces and voices (Vines *et al.*, 2006; Vatakis & Spence, 2008; Collignon *et al.*, 2008). We examine this possible facilitation further in experiments described in Chapter 5 where we apply a subset of our stimulus dataset for the study of multisensory integration.

While the audio-visual group performed better overall in identifying emotional interactions compared to the auditory group and the visual group, the auditory group was not far behind in the accuracy of their emotion judgements. The auditory group also gave more accurate judgements than the visual group in every stimulus condition. One possible reason for this is that the voice was not altered or modified in any particular manner, while the point-light displays were an impoverished format of body movement. Voice dialogues were fully intelligible - participants could clearly understand the sentences articulated, although the words we used in those sentences were neutral and did not gave much hint as to the weight of emotional expression. However, the argument that voice intelligibility influenced participants' judgements is weakened by the comparison of accuracy of emotion judgements between the two voice dialogue types we used: enquiry (question and answer) and deliberation (two affirmative sentences). We found no effect of the voice dialogue type in any of the groups indicating that the content of speech was not crucial for participants to make emotion judgements, and indicating that prosody itself may have been more influential. We examine this issue more closely in Chapter 4 when we degrade various parameters of movement and voice to decrease the reliability of both cues and observe how this affects the accuracy of emotion judgements.

We also observed that visual angry displays at low emotional intensity were repeatedly confused with happy displays. Further analysis revealed that the low identification accuracy for visual angry displays was largely due to variance in the way different actor pairs depicted the emotions. We found much higher variance in the judgements between actor pairs for angry interactions rather than happy interactions. It seems that some actors were simply better at portraying angry emotions than others. Such 'actor variance' was also visible in the context of their acting experience. Emotional interactions in the visual group portrayed by experienced actors were judged less accurately than interactions by inexperienced actors. This result supports earlier

suggestions by Ma *et al.* (2006), Busso (2008), and Roether *et al.* (2009) that experienced actors may exaggerate their emotional expression because they are more aware of the components of specific emotional expressions. Interestingly, this effect has not been observed in the auditory and audio-visual groups, indicating that voice may be easier to use as a tool for emotional expression by both experienced and inexperienced actors. Studies on deception and lying have shown that there are clear cues to deception in both voice and movement (Ekman *et al.*, 1991, 1999; DePaulo *et al.*, 2003). One explanation could be that the voice is easier to control than body movement when faking emotion in the acting context. While it is difficult to find comparison studies that used such an acting paradigm, other studies have demonstrated that there are no differences between face, voice and body movement in deception efficiency (Ekman & O'Sullivan, 1991). A more intuitive explanation of the lower accuracy for experienced actors' movement may be related to the notion of deep and surface acting. Drama studies show that affective delivery ratings are negatively related to surface acting but positively related to deep acting (Grandey, 2003). In our capture sessions, we used a realistic role-play scenario but it is possible that not every actor immersed themselves fully into the actual experience of emotions (deep acting). Anecdotally, we observed that some pairs of actors were more engaged in acting than others. Speculatively, it is possible that the difference we observed between interactions portrayed by experienced and inexperienced actors may have been due to variance in the level of engagement of the actors rather than variance in the actors' experience.

Another explanation for the low accuracy of judgements in low intensity angry displays may relate to differences in some features of movement in low intensity angry displays that were more ambiguous to the observers. In the qualitative analysis of movement features in Chapter 2, we reported that low intensity anger was typically portrayed by actors with small-scale arm movements, a small leaning of one actor towards another, or a frequently static expressions. In many cases, it was reported that such features were similar to low intensity happy expressions or neutral expressions. There is also a clear shift to more rapid, sharp and jerky movements for medium and high intensity angry interactions so it is possible that observers categorized low intensity angry displays as happy because of this contrast between low and medium/high intensity angry displays.

In the stimulus set, we also included a number of neutral interactions, although we did not explicitly ask participants to make neutral judgements. As a result, we looked at the proportions of neu-

tral displays judged as happy or angry in the three groups. We have found that there was no difference in the proportion of happy and angry judgements for neutral displays in the auditory and audio-visual groups, but more participants judged neutral displays as happy rather than angry in the visual group, although this difference was not very profound. The tendency to judge neutral visual displays as happy may relate to the same phenomenon as low intensity angry visual displays being frequently confused with happy displays. The low intensity, relatively passive movement information specific to low intensity displays or neutral displays creates a stronger indication of happy interaction due to lower dynamics of movement which is less likely for angry displays than for happy displays. This further adds to the argument that the visual group was faced with higher variance stimuli where judgements were strongly influenced by emotional valence, intensity and the experience of the actors who portrayed emotional interactions.

Looking at the result of the debriefing questionnaire, we found that participants underestimated the number of actors we used to record the displays, but the visual group had particular difficulties in estimating the number of actors - the judgements varied largely between 2 and 25. It is possible that movement on its own does not provide clear identity information and it is easier to establish identity from voice. While participants knew that more than one pair of actors was used to create the movies, they were not able to establish exactly how many which resulted in this very large variance in estimates. Point-light display studies on identity recognition showed that observers can recognize their friends from their point-light walk display only (Cutting & Kozlowski, 1977) and that observers can learn and subsequently recognize individuals from their arm movement (Hill & Pollick, 2000). However, it is clear that detection of different, unfamiliar identities from body movement only is a much more difficult task. In contrast, the voice has a large number of acoustic properties that figure prominently in the literature on talker recognition (Bricker & Pruzansky, 1976; Laver & Trudgill, 1979). These properties include the fundamental frequency of phonation, the typical frequencies of vocal tract resonances, the structure of glottal harmonics, and the fine-grained power spectra of nasals and vowels (Sheffert *et al.*, 2003). For example, vocal pitch is an extremely salient component of vocal quality and accounts for most of the variance in studies on talker recognition (Matsumoto *et al.*, 1973; Gelfer, 1988). In this context, the result that participants were most accurate in establishing the number of actors in the auditory and audio-visual groups were not surprising.

Looking at the number of specific words participants used to describe their strategy to carry out the experimental task, *speed* was the word most frequently used in the visual group, *tone* and *volume* in the auditory group and *tone* and *movement* were used equally frequently in the audio-visual group. A number of studies have found that an increase or decrease in velocity is one of the critical features driving the perception of emotions from body movement (Montepare & Zebrowitz-McArthur, 1988; Pollick *et al.*, 2001; Ada *et al.*, 2003; Roether *et al.*, 2009). Similarly, studies of voice perception have shown that tone of voice and loudness play an important role in emotion identification (Scherer, 2003; Hammerschmidt & Jürgens, 2007).

A number of questions were raised after the validation experiments described in this chapter. The high accuracy of emotion judgements in the auditory group suggested that the voice was a particularly salient cue for participants. However, this high accuracy could have been due to the intelligibility of the dialogue. In the next chapter, we describe conditions where intelligibility was degraded but prosody preserved to test whether intelligibility of dialogue is indeed a factor in emotion identification from the voice. Finally, we found some important differences between the groups that had access to auditory signal and those that did not. Overall, observers in the auditory and audio-visual groups were more accurate in identifying emotions compared to the group that only viewed body movement. We explored this result further in the experiments investigating perception of emotional interactions from multisensory signals described in Chapter 5.

4

THE EFFECT OF DEGRADING VISUAL AND AUDITORY INFORMATION ON THE PERCEPTION OF EMOTIONAL SOCIAL INTERACTIONS

4.1 INTRODUCTION

Let us imagine a situation: people on the street on a typical rainy day. Visibility is poor. From a distance we cannot see them clearly but we can distinguish whether they are fighting or just engaged in horseplay. We can imagine another situation - neighbours arguing behind a thick wall. We cannot understand their words but we clearly know, merely from the sound of their voices, whether they are arguing or having fun. In both situations, we are witnessing a social scene with emotional components in uncertain conditions. The goal of the study described in this chapter was to understand how observers perceive emotions from social stimuli when the visual and auditory components of such stimuli become distorted. One of the major advantages of the point-light display technique is that different components of the movement information can be parametrically manipulated. Therefore, if we want to examine the effect of any type of distortion on perception of movement, point-lights are the perfect tool for such a task. For distortion of point-light displays, we used two standard approaches taken from biological motion research: inversion and scrambling. For auditory stimuli, we examined the effect of auditory brown noise and low-pass filtering on the judgements of emotions.

4.1.1 *Inversion and scrambling of point-light displays*

Inversion of point-light displays disrupts the recognition of actions and emotions in those displays. Studies have shown that observers find it very difficult to identify walking direction (Pavlova & Sokolov, 2000; Bertenthal & Pinto, 1994), types of human action (Dittrich, 1993), the gender of the walker (Barclay *et al.*, 1978), and emotions in dance movements (Dittrich *et al.*, 1996) when point-light displays are presented upside-down. An inversion effect has also been studied in perception of emotional interactions from dyadic point-light displays. Clarke *et al.* (2005) found that people can still identify emotions when

they watch inverted displays of emotional interactions, although recognition performance diminished significantly. In general, the effect of inversion on the perception of body movement has been attributed to an impairment in configural processing which is similar to the face-inversion effect (Valentine, 1988; Farah *et al.*, 1995).

We were also interested in whether scrambling the location of points within the point-light displays of the interacting actors would have an effect on identification of emotions. A number of studies have found that observers can detect the direction of point-light walkers even if the displays are scrambled or embedded in an array of dynamic noise dots (Cutting *et al.*, 1988; Ahlström *et al.*, 1997; Bertenthal & Pinto, 1994; Ikeda *et al.*, 2005). Discrimination of activities from point-light displays is also possible when points are varied in contrast polarity, spatial frequency, or when some points are removed (Ahlström *et al.*, 1997). van Boxtel & Lu (2012) employed a visual search paradigm with threatening boxer targets among emotionally-neutral walker distractors, and vice versa. They found that the boxer was easier to detect for both unmodified and scrambled actions, whereas walkers were not. This indicates that emotionally expressive, threatening actions are better detected than emotionless walkers. In another study, when observers were asked to detect the presence of a point-light walker in a complex point-light mask, their performance depended upon the emotion conveyed by the point-light display (Chouchourelou *et al.*, 2006). Specifically, Chouchourelou *et al.* (2006) demonstrated that observers were best able to detect angry walkers and local velocity cues to anger accounted for high false alarm rates to the presence of angry gaits. These results support the hypothesis that the visual analysis of human action depends upon emotion processes. In this chapter, we employed the approaches of other authors by applying inversion and scrambling to dyadic point-light displays. However, there were two novel features in our approach: we used our stimulus set of displays with emotional interactions between two people instead of showing single actors; and we presented all inverted and scrambled point-light displays from two different viewpoints - side and oblique.

For non-face objects, the vast majority of studies using novel objects have found that performance is strongly viewpoint-dependent, falling off linearly as the difference between the to-be-matched views increases (see review by Hayward, 2003). There is also evidence that certain viewpoints are canonical. Participants label such views as 'better' and are quicker and more accurate in naming objects shown in these views (Palmer *et al.*, 1981; Hayward, 2003). Also, increasing the angle of rotation relative to the canonical perspective monotonically

increases reaction times, an effect often interpreted as evidence for mental rotation (Tarr & Pinker, 1989).

For face objects, studies which have examined viewpoint generalization have also uniformly found performance costs as viewpoint differences increase (Troje & Bülthoff, 1996; Hill *et al.*, 1997; Newell *et al.*, 1999; Lee *et al.*, 2006), and they have also drawn the implication that the neural representation of faces must therefore be viewpoint-specific. This conclusion is supported by the existence of view-specific face sensitive neurons in monkeys' inferotemporal cortex (Perrett *et al.*, 1992), and view-sensitive adaptation of neural activity in the human lateral occipital cortex when faces are used as adapting stimuli (Grill-Spector *et al.*, 1999). Of course, the view-specific face-sensitive neurons revealed by these studies are not necessarily involved in recognising particular individuals, since viewpoint is an important piece of information that might be encoded in its own right, and so they may not be at the heart of viewpoint-dependent performance in face recognition tasks (Burke *et al.*, 2007).

As with non-face and face objects, observers are viewpoint-sensitive in the perception of actions from body movement. Viewpoint is clearly important in the identity inference when observers watch their own movement versus the movement of others. Jokisch *et al.* (2006) displayed individuals' own gait and their friends' gait as point-lights in frontal, half-profile, and profile view. The identification of the friends' gaits was better for frontal and half-profile view than for profile view, whereas identification of each observer's own gait was viewpoint independent. These results show an enhanced visual sensitivity to self-generated movement independent of viewpoint. Other studies have also found that viewpoint can modulate the cortical response to visually presented actions (Perrett *et al.*, 1989; Olofsson *et al.*, 1997; Kilner *et al.*, 2006) and that representation of actions may be, to some extent, dependent on viewpoint (Verfaillie, 1993; Daems & Verfaillie, 1999). Kuhlmann *et al.* (2009) found that humans can discriminate between forward and reverse walking from almost every viewpoint even from a single frame of a point-light display. McAleer & Pollick (2008) demonstrated an advantage for viewing intentional motion from an overhead viewpoint rather than a side viewpoint when participants watched simple interactions between two actors. However, McAleer & Pollick (2008) represented activity with just a single point for each actor and such simplified animacy displays only provide limited information compared to full body point-light interactions. The main limitation of other existing studies investigating the role of viewpoint in perception of actions is that the actions used were typi-

cally simple and did not involve emotional components. It is unclear whether perception of emotional interactions from body movement would also be viewpoint dependent, and this question is examined in the following chapter.

4.1.2 *Distorting and filtering of voice*

In the validation experiment (Chapter 3), we found that observers were better at identifying emotions from voice dialogues rather than from body movement. Perception of voice dialogues was unaffected by the changes in emotional intensity and observers identified emotions in voice dialogues with high confidence. A high degree of accuracy in perception of emotions from voice was due to the strong reliability of the auditory signal. We hypothesized that this reliability might have been due to the robustness and the clear intelligibility of the voice dialogue. We therefore wanted to examine whether removing the intelligible content of dialogue, or degrading the overall quality, would have an effect on the accuracy of emotion identification.

Researchers have examined several means of deriving content-free speech such as playing speech samples backwards (Dirks, 1970), using foreign speech (Kramer, 1963; KRAMER, 1964) and randomized content-splicing (Scherer, 1971; Bezzoijen & Boves, 1986), and allowing the potential use of naturally occurring speech. Nonetheless, these approaches presented a variety of new problems; for example, playing speech samples backwards causes reversal of the acoustic intonation contours. In the experiment described in this chapter, we chose to apply low-pass filtering and brown noise as methods of degrading the quality and intelligibility of speech that overcome some issues with the methods mentioned above.

Low-pass filtering has been adapted from research into speech intelligibility (French & Steinberg, 1947; Pollack, 1948) on the assumption that such filtering would leave a large number of the emotional acoustic cues of the voice intact (Soskin & Kauffman, 1961; Knoll *et al.*, 2009). The applicability of low-pass filtering for speech research has been mainly assessed by comparing a single filter cut-off with the original speech samples (Starkweather, 1956), content-spliced speech (Scherer, 1971) and foreign speech (Kramer, 1963). In a recent study, Knoll *et al.* (2009) sampled natural infant-directed speech, foreigner-directed speech and British adult-directed speech with low-pass filtering at four different cut-offs (1200, 1000, 700 and 400 Hz). Affective ratings of these filtered samples were compared to those of the orig-

inal, unfiltered samples. The results showed that the perception of affect was still very accurate at the lowest cut-off of 400 Hz. In this context, it has been suggested that cues such as intonation contour, rate of speech, pause and rhythm, should remain virtually unaffected by the removal of the upper frequencies (Rogers *et al.*, 1971; Frick, 1985; Knoll *et al.*, 2009), since it is mainly the lower frequencies and tonal quality of the voice that appear to be important in communicating the speaker's emotional state (Scherer, 2003). Therefore, the use of the low-pass filtering method with a cut-off between 350-450 Hz seemed an ideal approach for the purpose of this thesis as we wanted to remove the intelligible content of speech but preserve or slightly distort the prosodic features.

Beside low-pass filtering, another method commonly used by researchers to distort the quality and intelligibility of emotional speech is white noise (You *et al.*, 2006; Hammerschmidt & Jürgens, 2007; Collignon *et al.*, 2008). Adding white or brown noise to speech typically distorts both the intelligibility of the voice and the prosodic features by introducing random signals with flat power spectral density[1]. Brown noise spectral density is inversely proportional to f^2, meaning it has more energy at lower frequencies (Diebold, 2006; Gardiner, 2009). It decreases in power by 6 dB per octave (20 dB per decade) and, when heard, has a 'damped' or 'soft' quality compared to white and pink noise (Barnes & Allan, 1966). We used brown noise in preference to white noise because it resembles the sound of heavy rainfall or a waterfall - there is less information at higher frequencies which makes it sound softer, more natural and less invasive to the listener than white noise (Gardiner, 2009). In fact, the use of both low-pass filtering and brown noise was guided by the principles of ecological validity - to choose a method of audio distortion that emulates real-life conditions. In this context, low-pass filtering makes the voice dialogue sound like neighbours arguing behind a thick wall, or like the sounds heard when submerged in water; the words are unintelligible but the emotion behind the words is detectable. Accordingly, brown noise emulates real-life conditions such as listening to other people's conversation on the street during heavy rainfall.

The major goal of the experiment described in this chapter was to understand how observers perceive emotions in social scenes from dyadic point-light displays and voice dialogues, when those displays and dialogues become distorted. For point-light displays, we wanted to examine the effects of inversion and scrambling on the perception

1 A signal with a flat power spectral density is a signal that contains equal power within any frequency band with a fixed width (Diebold, 2006).

of emotional interactions, and how perception changes when those displays are presented from different viewpoints. For voice dialogues, we wanted to examine whether separate application of noise and low-pass filtering would decrease the accuracy of emotional judgements when listening to these dialogues. Finally, we also wanted to look at the specific perception of neutral interactions. In the validation experiment (see Chapter 3), neutral displays were presented together with expressive emotional displays (angry and happy), but participants were only given the choice to make happy or angry judgements. In the noise experiment described in this chapter, we also gave participants a choice to make neutral judgements. We wanted to establish whether the identification accuracy of neutral displays would be different from happy and angry displays, and how the stimulus manipulations would affect perception of neutral interactions.

4.2 METHODS

4.2.1 *Participants*

A total of 18 participants were recruited for the experiment, 12 female, with a mean age of 22.5 years, ranging from 18 to 44 years. All participants were English-speaking and UK-born. They all reported normal hearing and normal or corrected-to-normal vision. All participants were naive to the purpose of the study and did not have any prior experience with point-light display movies or images. The study received ethical approval from the University of Glasgow's Faculty of Information and Mathematical Sciences Ethics Review Board and every participant signed a consent form.

4.2.2 *Stimuli*

From the original stimulus set described in Chapter 2 we selected eight angry and eight happy displays that were identified with an accuracy of 85% or higher, and an average confidence rating of five or higher. We also chose eight neutral displays that received an approximately equal number of happy and angry judgements (between 40-60%). We were specifically looking for displays which received similar number of correct judgements in both the visual and auditory conditions when validated in experiments described in Chapter 3. The difference in identification accuracy between visual and auditory component of the displays had to be less then 15%. Each display was selected from a different actor pair. The summary of accuracy of

judgements and confidence ratings from Chapter 3 for the selected stimulus subset is shown in Table A.3 in the Appendix.

The selected stimulus subset was broken down into two conditions: visual and auditory. The visual condition included unmodified dyadic point-light displays, inverted versions of displays and scrambled versions of displays. All the versions of displays are illustrated on Figure 24. The inverted displays were generated in Matlab (Mathworks, 2010) by rotating the unmodified displays through 180 degrees. The scrambled version of displays was generated by using the same local joint movements as present in the unmodified displays, but with starting positions of the joint movements scrambled within a kernel defined by the extent of the original figures. The advantage of this approach was that it contained the same local motion signals as the unmodified displays but without an original body shape (Grossman & Blake, 2002; Wuerger *et al.*, 2012).

After applying inversion and scrambling to each point-light display, we used Matlab (Mathworks, 2010) to create each display from two viewpoints: side and oblique. The side view was simply a profile view of actors, the same as used in the validation experiment (top row of Figure 24). For the oblique view, the actors were rotated to a viewpoint affording a 45 degree angle looking down from above, so that one actor was viewed from the back, and another from the front (bottom row of Figure 24). In summary, the visual condition consisted of: 3 stimuli types (unmodified, scrambled, inverted), 2 viewpoints (side, oblique), 3 emotions (happy, angry, neutral) and 8 actor pairs, giving a total of 144 visual displays.

The auditory condition consisted of unmodified voice dialogues, low-pass filtered (LPF) dialogues, and dialogues with brown noise applied to them. All dialogues were processed using Adobe Audition 3 (Adobe Systems, 2008). To create LPF versions of the dialogues, a LPF with a 400 Hz cut-off was applied to the unmodified dialogues attenuating signals with frequencies higher than the cutoff frequency. The actual amount of attenuation for each frequency varies depending on specific filter design. In the case of the LPF we applied, it attenuated spectral frequencies above 400 Hz, as seen in Figure 25. It is sometimes called a high-cut filter, or treble cut filter in audio applications (MacCallum *et al.*, 2011). To create noisy dialogues, brown noise was added to the unmodified clip. All clips were normalized to the same amplitude level of around 65dB. In summary, auditory stimuli consisted of 3 stimulus types (unmodifed, noisy, low-pass filtered), 3 emotions (happy, angry, neutral) and 8 actor pairs, that gave a total of 72 auditory displays.

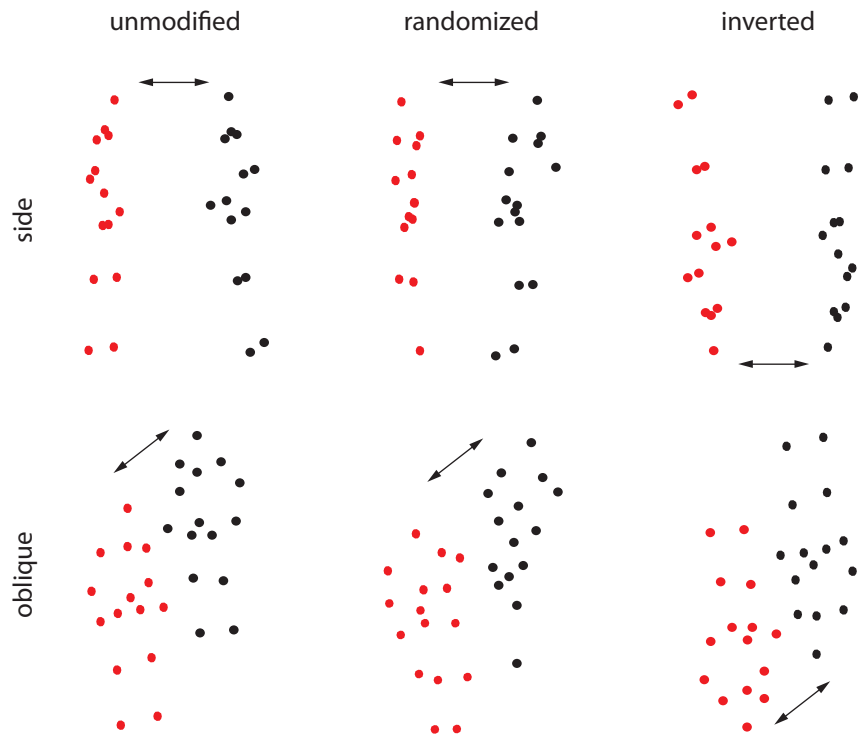


Figure 24: Visual stimuli used in the experiment unmodified, scrambled and inverted displays viewed from side (top row) and oblique (bottom row) viewpoints. Black and red points represent different actors for better visualization. Arrows show the direction of interaction and indicate where the head marker is. Original displays were white points on a black background.

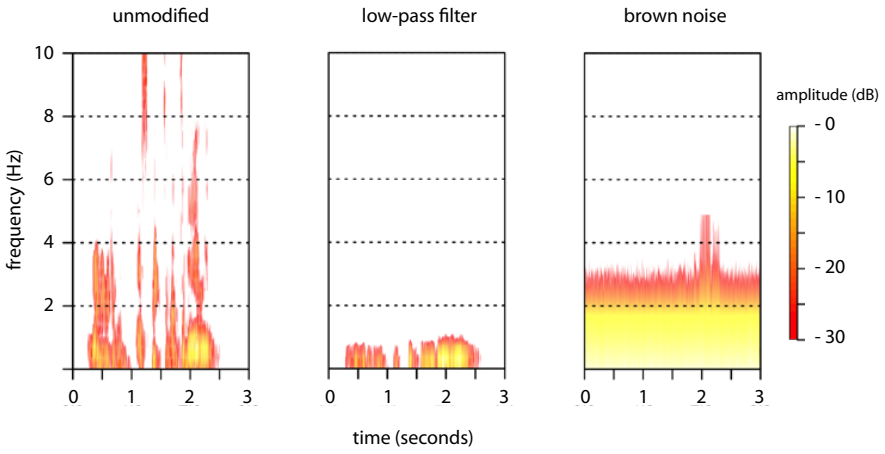


Figure 25: Two-dimensional spectrograph of auditory stimuli including unmodified 3 second, voice dialogue audio wave, and the same wave treated with a low-pass filter and a brown noise.

4.2.3 *Design & Procedure*

Participants were tested in a dark room, with only a small lamp to illuminate the keyboard. They were seated 65 cm from a 21" Cathode Ray Tube (CRT) monitor with resolution of 1024 by 768 pixels, and 60Hz refresh rate. Participants were presented with a total of 432 clips, that included 144 point-light displays (visual condition) and 72 voice dialogues (auditory condition), each display repeated two times. Point-light displays were presented in either unmodified, scrambled or inverted format, from side or oblique view, as white points on a black background. Point-lights subtended a maximum visual angle of approximately 8.5 degrees in height and 6 degrees in width. Voice dialogues were presented as unmodified, noisy or low-pass filtered (LPF) speech, and were accompanied by a white fixation cross shown during each display. Participants were asked to wear Beyer Dynamic DT Headphones, with an intensity at the sound source of 60 dB.

After each display, participants were asked to identify whether the interaction was happy, angry or neutral. They did so by choosing red **H** for happy, red **A** for angry or red **N** for neutral on the keyboard. Immediately after a response, the second screen was presented. Each display was presented twice and the order of all displays was randomized. We used Neurobehavioral Presentation 13.1 software (Neurobehavioral Systems, 2008) to present the displays and collect the responses.

4.3 RESULTS

With each repeated measure analysis of variance (ANOVA) we also calculated generalized *eta squared* (η^2_G) measures of effect size. η^2_G is preferred to *eta squared* and partial *eta squared* because it provides better comparability across between-subjects and within-subjects designs, and it can easily be computed from information provided by standard statistical packages (Olejnik & Algina, 2003; Bakeman, 2005). All p-values presented from posthoc Tukey analysis are presented after adjustment for the multiple comparisons.

4.3.1 *Visual condition*

Unmodified visual displays judged more accurately comparing to inverted and scrambled displays; displays from side view judged more accurately than displays from oblique view

The average number of correct responses were analysed by carrying out a repeated measure ANOVA with 'emotion' (happy, angry, neutral), 'stimuli' (unmodified, scrambled, inverted) and 'viewpoint' (oblique and side) as within factors. There was a significant main

effect of factor 'emotion' ($F(2,34) = 32.49$, $p < 0.001$, $\eta^2_G = 0.41$) indicating that some emotional displays were judged more accurately then others (Figure 26). The significant main effect of factor 'stimuli' ($F(2,34) = 48.21$, $p < 0.001$, $\eta^2_G = 0.13$) was also found, and across the emotions unmodified displays were judged more accurately compared to both inverted ($p < 0.001$) and scrambled ($p < 0.001$) displays (Figure 27), but there was no difference in accuracy of emotion judgements between inverted and scrambled displays ($p = 0.10$). There was a significant main effect of within factor 'viewpoint' ($F(1,17) = 10.68$, $p < 0.001$, $\eta^2_G = 0.03$) showing that overall displays presented from the side viewpoint were judged more accurately compared to displays presented from an oblique viewpoint ($p < 0.001$)[2].

We also found a significant interaction between factors 'emotion' and 'stimuli' ($F(4,68) = 16.59$, $p < 0.001$, $\eta^2_G = 0.15$), indicating that some emotional displays were more affected by stimulus manipulation then others (Figure 26). Indeed, a posthoc Tukey analysis showed that unmodified happy and angry displays were judged more accurately compared to inverted and noisy happy and angry displays ($p < 0.05$). However there was no difference between unmodified neutral and inverted neutral ($p = 0.11$) or noisy neutral ($p = 0.94$) displays.

Finally, a significant interaction between factors 'stimuli' and 'viewpoint' was also found ($F(2,34) = 8.52$, $p < 0.001$, $\eta^2_G = 0.03$) indicating that stimulus manipulation had a different effect on accuracy of emotion judgements depending on the viewpoint (Figure 26). Indeed, posthoc Tukey tests showed that unmodified side displays were judged more accurately compared to unmodified oblique displays ($p < 0.05$), and also more accurately compared to all other combinations of viewpoint and stimulus type ($p < 0.001$). Unmodified oblique displays were judged more accurately compared to scrambled oblique ($p < 0.001$), but similarly to inverted oblique ($p = 0.79$), inverted side ($p = 0.39$) and scrambled side displays ($p = 0.41$). There was no difference in accuracy of emotion judgements between scrambled side and oblique displays ($p = 0.18$), or inverted side and oblique displays ($p = 0.99$).

No interaction was found for factors 'emotion' and 'viewpoint' ($F(2,34) = 0.01$, $p = 0.99$, $\eta^2_G = 0$) indicating that all emotional displays were similarly affected by the change in viewpoint. Finally, there was no interaction for factors 'emotion', 'stimuli' and 'viewpoint' ($F(4,68) = 1.04$, $p = 0.39$, $\eta^2_G = 0.01$) indicating that all emotional displays with

2 To assure that borderline performance seen on Figure 26 for happy displays was above the level of chance, we looked at the overall accuracy of judgements by individual participants. Supplementary Figure A.5 shows that all participants made their judgement above the level of chance.

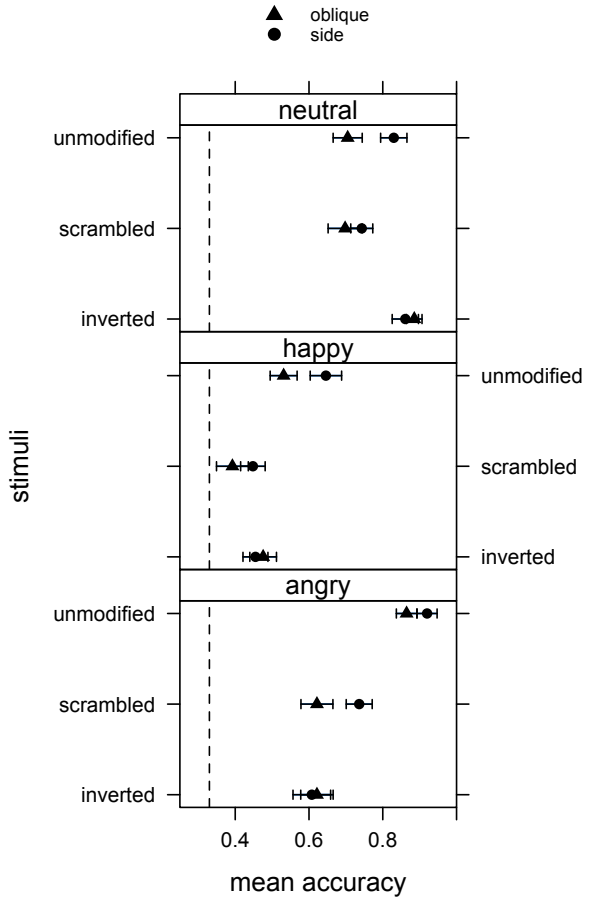


Figure 26: Mean accuracy of emotion judgments for happy, angry and neutral displays in both orientations (side, oblique) and for each stimuli type (unmodified, scrambled, inverted). The error bars represents one standard error of the mean and the dashed line shows the level of chance (0.33).

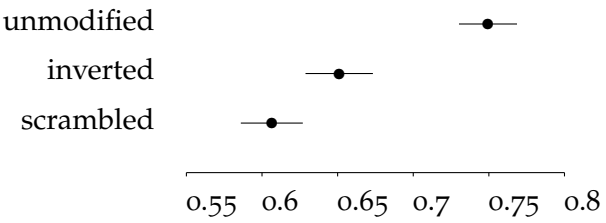


Figure 27: Mean accuracy of emotion judgments with each stimulus manipulation (unmodified, inverted, scrambled) collapsed across emotions and viewpoints. The error bars represents one standard error of the mean.

different stimuli manipulations applied to them were similarly affected by the change in viewpoint.

Participant were fastest when identified unmodified angry displays than all other stimuli conditions

We also looked at how fast participants' gave their responses in the visual condition. The mean reaction times were analysed by carrying out a repeated measures ANOVA with 'emotion' (happy, angry, neutral), 'stimuli' (unmodified, scrambled, inverted) and 'viewpoint' (oblique and side) as within factors. There was a significant main effect of factor 'emotion' ($F(2,34) = 18.97$, $p < 0.001$, $\eta^2_G = 0.07$) demonstrating lower response time in angry compared to happy ($p < 0.001$) and neutral ($p \leqslant 0.01$) displays (Figure 28). We also found a significant effect of factor 'stimuli' ($F(2,34) = 18.71$, $p < 0.001$, $\eta^2_G = 0.06$) showing that across emotions participants were faster when they viewed unmodified displays compared to inverted ($p < 0.001$) or scrambled ($p < 0.001$) conditions (Figure 28). However, ANOVA also revealed a significant interaction between factors 'emotion' and 'stimuli' ($F(4,68) = 5.41$, $p < 0.001$, $\eta^2_G = 0.02$) indicating that participants' response time to some emotional displays was more affected by stimulus conditions than to other emotional displays (Figure 28). Indeed, a posthoc Tukey analysis showed that participants were faster in giving their responses to unmodified angry displays compared to all other conditions (i.e. angry inverted ($p \leqslant 0.01$), angry scrambled ($p \leqslant 0.02$), happy unmodified ($p < 0.001$), happy inverted ($p < 0.001$), happy scrambled ($p < 0.001$), neutral unmodified ($p \leqslant 0.02$), neutral inverted ($p < 0.001$), neutral scrambled ($p < 0.001$)). No significant effect was found for factor 'viewpoint' ($F(1,17) = 0.79$, $p = 0.39$, $\eta^2_G = 0$), and there was no interaction between factors 'emotion' and 'viewpoint' ($F(2,34) = 0.93$, $p = 0.40$, $\eta^2_G = 0$), 'stimuli' and 'viewpoint' ($F(2,34) = 0.22$, $p = 0.80$, $\eta^2_G = 0$), or 'emotion', 'stimuli' and 'viewpoint' ($F(4,68) = 0.85$, $p = 0.50$, $\eta^2_G = 0$), indicating that viewpoint did not have any influence on the participants' response time.

4.3.2 *Auditory condition*

Unmodified angry and happy (not neutral) auditory displays judged more accurately comparing to low-pass filtered and brown noise filtered displays

The mean number of correct responses were analysed by carrying out a repeated measure ANOVA with 'emotion' (happy, angry and neutral) and 'stimuli' (unmodified, low-pass filtered - LPF and noise) as within factors. There was a significant main effect of factor 'emotion' ($F(2,34) = 37.54$, $p < 0.001$, $\eta^2_G = 0.51$) indicating that some emotional dialogues were judged more accurately then other dialogues (Figure 29). The significant main effect of factor 'stimuli' ($F(2,34) = 20.70$, $p < 0.001$, $\eta^2_G = 0.10$) was also found and across emotions unmodified dialogues were judged more accurately compared to both LPF

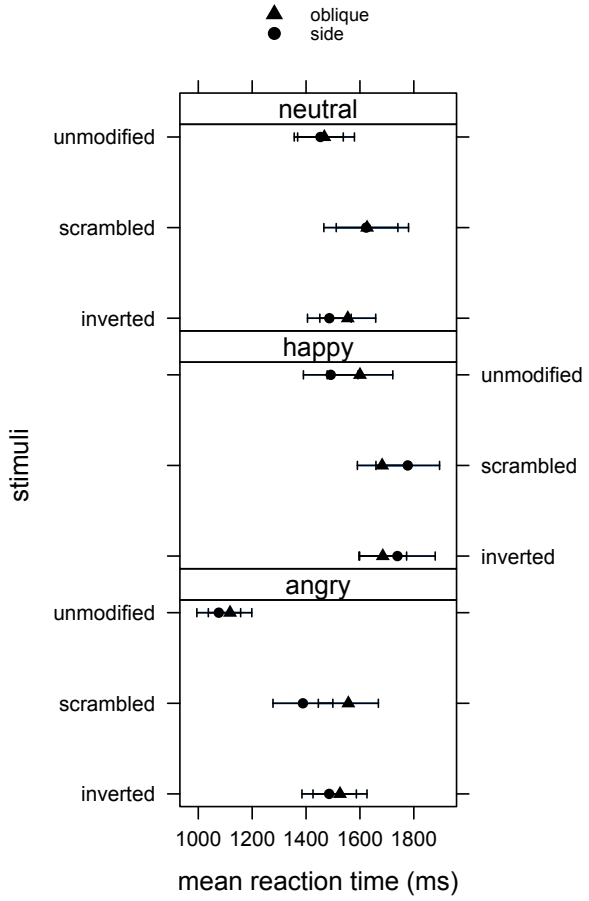


Figure 28: Mean reaction times (ms) obtained during emotion judgments for happy, angry and neutral displays in both orientations (side, oblique) and for each stimulus type (unmodified, scrambled, inverted). The error bars represents one standard error of the mean.

($p < 0.001$) and noisy ($p < 0.001$) dialogues, but there was no difference in accuracy of emotion judgements between LPF and noisy dialogues ($p = 0.94$), as seen on Figure 30. Finally, we found a significant interaction between the factors 'emotion' and 'stimuli' ($F(4,68) = 11.95$, $p < 0.001$, $\eta^2_G = 0.09$) indicating that some emotional displays were more affected by the stimulus manipulation than others (Figure 29). Posthoc tukey analysis showed that happy unmodified displays were judged more accurately compared to happy noisy ($p < 0.001$) and happy LPF ($p < 0.001$) displays. Similarily, angry unmodified dialogues were judged more accurately then angry noisy ($p \leqslant 0.01$) and angry LPF ($p \leqslant 0.05$), while neutral unmodified were judged with similar accuracy to neutral noisy ($p = 0.7$) and neutral LPF ($p = 0.9$) displays. These interactions are clearly visible on Figure 29 where there is not much difference in accuracy of emotion judgements between angry and neutral displays with different stimulus manipulations applied[3].

Participant were fastest when identified unmodified angry and happy (but not neutral) displays than all other stimuli conditions

We also looked at how fast participants gave their responses in the auditory condition. The mean reaction times were analysed by carrying out a repeated measure ANOVA with 'emotion' (happy, angry, neutral), 'stimuli' (unmodified, LPF, noise) as within factors. There was a significant main effect of factor 'emotion' ($F(2,34) = 8.03$, $p < 0.001$, $\eta^2_G = 0.06$) demonstrating lower response time in angry compared to happy ($p \leqslant 0.001$) and neutral ($p \leqslant 0.04$) dialogues (Figure 31). We also found a significant effect of factor 'stimuli' ($F(2,34) = 5.64$, $p < 0.05$, $\eta^2_G = 0.04$) indicating that across emotions stimulus condition had some effect on the participants' response time. Indeed, we also found a weak significant interaction between factors 'emotion' and 'stimuli' ($F(4,68) = 3.93$, $p \leqslant 0.01$, $\eta^2_G = 0.02$) indicating that participants' response time to some emotional dialogues was more affected by stimuli conditions then to other emotional dialogues (Figure 31). Posthoc Tukey analysis showed that participants were faster in giving their responses to unmodified angry and happy dialogues compared to corresponding noisy and LPF angry and happy dialogues ($p < 0.05$), although there was no such difference for neutral displays ($p > 0.05$). Participants were also faster in their response to angry unmodified comparing to happy unmodified and neutral unmodified displays ($p < 0.05$).

3 To assure that borderline performance seen on Figure 29 for happy displays was above the level of chance, we looked at the overall accuracy of judgements by individual participants. Supplementary Figure A.6 shows that all participants made their judgement above the level of chance.

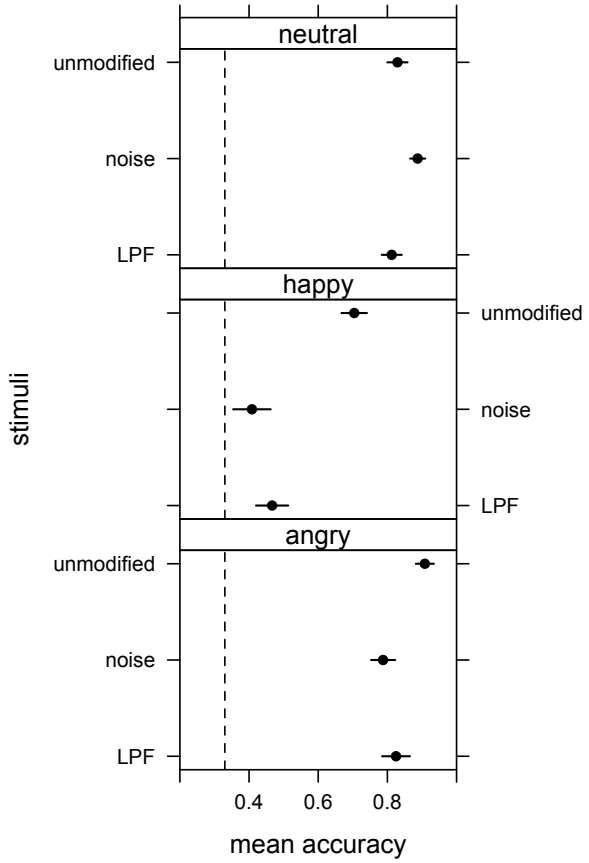


Figure 29: Mean accuracy of emotion judgments for happy, angry and neutral displays with each stimulus manipulation (unmodified, noise, LPF). The error bars represents one standard error of the mean and the dashed line shows the level of chance (0.33).

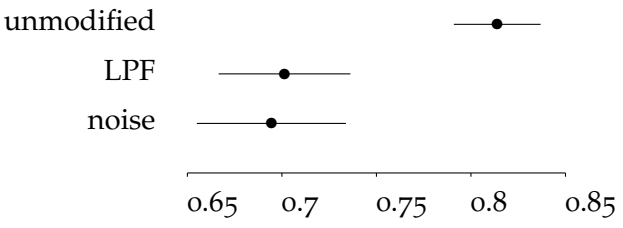


Figure 30: Mean accuracy of emotion judgments with each stimulus manipulation (unmodified, noise, LPF). The error bars represents one standard error of the mean.

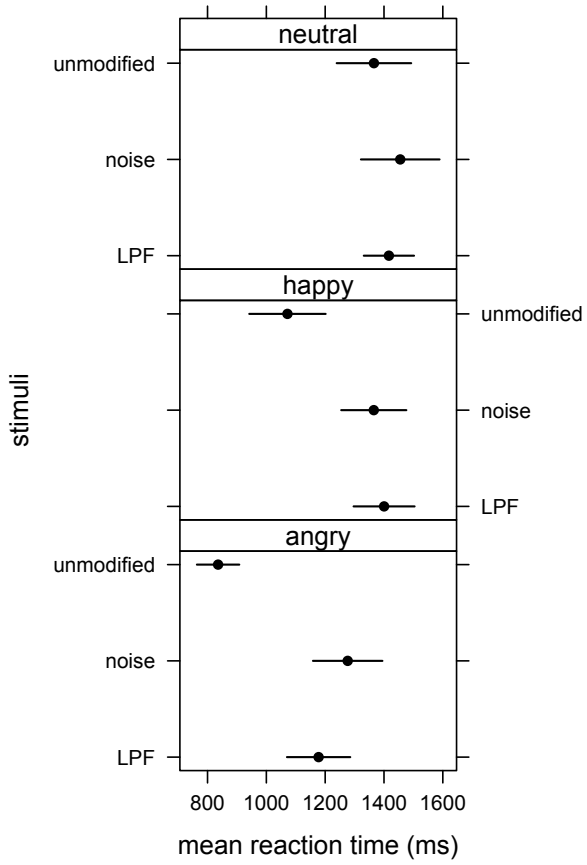


Figure 31: Mean reaction times (ms) obtained during emotion judgments for happy, angry and neutral displays for each stimulus type (unmodified, noise, LPF). The error bars represents one standard error of the mean.

4.4 DISCUSSION

Previous studies have indicated that observers can recognize a range of emotions from just a few point-lights representing a single person, although this recognition is impaired if the point-lights are scrambled (Chouchourelou *et al.*, 2006; van Boxtel & Lu, 2012), inverted (Dittrich *et al.*, 1996; Clarke *et al.*, 2005) or presented from different viewpoints (Kuhlmann *et al.*, 2009). Similarly, listeners can recognize emotions from voices, although methods such as white noise (Collignon *et al.*, 2008) or low-pass filtering (Knoll *et al.*, 2009) applied to speech impair the intelligibility of speech and emotion recognition. However, it is unclear how social aspects such as dyadic interactions affect this robustness of visual and auditory stimuli. In this chapter we investigated whether observers could identify emotions from the point-light displays and voice dialogues of two actors engaged in social interactions, even when body movement and voice information were distorted. For visual stimuli, we focused on examining how inversion and scrambling impacted the judgements of emotional dyadic interaction. We also wanted to examine how the change in presentation viewpoint of point-light displays between side and oblique views impacted the judgements. Correspondingly, we were interested in how typical methods of audio distortion, such as the addition of brown noise or low-pass filtering, would affect judgements of emotional voice dialogues.

For the experiments described in this chapter, we used a subset of the stimulus set validated in Chapter 3. We chose only angry and happy displays with high emotional intensity. The reason for only using a subset of the original, full stimulus set was that we wanted to reject a number of displays that were repeatedly confused by participants with the opposite emotion. We chose high intensity displays so that we could clearly observe any distortion effect on emotion judgements. Finally, we wanted to limit the number of displays in the stimulus subset to allow for a wider range of within-stimulus distortion conditions to be introduced.

We found that participants were fastest and most accurate in identifying angry displays regardless of whether they were presented as point-light displays or voice dialogues. This contrasted with the results from the validation experiments described in Chapter 3 where we found that happy displays were better identified than angry ones, although we have already discussed that this effect was mainly due to differences in intensity levels. As explained in the discussion of Chapter 3, angry displays were strongly affected by intensity, with

overall accuracy decreasing at lower levels of intensity. In the current noise experiment, we selected only high intensity displays which had been identified with high accuracy.

In general, the results that angry displays are easier and faster to identify than happy displays are consistent with the studies arguing that humans have a high sensitivity to negative affect as an indicator of threat (Pichon *et al.*, 2008). Strong sensitivity to anger displays has been shown for emotional faces (Fox *et al.*, 2000), voices (Green *et al.*, 2010) and body movement (Ikeda & Watanabe, 2009). Pichon *et al.* (2008) found that whole-body expressions of anger elicited activity in regions including the amygdala and the lateral orbitofrontal cortex which play a role in the affective evaluation of the stimuli. Additionally, Pichon *et al.* (2008) have shown that the perception of anger engaged the hypothalamus, the ventromedial prefrontal cortex, the temporal pole and the premotor cortex. These regions of the brain are linked to autonomic reactions and motor responses related to defensive behaviours. In a series of separate experiments, Fox *et al.* (2000) and Ohman *et al.* (2001) showed that detection of angry facial expressions was faster and more efficient than for happy faces. Clarke *et al.* (2005) also found that angry point-light interactions were identified more accurately compared to other positive and negative emotional interactions such as joy, love, sadness or fear. In the perception of emotions from voice, anger is generally best recognized, followed by sadness and fear (Scherer, 2003). As Scherer (2003) points out, there is a clear adaptive advantage in being able to threaten foes in anger over large distances - something for which the voice is ideally suited.

Inversion and scrambling of dyadic point-light displays decreased identification accuracy by approximately 15-20%, although participants were still able to identify the emotions above the level of chance. These results are in line with existing research. In a series of psychophysical studies, Chouchourelou *et al.* (2006) reported that observers were able to identify emotions of happiness, fear, sadness and anger above chance level from scrambled point-light walkers. Ikeda & Watanabe (2009) also showed that angry and happy point-light walkers can be detected behind a scrambled mask, and again detection of anger was stronger, a point which is also supported by our findings. Pollick *et al.* (2001) demonstrated that even when the phase and position relationships were distorted, participants categorized the stimuli under correct emotions above the chance level despite the fact that the stimuli did not resemble humans. One explanation is that the detection of emotion from dynamic stimuli can be sustained by kinematics of the body such as the local velocity sig-

nals and that it does not require much information regarding body structure. However, Thurman & Lu (2013) argues that perception of animacy in scrambled biological motion involves not only analysis of local intrinsic motion, but also its congruency with global extrinsic motion and global spatial structure. Thus, Thurman & Lu (2013) suggest a strong influence of prior knowledge of characteristic features of creatures in the natural environment. Additionally, Chang & Troje (2008) found that spatially scrambled point-light creatures were perceived as being animate despite disruption to the canonical biological form, particularly when the local trajectories represented upright, rather than inverted, movements. Furthermore, animacy ratings correlated significantly with the ability to subsequently discriminate the walking direction of the scrambled animations. This finding suggests that animacy may be associated with basic perceptual mechanisms, such as the proposed "life detector" (Troje & Westhoff, 2006), which enable detection of vital intrinsic motion information signalling directionality.

The inversion effect we found was similar to what Clarke *et al.* (2005) showed in their study with dyadic point-light interactions, and Dittrich *et al.* (1996) showed with point-light displays of professional dancers, i.e. participants were much less accurate in identifying emotions from upside-down displays, but still above the level of chance. Our results showed that the effect of inversion was less evident than what Clarke *et al.* (2005) and Dittrich *et al.* (1996) found, with recognition of inverted angry and happy point-light displays being well above the level of chance. One suggestion is that the presence of a second agent made it easier for observers to recognize emotion from the inverted orientation. Such a view would be consistent with the suggestions of Neri *et al.* (2006) and Manera *et al.* (2011) that observation of communicative interactions improves detection of agents and detection of the emotions they express. Our results on inversion also suggest that the effect of local cues is more powerful than assumed. Inversion might distort perception of walking direction, but kinetic features such as velocity and direction are still highly recognizable, even from severely distorted interactions.

Our results also indicate that perception of emotions from unmodified point-light interactions is viewpoint dependent. Participants found it easier to identify angry, happy and neutral displays from a side view rather than an oblique view, although different viewpoints did not affect the response time. Placing these results in the context of existing research is challenging mainly because of the different modes of viewpoint investigated by other researchers. The majority of stud-

ies on the effect of viewpoint argue for the advantage of frontal views compared to profile (i.e. side) and half-profile views (Mather & Murdoch, 1994; Troje, 2002; Troje *et al.*, 2005). However, all of these studies focus on the perception of a single actor rather than the interaction between two actors. When looking at the interactions, it is intuitively plausible to assume that a side view would be the optimal viewpoint. When we observe people interacting around us, they are most likely facing each other rather than us. Our result supports this intuitive assumption that a side viewpoint is optimal in the observation of interactions between other people.

We also found that inversion and scrambling had a different effect depending on the viewpoint from which those displays were presented. In the case of both inverted and scrambled displays, there was no difference between the two viewpoints. It is possible that the viewpoint effect only applied to unmodified displays because the oblique view presented a more complex motion scene compared to a side view, which is more commonly encountered when we observe other people interacting around us. The viewpoint became less relevant when inversion impaired configural processing and scrambled displays lacked coherent global structures. It is also possible that in the oblique viewpoint, there was an overlap and occlusion between points while actors were dynamically interacting. In the context of single agent studies, Coulson (2004) also highlights the role of occluding effects of particular viewpoints on some classes of stimulus. For example, closed and downward looking postures for disgust, fear and sadness appear smaller from the front and present less information to the viewer than from side and rear views. The overall preference for frontal views suggests that attributing emotion to a body posture is a great deal easier when the person adopting the posture is facing the perceiver. Such an orientation, while not necessarily ideal for perceiving the three-dimensional relationships between body segments, may nonetheless enhance recognition due to its interpersonal significance (Coulson, 2004). While optimal perception from a frontal viewpoint applies to interpersonal perception, it is different in the context of intrapersonal perception. Our finding clearly suggests that a side view may be the optimal viewpoint for intrapersonal identification of emotions.

Our results also showed that neutral displays were identified with very high accuracy and identification was not affected by any stimuli manipulation in both visual and auditory conditions. At the same time, participants were slower in response to neutral stimuli compared to angry, but not happy, stimuli in both visual and auditory

conditions. Such a finding is in line with other studies on the perception of neutral versus emotionally expressive stimuli (Chouchourelou *et al.*, 2006; Atkinson *et al.*, 2007). Neutral body movements are typically less dynamic and the actors produce significantly more movements in angry interactions than in neutral interactions. Neutral voice dialogue is also inexpressive. Atkinson *et al.* (2007) found that inversion and reversal impaired the classification of fear and disgust more than it did the neutral expressions. Chouchourelou *et al.* (2006) found that detection performance with neutral gaits was lower than with angry gaits. Studies using static facial expressions revealed that neutral faces are harder to detect amongst other emotional faces, especially angry ones (Hansen & Hansen, 1988; Fox *et al.*, 2000; Eastwood *et al.*, 2001; Ohman *et al.*, 2001). Ferri *et al.* (2013) showed that the observation of an action embedded in an emotional context (i.e. angry or happy facial expression), compared with the observation of the same action embedded in a neutral context, elicited higher neural response at the level of motor frontal cortices, temporal and occipital cortices, bilaterally. These findings suggest that emotions exert a modulatory role on action observation in different cortical areas involved in action processing, in contrast with the neutral actions.

Moving from the results on degrading visual to auditory stimuli, it is important to highlight that our main goal was to test whether distortion of intelligibility of dialogues as well as its prosodic quality will have an effect on participants' judgements. Overall, it is clear that auditory dialogues showed high robustness against different filtering methods. Although low-pass filtering and brown noise decreased the accuracy of emotion judgements from voice, participants were still able to judge the distorted voices above the level of chance. Looking at the interaction between emotions in dialogue and the filtering method used, we found that filtering affected identification accuracy more for happy than angry displays. Specifically, identification accuracy for filtered dialogues dropped by around 20% for happy displays comparing to around 8% for angry displays. It is possible that filtering methods we used were more efficient with happy displays due to removal of specific cues characteristic to the perception of happiness. For example, increase in levels of high-frequency energy is frequently attributed as a cue to perception of elated joy, enjoyment or happiness (e.g. Scherer, 1986; Banse & Scherer, 1996), and this cue might have been removed or severely distorted when filtered with LPF or brown noise. At the same time there was no difference in accuracy of judgements for neutral displays between unmodified, noisy and LPF conditions, supporting corresponding results from visual conditions that

identification of neutral interactions was not influenced by distortion methods due to its passive and non-expressive format.

To summarize, it is clear that observers are good at detecting emotional signals from social scenes even in inherently uncertain and unreliable environments. In spite of using different methods to distort the quality of visual and auditory social stimuli, we found that observers could still identify emotions well above the level of chance. It is also clear that voice dialogues play an important role during the perception of emotional social interactions and that observers are very sensitive to voice information. In next Chapter 5, we focus on the final stage of this thesis project: the effect of audio-visual integration of emotional and social signals and the application of our stimulus set for multisensory studies of emotional social interactions.

5

MULTIMODAL INTEGRATION OF EMOTIONAL SIGNALS FROM DYADIC DISPLAYS OF BODY MOVEMENT AND VOICE.

5.1 INTRODUCTION

This chapter describes four experiments aiming at the application of the stimulus set developed and validated in Chapters 2, 3 and 4 for the study of multimodal integration of emotional social signals. A short overview is given of existing studies using multimodal social stimuli as well as the motivation behind validating stimuli described in this thesis in the context of existing literature. This is followed by a detailed description of the methods and results used in four consecutive experiments with the methodological approach similar to Collignon *et al.* (2008) and Petrini *et al.* (2010) (Experiments 1, 2, 3) as well as Alais & Burr (2004) and Ernst & Banks (2002) (Experiment 4). We conclude this chapter with a summary of the results and a discussion of them in the context of existing literature on multimodal integration and cue combination.

5.1.1 *Brief overview of multisensory studies with emotional stimuli*

Perception of emotions is a multimodal event. By integrating signals from facial expressions, body movements, vocal prosody and other cues, we make emotional judgements about others. This multisensory integration of emotional expressions has been studied with faces and voices (de Gelder & Vroomen, 2000; Kreifelts *et al.*, 2007; Collignon *et al.*, 2008), body expression and faces (Meeren *et al.*, 2005; Van den Stock *et al.*, 2007), body expression with sound stimuli (Vines *et al.*, 2006; Petrini *et al.*, 2010), and body expressions and voices (Pichon *et al.*, 2008; Stienen *et al.*, 2011). A number of studies investigating the perception of emotions from facial expression and voices suggest strong bidirectional links between emotion detection processes in vision and audition (Massaro & Egan, 1996; de Gelder & Vroomen, 2000; Collignon *et al.*, 2008; Jessen *et al.*, 2012). For example, de Gelder & Vroomen (2000) presented static images of facial expressions that were morphed on the continuum between happy and sad, while at the same time presenting a short emotional vocal sentence. Partici-

pants were instructed to categorize only emotional expression in the face and ignore the voice. The results showed a clear influence of the task-irrelevant auditory modality (voice) on the target visual modality (facial expression). Similarly, Collignon *et al.* (2008) asked participants to identify fearful and disgust expressions from dynamic faces combined with short vocalizations, or presented unimodally. Participants were faster and more accurate in their responses when they viewed faces combined with voices, rather than when both cues were presented unimodally. When participants were presented with emotionally incongruent combinations between faces and voices (e.g. fearful faces with disgust voices), they made their emotion judgements based on facial expressions rather than voices. However, when the visual quality of the facial expression was diminished, participants categorized the emotion using the more reliable voice.

Studies using stimuli that were combinations between body expression and voice have indicated that recognition performance for body movements and voices is similar to that found for faces and voices (Van den Stock *et al.*, 2008; Stienen *et al.*, 2011). Van den Stock *et al.* (2008) used whole-body video images with the facial expression blanked and included human as well as animal sounds.The authors asked participants to attend to the action displayed by the body and to categorize the expressed emotion. The results revealed that recognition of body language was biased towards the emotion expressed by the simultaneously presented auditory information supporting similar findings with faces and voices (de Gelder & Vroomen, 2000). Stienen *et al.* (2011) asked participants to make happy and fearful judgements from static frames of body expressions and short vocalizations. Similar to Collignon *et al.* (2008), the authors used emotionally incongruent displays and demonstrated that congruency between the voice and the bodily expression influenced the perception of emotion.

As we discussed in the very first paragraphs of Chapter 1 using the pub example, our environment is inherently noisy in the sense that we frequently face uncertain sensory conditions in social situations and these conditions can impair our judgement of emotions. In the pub example (see Section 1.1 of Chapter 1), these uncertain sensory conditions involved poor visibility of people observed from a further distance in low light, or poor auditory reception due to the background noise coming from surrounding conversations. Scaling this situation to controlled experimental conditions, we wanted to examine whether observers were able to optimally combine visual and auditory cues in order to reduce sensory uncertainty for high-level factors such as perceived emotions. Cue combination in

conditions of sensory uncertainty has already been examined in a number of studies with visual and auditory stimuli (Alais & Burr, 2004), visual and haptic stimuli (Ernst & Banks, 2002), or haptic and auditory stimuli (Bresciani & Ernst, 2007). Alais & Burr (2004) used the well-known "ventriloquist effect"[1] by investigating a spatial localization of audio-visual stimuli. The authors showed that when visual localization was good, vision dominated and captured sound, but for severely blurred visual stimuli that were poorly localized, the reverse holds true: sound captures vision. For less blurred stimuli, neither sense dominates and perception follows the mean position. Precision of bimodal localization was usually better than either the visual or the auditory unimodal presentation. All the results were well explained not by one sense capturing the other, but by a simple model of optimal combination of visual and auditory information. Ernst & Banks (2002) showed the same effects but by using visual and haptic stimuli where participants had to estimate the thickness of a vertical bar displayed behind visual white noise.

5.1.2 *Goals and motivation behind multisensory studies with emotional social interactions*

Ernst & Banks (2002), Alais & Burr (2004) and Bresciani & Ernst (2007) all used very simple stimulus conditions without any emotional component. In contrast, we were more interested in whether combining visual and auditory cues reduces sensory uncertainty for high-level factors such as perceived emotions. Additionally, in all the studies described above, observers were always presented with a single agent's expression rather than with interactions between multiple agents. Experiments described in this chapter aimed to extend the investigation of perception of emotions to a more social context, using audio-visual stimuli of dyadic point-light displays combined with voice dialogues, at the same time preserving the same theoretical and methodological framework as studies by Ernst & Banks (2002), Collignon *et al.* (2008), Stienen *et al.* (2011), and Petrini *et al.* (2010). Chapter 3 has already discussed that one of the advantages of the developed stimulus set was that we captured both body movement and voice dialogues between interacting actors in a synchronized manner. We have already observed that both voice and body movement are highly salient cues on their own. We also showed that observers were highly accurate when presented with combination of body movement and voice, al-

1 Ventriloquism is the ancient art of making one's voice appear to come from elsewhere (Alais & Burr, 2004, p. 257).

though in Chapter 3 it was only tested with a very large stimuli set, and we were uncertain whether this high accuracy was due to audio-visual facilitation or rather the quality of different displays in a large, non-validated stimuli set used in Chapter 3. Our primary goal for the experiments described in this chapter was to make the stimulus set we developed applicable to the multisensory studies. This is mainly because such complex stimuli as the one we developed in this thesis have never before been used in multisensory studies. We conducted four separate experiments as described in this chapter, replicating the experimental design of some well established studies in the field of multisensory integration. In Experiments 1 and 2, we addressed the basic questions of how well participants integrate emotional signals from movement and voice. In Experiment 3, we requested that the participants pay attention to only one modality at a time to ascertain whether any multimodal effects found in Experiments 1 and 2 were due to automatic processes and would not disappear when participants were asked to ignore one of the two modalities. In Experiment 4, we investigated whether combining visual and auditory cues reduces sensory uncertainty for high-level factors such as perceived emotions. The remaining sections of this chapter describe the methods and results obtained from all four experiments highlighted above in detail. Subsequently, there will be a focus on discussion of these results in the context of the broader literature on multisensory integration of emotional signals and cue integration.

5.2 EXPERIMENT 1 AND 2: FILTERING AUDIO

Filtering audio with brown noise (Exp. 1) or LPF (Exp. 2).

The first two experiments addressed the basic questions of how well participants integrate emotional signals from movement and voice. In these two experiments, we used a similar procedure to the one applied by Collignon *et al.* (2008) and Petrini *et al.* (2010). The participants were required to discriminate between angry and happy expressions either displayed aurally, visually or audio-visually, in a congruent (the same expressions in the two modalities) or incongruent way (different expressions in the two modalities). This method allows us to investigate whether the presentation of bimodal congruent stimuli improves the participants' performance and which modality dominates in a conflicting situation. Since we observed in previous chapters that auditory information presented unimodally was judged with higher accuracy than visual information, we included conditions in which the reliability of the auditory stimuli was decreased to a level similar to the visual stimuli. We used two filtering methods

described in Chapter 4 - brown noise and low-pass filtering (LPF). We tested these two methods in separate experiments to avoid carry-over effects and to establish whether these two filtering methods were similar or different in how they impact integration of auditory information. Therefore, both Experiments 1 and 2 used the same stimuli, design and procedure, except that we filtered audio with brown noise in Experiment 1 and with LPF in Experiment 2, and we used different groups of participants in both experiments.

Note that in Chapter 4 brown noise and LPF were only tested unimodally

5.2.1 *Methods*

5.2.1.1 *Participants*

A total of 16 participants were recruited for Experiment 1: 8 female and 8 male, with a mean age of 22 years, ranging from 18 to 34 years. A total of 15 participated in Experiment 2: 7 female and 8 male, with a mean age of 22 years, ranging from 17 to 32 years. All participants were English speakers and UK born. All reported normal hearing and normal or corrected-to-normal vision. All participants were naive to the purpose of the study and had no prior experience with point-light display movies or images. The study received ethical approval from the University of Glasgow's Faculty of Information and Mathematical Sciences Ethics Review Board and every participant signed a consent form.

5.2.1.2 *Stimuli*

From the original stimulus set described in Chapter 2 we selected eight angry and eight happy displays that were identified with an accuracy of 75% or higher. We were specifically looking for displays which received similar numbers of correct judgements for both visual and auditory modalities during the experiment described in Chapter 3.

The visual stimuli were the same for both Experiment 1 and 2 described in this chapter using unmodified, side view, dyadic point-light displays. The auditory stimuli differed between Experiment 1 and 2. Since preliminary results described in Chapter 4 demonstrated a high accuracy of judgements with the auditory displays, we decreased the reliability of the auditory target to the level similar to the visual displays. In Experiment 1 we used unmodified voice dialogues, and dialogues with brown noise applied to them. In Experiment 2 we used unmodified dialogues, and low-pass filtered (LPF) dialogues. The methods that were used to create distorted dialogues

are described in Chapter 4.2.2. The reason we decided to use both brown noise and low-pass filtering in separate experiments was that both methods were similarly effective in degrading reliability of the auditory signal, when tested unimodally (see description of results in Section 4.3.2 or Figure 30 of Chapter 4). However, both methods were qualitatively different in the way they distorted audio, so we wanted to make sure that any effects, such as the lack of intelligibility, were tested when presented in the bimodal context.

The bimodal stimuli were obtained by combining corresponding point-light displays with voice dialogues. The matching could either be 'congruent', with the use of point-light displays and voice dialogues expressing the same emotion (e.g. angry point-lights/angry voices), or 'incongruent', with point-light displays and voice dialogues expressing different emotions (e.g. happy point-lights/angry voices). We created two incongruent versions of bimodal stimuli: point-light displays combined with unmodified voice dialogues, and point-light displays combined with dialogues filtered with brown noise (Experiment 1) or LPF (Experiment 2). A schematic explanation of how bimodal incongruent stimuli were created is shown on Figure 32. During creation of 'incongruent' combinations, we were specifically looking for different actor pairs who had closely corresponding timings of interactions. This was done to avoid temporal desynchronization between point-light displays and voice dialogues and was done manually using Adobe Audition 3 (Adobe Systems, 2008).

5.2.1.3 *Design & Procedure*

In both experiments, participants were tested in a dark room, with only a small lamp to illuminate the keyboard. They were seated approximately 65 cm from a 21" Cathode Ray Tube (CRT) monitor with resolution of 1024 by 768 pixels, and 60Hz refresh rate. Point-light displays subtended a maximum visual angle of approximately 8.5 degrees in height and 6 degrees in width. Voice dialogues were presented simultaneously with a white fixation cross shown during each display. Participants wore headphones (Beyer Dynamic DT Headphones), with an intensity at the sound source of 60 dB. We used Neurobehavioral Presentation 13.1 software (Neurobehavioral Systems, 2008) to present the displays and collect the responses. After each display, participants were asked to identify whether the presented interaction was happy or angry. They did so by choosing red **H** for happy, or red **A** for angry on the keyboard.

In both experiments, participants were presented with a total of 336 displays that included three repetitions of all conditions randomly in-

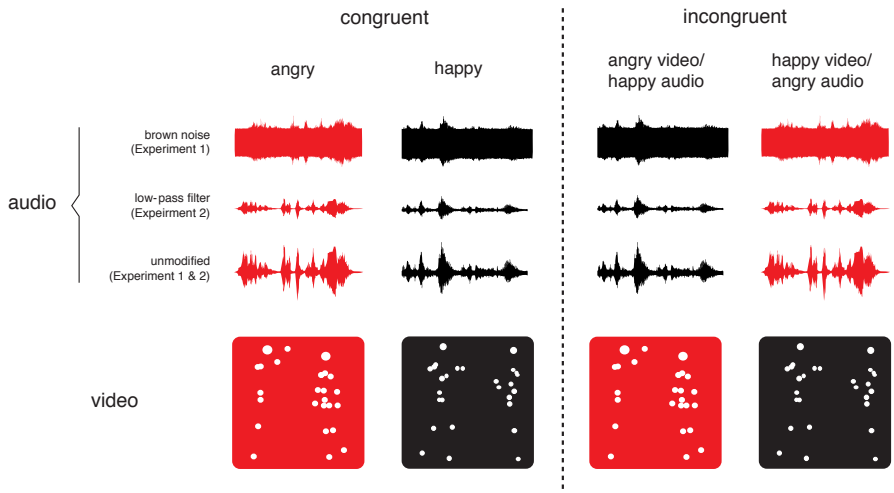


Figure 32: Schematic explanation of creating bimodal incongruent stimuli. Visual angry displays were combined with auditory happy displays, while visual happy were combined with auditory angry displays. Two types of auditory stimuli are also highlighted. For illustrative purposes, red represents angry displays and black - happy displays.

terleaved in 3 separate blocks of 112 stimuli. Those stimuli consisted of: 2 emotions (happy, angry), 7 stimulus types (visual, auditory unmodified, auditory filtered, bimodal congruent with unmodified dialogue, bimodal congruent with filtered dialogue, bimodal incongruent with unmodified dialogue, bimodal incongruent with filtered dialogue), and 8 actor pairs. Auditory filtered stimuli were presented either with addition of brown noise (Experiment 1), or filtered with LPF (Experiment 2).

5.2.2 *Data analysis*

In order to take both accuracy and response speed into account, Inverse Efficiency (IE) scores were derived by dividing the response times by correct response rates separately for each condition; higher values of IE indicated worse performance. IE scores are a standard method for combining a reaction-times and accuracy measures of performance (Townsend & Ashby, 1978, 1983). This approach can be considered as 'corrected reaction times' that discount possible criterion shift or speed/accuracy tradeoffs (Spence *et al.*, 2001; Röder *et al.*, 2007). IE scores were submitted to repeated measures analysis of variance (ANOVA).

To analyze responses for incongruent bimodal stimuli we had to use a different approach, as there were no 'correct' responses for this stimulus. We chose an approach similar to one used by Collignon *et al.* (2008) and Petrini *et al.* (2010). We calculated a tendency to respond either 'angry' or 'happy' by subtracting the proportion of 'happy' judgements from the proportion of 'angry' judgements (pAngry - pHappy) in the four incongruent stimulus conditions (happy point-light display/angry unmodified voice; happy point-light display/angry degraded voice; angry point-light display/happy unmodified voice; and angry point-light display/happy degraded voice). The index, which varied between -1 (subject always responded 'happy') to 1 (subject always responded 'angry') was then submitted to ANOVA.

With each ANOVA we also calculated generalized *eta squared* (η^2_G) measures of effect size. η^2_G is preferential to *eta squared* and partial *eta squared* because it provides better comparability across between-subjects and within-subjects designs. It can also easily be computed from information provided by standard statistical packages (Olejnik & Algina, 2003; Bakeman, 2005). All p-values presented from posthoc Tukey analysis were given after adjustment for the multiple comparisons.

5.2.3 *Results for Experiment 1*

'Filtering' refers to brown noise in Experiment 1

The IE scores[2] averaged for each condition were submitted to a repeated measure ANOVA with 'emotion' (happy and angry), 'filtering' (unmodified, filtered) and 'stimuli' (visual, auditory, and bimodal congruent) as within factors. Primarily, we obtained a main effect of the factor 'stimuli' ($F(2,30) = 24.50$, $p < 0.001$, $\eta^2_G = 0.14$) demonstrating higher IE scores with visual stimuli compared to bimodal stimuli ($p < 0.001$), and auditory ($p < 0.001$) stimuli, although there was no difference between bimodal and auditory stimuli ($p = 0.07$), as seen on Figure 33. We also obtained a main effect of factor 'filtering' ($F(1,15) = 12.83$, $p < 0.001$, $\eta^2_G = 0.03$), as well as weak interaction between factors 'stimuli' and 'filtering' ($F(2,30) = 5.53$, $p \leqslant 0.01$, $\eta^2_G = 0.02$). Posthoc Tukey analysis showed narrowly better performance when participants viewed bimodal congruent unmodified rather than bimodal congruent filtered ($p \leqslant 0.05$) and auditory filtered ($p \leqslant 0.01$), but no difference in performance when viewing auditory unmodified rather than auditory filtered ($p = 0.18$). We found no main effect of factor 'emotion' ($F(1,15) = 1.65$, $p = 0.22$, $\eta^2_G = 0.01$), and no other

Visual worse than auditory and bimodal; no difference between auditory and bimodal; weak effect of brown noise filtering mainly for bimodal

2 See Figure A.7 and A.8 in the Appendix for mean accuracy and reaction times data from which IE scores were derived.

interactions between factors: 'emotion' and 'stimuli' ($F(2,30) = 0.75$, $p = 0.48$, $\eta^2_G = 0$); 'emotion' and 'filtering' ($F(1,15) = 0.88$, $p = 0.36$, $\eta^2_G = 0$); 'emotion', 'stimuli' and 'filtering' ($F(2,30) = 0.78$, $p = 0.47$, $\eta^2_G = 0$).

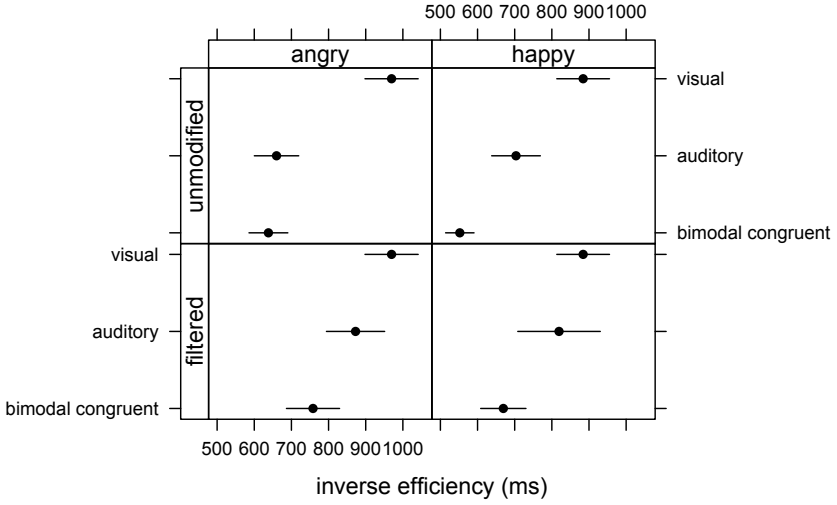


Figure 33: Mean IE scores and standard errors obtained in Experiment 1 for unimodal stimuli (auditory and visual) and congruent bimodal stimuli for both emotional expressions. The figure displays the results obtained with unmodified auditory stimuli (top row labelled *unmodified*) and brown noise filtered auditory stimuli (bottom row labelled *filtered*). IE scores are obtained by dividing RTs by correct response rates, thus eliminating any potential speed/accuracy tradeoff effects in the data; the lower the score the more efficient the performance.

We also looked at the tendency to choose happy or angry emotions when observers were presented with incongruent displays. The index calculated for incongruent displays, which varied between -1 (subject always responded 'happy') to 1 (subject always responded 'angry'), was analyzed by means of a two-way ANOVA with 'auditory emotion' (happy or angry) and 'auditory filtering' (filtered or unmodified) as within-subject factors. There was no significant effect of factor 'auditory filtering' ($F(1,15) = 0.86$, $p = 0.37$, $\eta^2_G = 0$), but we found a significant effect of factor 'auditory emotion' ($F(1,15) = 161.45$, $p < 0.01$, $\eta^2_G = 0.61$) and significant interaction between factors 'auditory emotion' and 'auditory filtering' ($F(1,15) = 48.95$, $p < 0.01$, $\eta^2_G = 0.20$). Posthoc Tukey analysis revealed that the index was more positive with 'visual happy/auditory angry unmodified' stimuli than with 'visual happy/auditory angry filtered' ($p < 0.05$), and that the index was more negative with 'visual angry/auditory happy unmodified' stimuli than with 'visual angry/auditory happy filtered'

With bimodal incongruent stimuli, participants oriented their emotion response towards the auditory rather than visual modality, although brown noise filtering weakened this tendency

stimuli ($p < 0.05$). Figure 34 clearly shows that for all bimodal incongruent combinations with unmodified and filtered auditory stimuli, participants oriented their response towards the auditory modality, but this tendency was weaker when the brown noise was present in the auditory signal. Posthoc Tukey analysis also revealed that there was a higher tendency to respond 'angry' for displays with angry unmodified audio rather than happy unmodified ($p < 0.01$) and happy filtered audio ($p < 0.01$). Participants also showed a tendency to respond 'happy' for displays with happy unmodified audio rather than happy filtered audio ($p < 0.05$), but there was no such difference between displays with angry unmodified and angry filtered audio ($p = 0.09$). Figure 34 clearly shows that with both unmodified and filtered auditory stimuli, participants oriented their emotion response towards the auditory rather than visual modality.

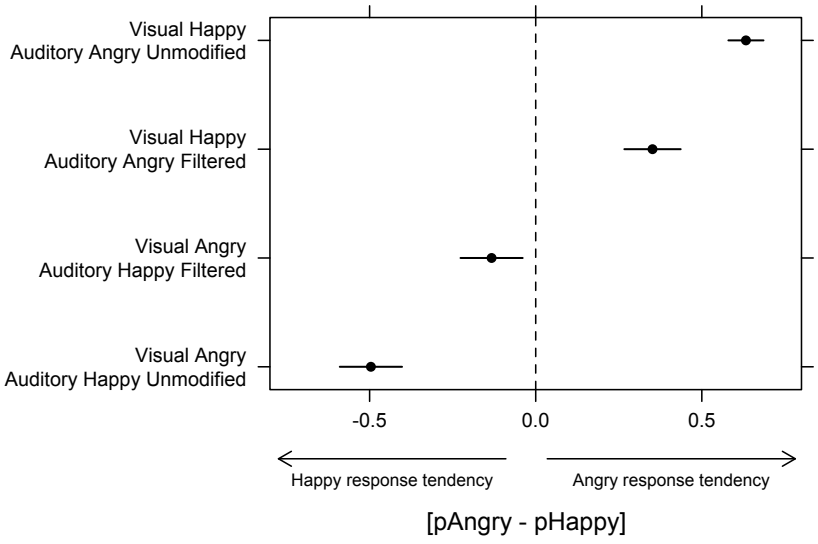


Figure 34: Bias to respond either 'happy' or 'angry' in bimodal incongruent conditions was estimated by subtracting the proportion of 'happy' responses from the proportion of 'angry' responses (pAngry - pHappy) in Experiment 1. Participants tend to report the emotion expressed in the auditory modality with both unmodified and brown noise filtered stimuli. Error bars represent one standard error of the mean.

5.2.4 *Results for Experiment 2*

'Filtering' refers to low-pass filtering in Experiment 2

The IE scores[3] averaged for each condition were submitted to a repeated measure ANOVA with 'emotion' (happy and angry), 'filtering' (unmodified, filtered) and 'stimuli' (visual, auditory, and bimodal congruent) as within factors. Primarily, we obtained a main effect of the factor 'stimuli' ($F(2,28) = 30.44$, $p < 0.001$, $\eta^2_G = 0.19$) demonstrating higher IE scores with visual stimuli comparing to bimodal stimuli ($p < 0.001$), and auditory ($p < 0.001$) stimuli, although there was no difference between bimodal and auditory stimuli ($p = 0.40$). We also obtained a main effect of factor 'emotion' ($F(1,14) = 5.61$, $p \leqslant 0.03$, $\eta^2_G = 0.02$) showing that angry displays were narrowly better recognized than happy displays (Figure 35). Finally, we obtained a main effect of factor 'filtering' ($F(1,14) = 5.11$, $p \leqslant 0.04$, $\eta^2_G = 0.01$) showing narrowly better performance with unmodified rather than filtered stimuli. We found no interaction whatsoever between any factors we analyzed: 'emotion' and 'stimuli' ($F(2,28) = 0.65$, $p = 0.53$, $\eta^2_G = 0$); 'emotion' and 'filtering' ($F(1,14) = 0.57$, $p = 0.46$, $\eta^2_G = 0$); 'stimuli' and 'filtering' ($F(2,28) = 1.87$, $p = 0.17$, $\eta^2_G = 0$); 'emotion', 'stimuli' and 'filtering' ($F(2,28) = 1.08$, $p = 0.35$, $\eta^2_G = 0$).

Visual worse than auditory and bimodal, no difference between auditory and bimodal; performance slightly worse with low-pass filtered stimuli comparing to unmodified

We also looked at the tendency to choose happy or angry responses when observers were presented with incongruent displays (Figure 36). The index calculated for incongruent displays was analyzed by means of a two-way ANOVA with 'auditory emotion' (happy or angry) and 'auditory filtering' (unmodified or filtered) as within-subject factors. There was no significant effect of factor 'auditory filtering' ($F(1,14) = 0.74$, $p = 0.40$, $\eta^2_G = 0$), but we found a significant effect of factor 'auditory emotion' ($F(1,14) = 57.25$, $p < 0.001$, $\eta^2_G = 0.70$) and significant interaction between factors 'auditory emotion' and 'auditory filtering' ($F(1,14) = 54.43$, $p < 0.001$, $\eta^2_G = 0.11$). Posthoc Tukey analysis revealed that the index was more positive with 'visual happy/auditory angry unmodified' stimuli than with 'visual happy/auditory angry filtered' ($p < 0.05$), and that the index was more negative with 'visual angry/auditory happy unmodified' stimuli than with 'visual angry/auditory happy filtered' stimuli ($p < 0.05$). With both unmodified and filtered auditory stimuli, the participants oriented their response towards the auditory modality, but this tendency was weaker when the filtering was present in the auditory signal, as seen on Figure 36.

With bimodal incongruent stimuli, participants oriented their emotion response towards the auditory rather than visual modality, although low-pass filtering weakened this tendency

3 See Figure A.9 and A.10 in the Appendix for mean accuracy and reaction times data from which IE scores were derived.

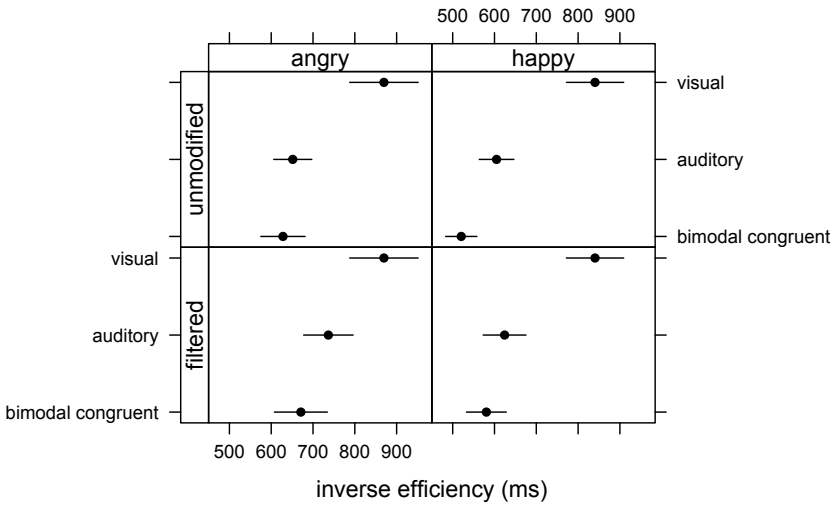


Figure 35: Mean inverse efficiency scores and standard errors obtained in Experiment 2 for unimodal stimuli (auditory and visual) and congruent bimodal stimuli for both emotional expressions. The figure displays the results obtained with unmodified auditory stimuli (top row labeled *unmodified*) and low-pass filtered auditory stimuli (bottom row labeled *filtered*). IE scores are obtained by dividing RTs by correct response rates, thus eliminating any potential speed/accuracy tradeoff effects in the data; the lower the score the more efficient the performance.

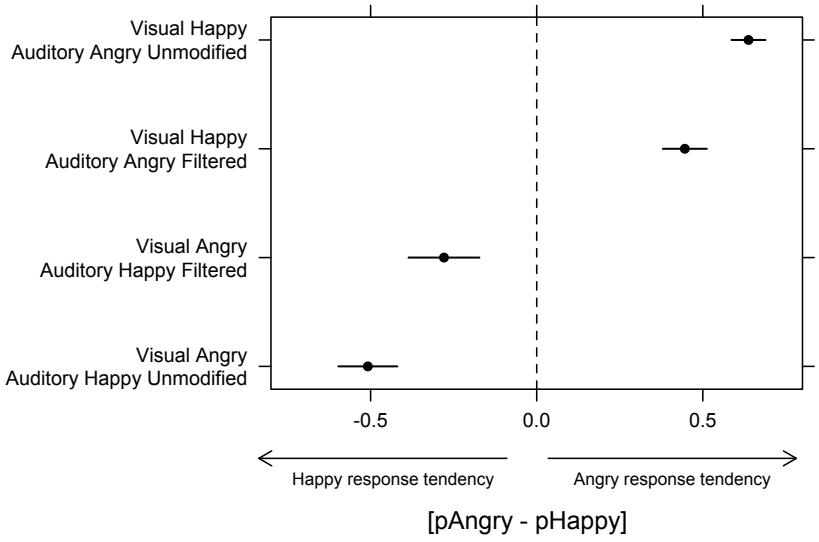


Figure 36: Bias to respond either 'happy' or 'angry' in bimodal incongruent conditions was estimated by subtracting the proportion of 'happy' responses from the proportion of 'angry' responses (pAngry - pHappy) in Experiment 2. Participants tend to report the emotion expressed in the auditory modality with both unmodified and degraded stimuli. Error bars represent one standard error of the mean.

5.3 COMPARISON OF FILTERING METHOD BETWEEN EXPERIMENT 1 AND 2

No difference between two filtering methods: brown noise (Exp. 1) and low-pass filter (Exp. 2)

Experiment 1 and 2 gave similar patterns of results. In both experiments participants' performance was better in auditory-only and bimodal congruent condition than the visual-only condition. In both experiments there was no difference between auditory and bimodal congruent conditions. When presented with bimodal incongruent stimuli, participants showed a tendency to report the emotion expressed in the auditory modality although in both experiments this tendency weakened with the filtering. Both filtering methods (brown noise [Exp. 1] and low-pass filtering [Exp. 2]) did lower the performance compared to unmodified stimuli. However, we were also interested in whether either of these two filtering methods was particularly better in degrading participants' performance. We conducted two-sample t tests on the averaged IE scores to establish whether there was a difference in correct discriminations when participants were presented with the auditory condition filtered with a low-pass filter (Experiment 1) rather than brown noise (Experiment 2). Results showed that there was no significant difference in participants' performance between the two filtering methods ($t(10) = 1.69$, $p = 0.12$), as seen in Figure 37.

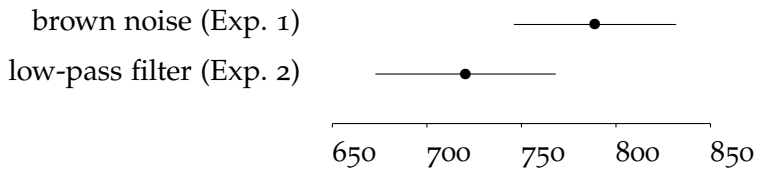


Figure 37: Comparison of mean inverse efficiency scores and standard errors obtained between Experiment 1 and 2 for two filtering methods. Error bars represent one standard error of the mean.

5.4 EXPERIMENT 3: FOCUS ON MODALITY

In Experiment 3, we requested the participants to pay attention to only one modality at a time to ascertain whether any multimodal effects found in Experiments 1 and 2 were due to automatic processes and would not disappear when participants were asked to ignore one of the two modalities. The underlying idea was that if audio- visual integration operates in an automatic fashion, multisensory influence should occur even if the participants only focus their attention to-

wards one single modality (de Gelder & Vroomen, 2000; Vroomen & de Gelder, 2000). As Collignon *et al.* (2008, p. 132) explain:

> This (procedure) could be related to a kind of "emotional Stroop" where the automatic nature of integration in bimodal emotion induces an inability for participants to focus on only one modality, even if instructed to do so.

5.4.1 *Methods*

5.4.1.1 *Participants*

Sixteen participants were recruited for Experiment 3: 6 female and 10 male, with a mean age of 22.7 years, ranging from 18 to 36 years. All participants were English speakers and UK born. All reported normal hearing and normal or corrected-to-normal vision. All participants were naive to the purpose of the study and had no prior experience with point-light display movies or images. The study received ethical approval from the University of Glasgow's Faculty of Information and Mathematical Sciences Ethics Review Board and every participant signed a consent form.

5.4.1.2 *Stimuli*

The stimulus set used in Experiment 3 was exactly the same as in Experiments 2 described in detail in Section 5.2.1.2. In Experiment 3 audio was filtered with LPF - exactly the same as in Experiment 2.

5.4.1.3 *Design & Procedure*

In Experiment 3 participants also performed an emotion identification task but were explicitly asked to focus their attention on one sensory modality at a time, ignoring the other modality. We used this procedure to test whether there would be any multisensory interaction in the processing of emotions even if the participants' attention was focused on a single modality (de Gelder & Vroomen, 2000). As a result we introduced two separate focus blocks in Experiment 3: a visual and an auditory block. The visual block included 2 emotions (happy, angry), 5 stimulus types (visual, bimodal congruent with unmodified dialogue, bimodal congruent with filtered dialogue, bimodal incongruent with unmodified dialogue, bimodal incongruent with filtered dialogue), and 8 actor pairs. The auditory block included 2 emotions (happy, angry), 5 stimulus types (auditory filtered, bimodal congruent with unmodified dialogue, bimodal congruent with filtered dia-

logue, bimodal incongruent with unmodified dialogue, bimodal incongruent with filtered dialogue), and 8 actor pairs. The participants were presented with a total of 480 stimuli. Each focus block (i.e. auditory and visual) consisted of 240 stimuli which included three repetitions of 80 stimulus conditions randomly interleaved within three separate blocks. Before the exposure to the visual focus block, participants were instructed to focus their attention on the visual displays and ignore the audio. Respectively, before exposure to the auditory focus block, participants were instructed to focus their attention on the audio and ignore the visual displays. The order of visual and auditory blocks was counterbalanced across participants.

5.4.2 *Data analysis*

As in Experiment 1 and 2, Inverse Efficiency (IE) scores were derived by dividing the response times by correct response rates separately for each condition. IE scores were submitted to repeated measures analysis of variance (ANOVA) and with each ANOVA we also calculated generalized *eta squared* (η^2_G) measures of effect size. Refer to Section 5.2.2 for details of the analysis.

5.4.3 *Results for Experiment 3*

Adding incongruent visual information to either unmodified or filtered audio made the participants' performance worst comparing to the congruent bimodal condition

We conducted two separate analyses with the data obtained from Experiment 3. In the first analysis we wanted to examine how filtering affected judgements in bimodal conditions, depending on whether participants attended to visual or auditory modalities. Therefore we only included both unimodal conditions, as well as congruent and incongruent bimodal filtered conditions. The IE scores[4] averaged for each condition were submitted to a repeated measures ANOVA with 'attention' (attend visual, attend auditory), and 'stimuli' (unimodal, bimodal congruent filtered, bimodal incongruent filtered) as within factors (Figure 38). We only found a significant main effect of factor 'stimuli' ($F(2,30) = 7.41$, $p < 0.001$, $\eta^2_G = 0.06$), but no significant effect of factor 'attention' ($F(1,15) = 0.5$, $p = 0.49$, $\eta^2_G = 0$) and no interaction between factors 'attention' and 'stimuli' ($F(2,30) = 2.02$, $p = 0.15$, $\eta^2_G = 0.01$). Posthoc Tukey analysis revealed that while there was no difference in IE when comparing auditory-only to the bimodal conditions ($p = 0.17$), there was a difference between the two bimodal conditions ($p \leqslant 0.01$). Therefore, the visual information in the bimodal condi-

4 See Figure A.11 and A.12 in the Appendix for mean accuracy and reaction times data from which IE scores were derived.

tion did have an effect on the performance increasing IE when the visual information was incongruent, as seen in Figure 38. This effect occurred despite asking participants to attend to the auditory signal and ignore the visual, as well as to attend the visual signal and ignore the auditory.

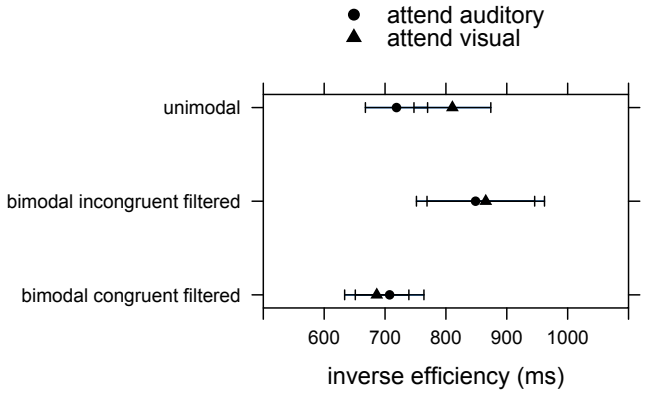


Figure 38: Mean IE scores and standard errors obtained in Experiment 3 for unimodal, and congruent and incongruent bimodal filtered stimuli. Circles display performance when participants were instructed to attend the emotion expressed aurally, while triangles display performance when participants were instructed to attend the emotion expressed visually.

In a second analysis for Experiment 3 we wanted to examine how congruency in bimodal unmodified stimuli affected the judgements, depending on whether participants attended to visual or auditory modalities. The IE scores[5] averaged for each condition were submitted to a repeated measure ANOVA with 'attention' (attend visual, attend auditory), and 'stimuli' (bimodal congruent unmodified, bimodal incongruent unmodified) as within factors (Figure 39). We found a significant main effect of factor 'stimuli' ($F(1,15) = 8.13$, $p \leqslant 0.01$, $\eta^2_G = 0.08$), but no interaction between factors 'attention' and 'stimuli' ($F(1,15) = 2.46$, $p = 0.14$, $\eta^2_G = 0.01$). The results replicated what we found for the bimodal filtered stimuli, in that there was a difference between the two bimodal conditions in the same direction. That is, adding visual incongruent information to the sound made the participants' performance worse (greater IE). However, because the reliability of the sound with unmodified stimuli was much higher than in the filtered case, there was also a significant effect of attended cue (i.e. 'attention': $F(1,15) = 7.39$, $p \leqslant 0.02$, $\eta^2_G = 0.05$). That is, partici-

5 See Figure A.13 and A.14 in the Appendix for mean accuracy and reaction times data from which IE scores were derived.

pants attended more to the sound than vision in the condition where the sound reliability was much higher (Figure 39).

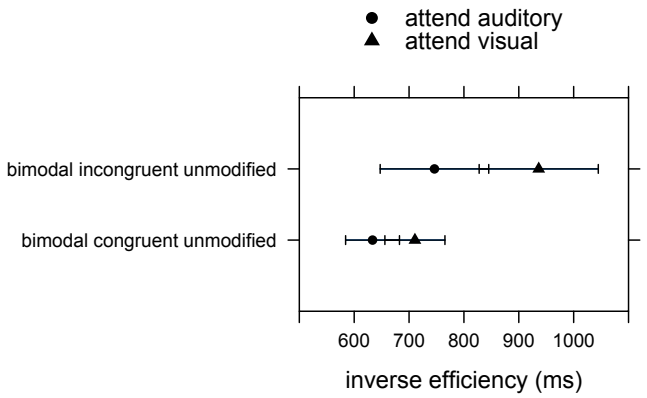


Figure 39: Mean IE scores and standard errors obtained in Experiment 3 for congruent and incongruent bimodal unmodified stimuli. Circles display performance when participants were instructed to attend the emotion expressed aurally, while triangles display performance when participants were instructed to attend the emotion expressed visually.

5.5 EXPERIMENT 4: CUE COMBINATION

In Experiment 4, we investigated whether combining visual and auditory cues reduces sensory uncertainty for high-level factors such as perceived emotions. To investigate cue combination quantitatively, we first measured the variances associated with visual and auditory estimation of the "angriness" of the interaction. We then used these measurements to construct a maximum-likelihood integrator. This was the same approach as the one used by (Ernst & Banks, 2002) with haptic and visual cues, and by Petrini *et al.* (2012) with haptic and auditory cues. Different levels of "angriness" were achieved by manipulating the frame rate of the visual (body movement) and auditory (voice) displays.

5.5.1 *Methods*

5.5.1.1 *Participants*

A total of 16 participants were recruited for the experiment, 9 female and 7 male, with a mean age of 24 years, ranging from 17 to 38 years. All participants were English speakers and UK born. All reported normal hearing and normal or corrected-to-normal vision. All par-

ticipants were naive to the purpose of the study and had no prior experience with point-light display movies or images. The study received ethical approval from the University of Glasgow's Faculty of Information and Mathematical Sciences Ethics Review Board and every participant signed a consent form.

5.5.1.2 *Stimuli*

A single, bimodal, high-intensity angry display has been used in Exp. 4

Initially, we applied a very similar procedure of stimulus selection to the one described in Section 5.2.1.2 of this Chapter, but we only selected a single angry display that was identified with an accuracy of 85% or higher. We were specifically looking for displays which received similar number of correct judgements for both visual and auditory modalities during the experiment described in Chapter 3. Since preliminary results described in Chapter 4 demonstrated a high accuracy of judgements with the auditory displays, we decreased the reliability of the auditory target. To do so, we applied the low-pass filtering (LPF) method used during experiments described in Section 4.2.2 of Chapter 4. Therefore, the auditory signal was always presented in the filtered format.

To manipulate the 'angriness' of the clips, we changed the speed of the movement and voice. Speed has been frequently reported in the studies looking into the factors determining whether movement or voice is attributed as less or more angry (Pollick *et al.*, 2002; Scherer, 2003; Roether *et al.*, 2009) . The speed was manipulated indirectly by parametrically manipulating the frame rate for both visual and auditory components of the selected angry display. The original display was coded with the frame rate of 60 frames per second (fps), in Audio Video Interleave (*.avi*) movie clip format. We parametrically decreased the frame rate to speed up the video and audio (48 fps, 51 fps, 54 fps and 57 fps), or increased the frame rate to slow it down (63 fps, 66 fps, 69 fps, 72 fps). The change of frame rate was conducted using scripts written in *FFmpeg* and *MEncoder*[6]. Table 5 summarizes all levels of comparison stimuli we created.

5.5.1.3 *Design & Procedure*

Participants were tested in a dark room, with only a small lamp to illuminate the keyboard. They were seated approximately 65 cm from a 21" Cathode Ray Tube (CRT) monitor with resolution of 1024 by

6 *FFmpeg* is an open source, cross-platform solution to record, convert and stream audio and video. *MEncoder* is an open source command line video decoding, encoding and filtering tool, and can convert audio and video clips into a variety of compressed and uncompressed formats using different codecs.

frame rate (fps)	rel.frame rate (fps)	stimuli speed
48	-12	fastest
51	-9	
54	-6	↑
57	-3	
60	0	standard
63	+3	
66	+6	↓
69	+9	
72	+12	slowest

Table 5: Different levels of stimuli created for Experiment 4. First column shows different frame rates in frames per second (fps), second column - relative frame rate (in fps) in relation to standard, and the last columns highlights which levels of stimuli were perceived as fastest and slowest as a result of frame rate manipulation.

768 pixels, and 60Hz refresh rate. Point-light displays subtended a maximum visual angle of approximately 8.5 degrees in height and 6 degrees in width. Participants wore headphones (Beyer Dynamic DT Headphones), with an intensity at the sound source of 60 dB. We used PsychoPy Psychophysics Software (Peirce, 2007) to program the experiment, present the displays and collect the responses.

Participants were instructed that they would be presented with pairs of short clips (around 3 seconds each) of two people engaged in angry conversation. They were told that both clips would be presented either as visual point-light displays, auditory dialogues or audio-visual combinations between point-light displays and dialogues. After viewing a single pair of clips, participants were asked to respond which of the two clips was 'angrier' and to press '1' or '2' on the keyboard corresponding to the clip they thought was 'angrier'. They were also instructed to wait until both clips from every pair were displayed, but then be fast in giving their answer.

A single pair of presented clips matched one of the five different conditions, as shown in Figure 40. One clip was always a standard stimulus (unmodified frame rate), and another one was a comparison stimulus (variable frame rate) but the nature of standard and comparison stimuli varied depending on the condition. In two unimodal conditions participants were presented with a pair of voice

dialogues (auditory condition) or a pair of point-light interactions (visual condition) where the frame rate in the standard stimuli was not modified but the frame rate in the comparison stimuli varied between nine different levels (Figure 40a). In the bimodal congruent condition participants were presented with a pair of clips with combined voice and movement, where the frame rate of both voice and movement in standard stimuli was not modified but the frame rate of both voice and movement in the comparison stimuli was varied across nine levels (Figure 40b).

In two bimodal incongruent conditions, auditory and visual cues provided by the standard stimuli were in conflict, but the frame rate averaged to the same level as the standard stimuli in the bimodal congruent condition. In one bimodal incongruent condition the voice of the standard stimuli had a frame rate equal to 60+6 fps, and movement equal to 60-6 fps (Figure 40c). In the other incongruent condition it was the opposite: the voice had a frame rate level equal to 60-6 fps, and movement equal to 60+6 fps (Figure 40d). The comparison stimuli in both bimodal incongruent conditions was the same as in the bimodal congruent condition that is - a bimodal congruent stimulus where the frame rate of both voice and movement varied across nine levels.

In summary, participants were presented with a total of 240 displays across 5 separate blocks: visual-only, auditory-only, bimodal congruent, bimodal incongruent 1 with slow audio and fast movement, and bimodal incongruent 2 with fast audio and slow movement. Each block consisted of 45 displays that included 9 randomly presented conditions, each from a different frame rate comparison level. Each condition was repeated 6 times and randomized across the block. The order of block presentation was counterbalanced across participants.

5.5.2 *Results for Experiment 4*

We first ran the analysis for each participant separately. Psychometric functions were fitted to the proportion of 'angrier' responses given by each participant as a function of comparison stimulus relative frame rate (Figure 41). The estimate of each individual's function's mean (i.e., the point at which the psychometric function cuts the 50% of 'angrier' responses) indicated the Point of Subjective Equality (PSE). The "angriness" discrimination threshold was given by the standard deviation of the psychometric function (i.e., the slope of the function). For the fitting we used Psignifit 2.5.6 (Fründ *et al.*, 2011) - a

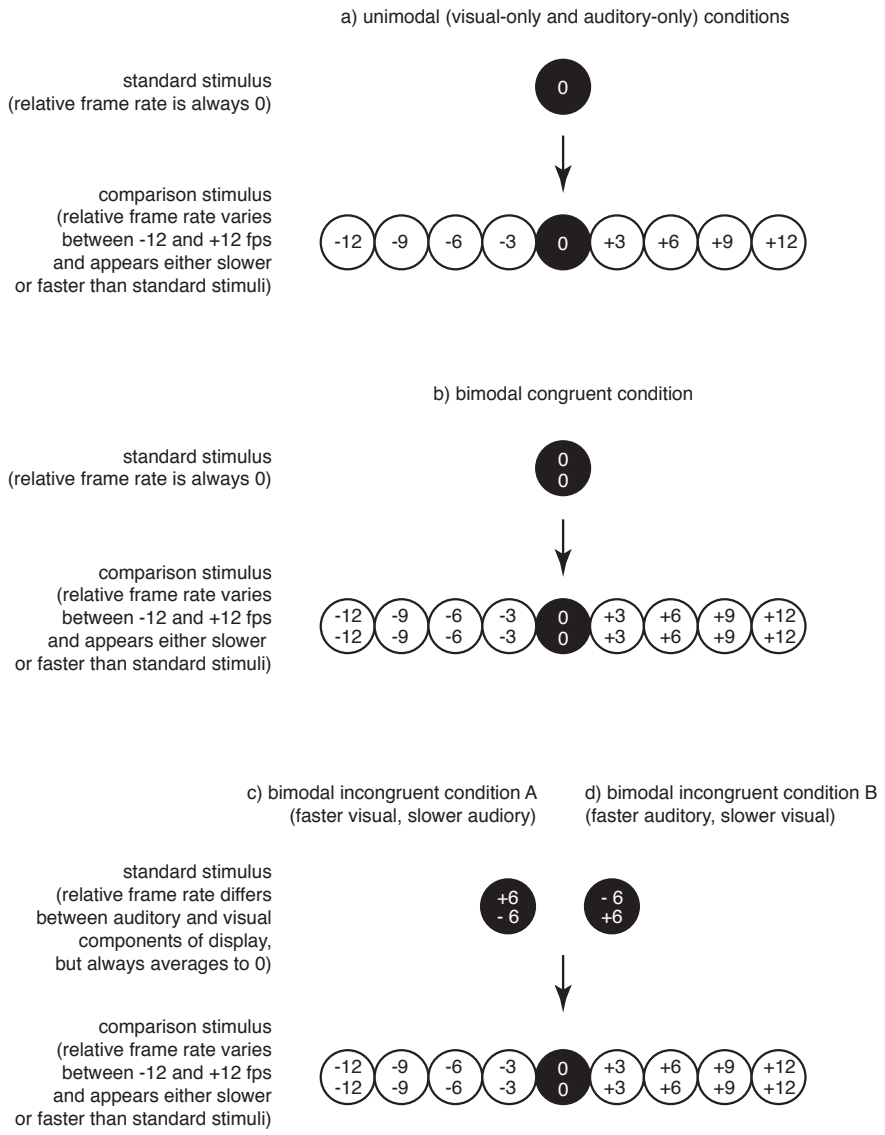


Figure 40: Illustration of different conditions used in Experiment 4. In unimodal (a) and bimodal congruent (b) conditions, standard stimuli always had relative frame rate equal to 0, while the frame rate of comparison stimuli varied between -12 and +12 fps. In case of bimodal incongruent conditions (c,d), standard stimulus had different relative frame rate between visual and auditory components, while the frame rate of comparison stimuli varied between -12 and +12 fps.

software package for Matlab (Mathworks, 2007) that implements the maximum-likelihood method.

On the basis of this initial analysis we excluded six participants due to their inability to do the visual-only task (panels coloured black on Figure 41). Specifically, they were not able to discriminate the standard stimuli from the comparison stimuli and their PSE or threshold fell outside of the chosen range of stimuli (see Figure 42 for the summary of individual thresholds). In the final step we averaged the fitting obtained from 10 non-excluded participants to get the averaged measure of PSEs (vertical lines on Figure 43) and discrimination thresholds (Figure 44).

The results showed that participants mainly used the auditory information and ignored the visual information. Figure 43 shows that the 'angriness' discrimination threshold determined by the slope was very similar for the auditory and bimodal congruent conditions, but much steeper for the visual condition. In short, the auditory information was much more reliable than the visual information to judge the level of 'angriness', and therefore participants ignored the visual signal. This is further confirmed by Figure 44 showing that both bimodal congruent and auditory-only discrimination thresholds were both well predicted by the optimal estimate, calculated by entering the unimodal discrimination thresholds into the maximum likelihood (MLE) model (see Figure 44 for details of threshold calculations). This observation was further supported by a series of one-tailed t-tests showing that although the bimodal threshold was lower than the visual threshold ($t(18) = -5.02$, $p < 0.01$) it did not differ from the auditory threshold ($t(18) = 0.49$, $p = 0.63$) or the predicted optimal bimodal threshold ($t(18) = 0.7$, $p = 0.49$).

We also looked at participants' responses to incongruent stimuli, summarized in Figure 45. Those results showed that participants were relying more on auditory rather than visual cues when they were presented with conflicting cues, as illustrated by a strong shift of the dotted line with triangles towards positive values on Figure 45.

5.6 DISCUSSION

The experiments described in this chapter were designed to apply our stimulus set to the study of audio-visual integration of emotional signals from body movement and voice in the context of dyadic interaction. We used an experimental design frequently utilized in emotional face-voice (de Gelder & Vroomen, 2000; Collignon *et al.*, 2008), and body movement and sound studies (Petrini *et al.*, 2010), but not

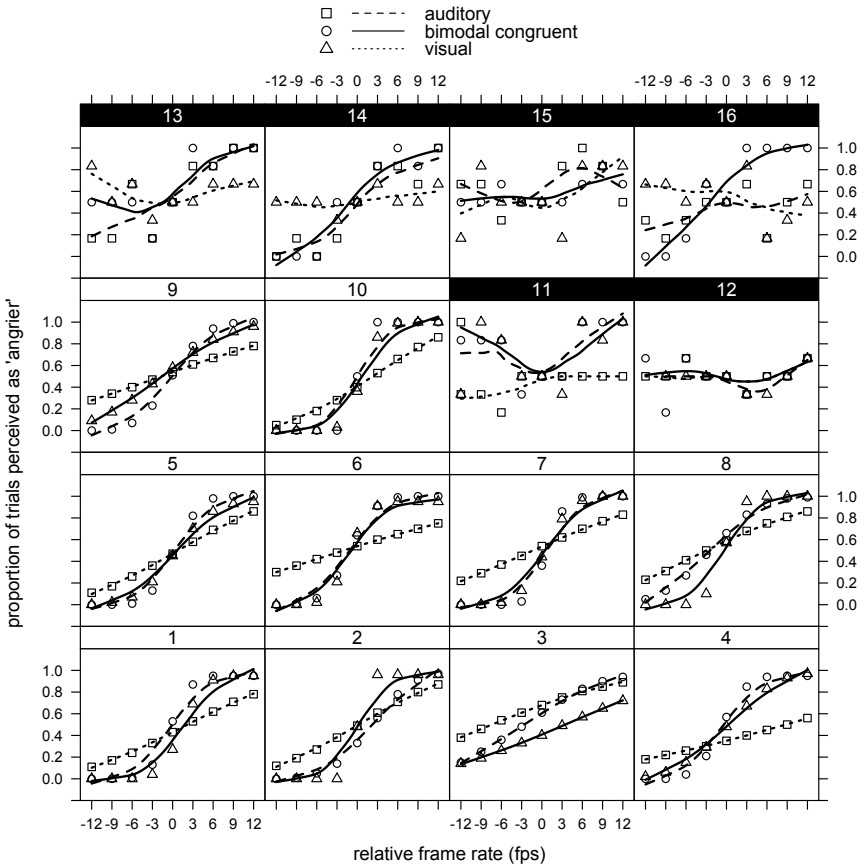


Figure 41: Proportion of trials in which a comparison was perceived as "angrier" than the standard stimulus is plotted against the relative frame rate (fps) of the comparison stimulus. Results are presented separately for each participant performing the auditory-only, visual-only and bimodal congruent condition (i.e. no conflict between the cues). The dashed curve with square symbols refers to the mean results for the auditory-only condition, the dotted curve with triangle symbols refers to the visual-only condition, and the solid curve with circle symbols - to the congruent bimodal condition. The point at which the psychometric function cuts the 50% point on the ordinate is the mean or PSE. The slope of the functions is used to estimate the standard deviation or 'angriness' discrimination threshold, such that the steeper the slope the lower is the variability and consequently the threshold. Black panels represent participants who were excluded from further analysis due to their inability to do the visual-only task.

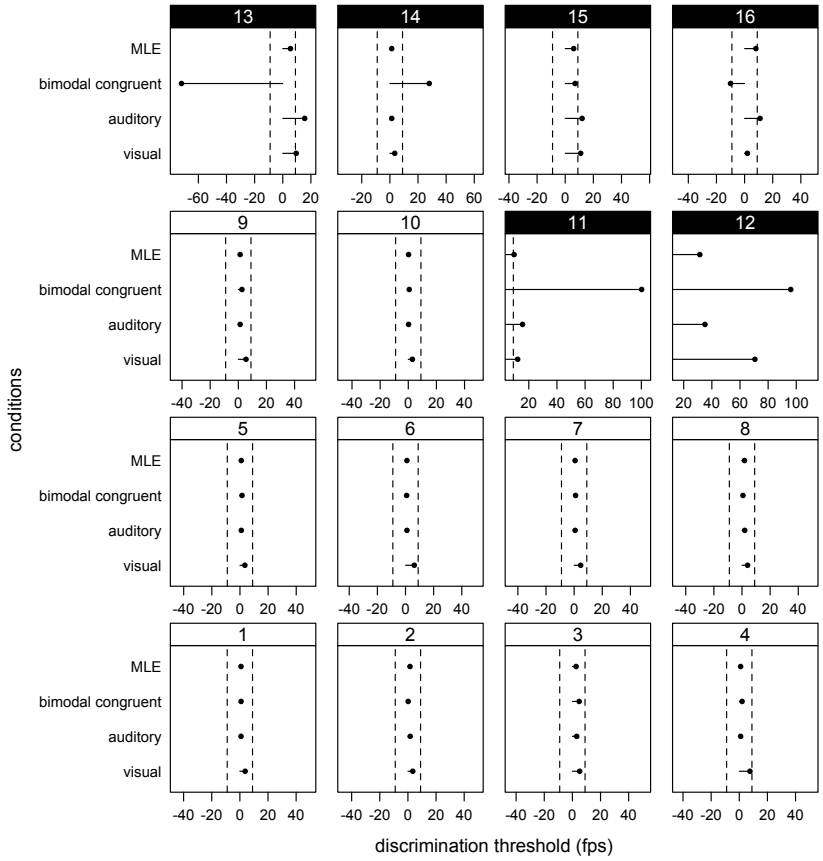


Figure 42: Discrimination thresholds for each participant for visual, auditory, and bimodal congruent conditions plotted together with the maximum likelihood (MLE) model predictions for the bimodal condition (MLE was calculated individually for each participant). Discriminations thresholds for participants 11 to 16 fell outside of the chosen range of stimuli (vertical dashed lines) and therefore those participants were excluded from further analysis (black panels).

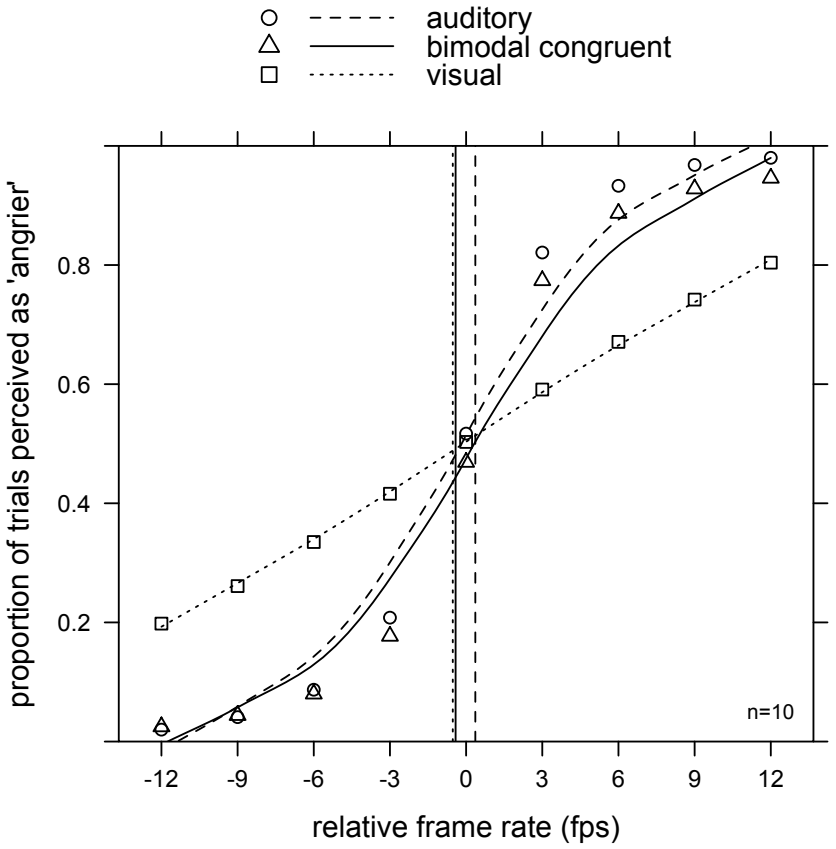


Figure 43: Proportion of trials in which a comparison was perceived as "angrier" than the standard stimuli is plotted against the relative frame rate (fps) of the comparison stimuli. Presented are mean results for the group of participants who were not excluded from the analysis (n=10) performing the auditory-only, visual-only and bimodal congruent condition (i.e. no conflict between the cues). The dashed curve with square symbols refer to the average results for the auditory-only condition, the dotted curve with triangle symbols refers to the visual-only condition, and the solid curve with circle symbols - to the congruent bimodal condition. The point at which the psychometric function cuts the 50% point on the ordinate is the mean or PSE. The vertical lines indicate the average PSEs for specific conditions. The slope of the functions is used to estimate the standard deviation or 'angriness' discrimination threshold, such that the steeper the slope the lower is the variability and consequently the threshold.

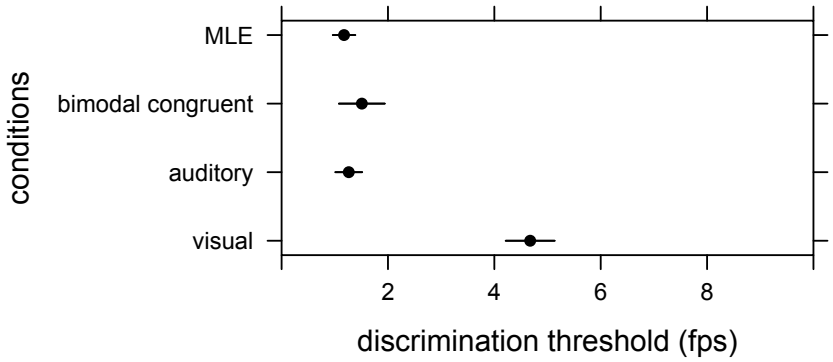


Figure 44: Mean discrimination thresholds for visual, auditory, and bimodal congruent conditions plotted together with the average MLE model predictions for the bimodal condition. The predicted bimodal threshold (δ_{AV}) was calculated individually for each participant, and then averaged, by entering the individual auditory (δ_A) and visual (δ_V) thresholds into the equation $\delta^2_{AV} = \frac{\delta^2_A \delta^2_V}{\delta^2_A + \delta^2_V}$. Error bars represent one standard error of the mean.

used before with body movement and voice. In Experiments 1 and 2, the participants were required to discriminate between happy and angry interactions either displayed aurally, visually or audio-visually, in a congruent (same emotion in the two modalities) or incongruent way (different emotion in the two modalities). This method allowed us to investigate whether the presentation of audio-visual congruent stimuli improves the participants' performance, and which modality dominates in the conflicting condition. Since we observed a higher accuracy of judgements with auditory stimuli in our previous experiments (see Chapters 3 and 4), we used brown noise (Experiment 1) and low-pass filtering (Experiment 2) to filter voice dialogues in order to adjust the reliability of auditory stimuli to a level similar to that of visual stimuli. Although these two methods were used in the studies described in Chapter 4 to filter voice stimulus, and were found to decrease the accuracy of participants' judgements, we wanted to examine them in a bimodal context rather than a unimodal context. This was motivated by the potentially different nature of the interactions between auditory and visual signals in relation to these different filtering methods.

In Experiment 3, we used the same procedure and stimuli as in Experiments 1 and 2, but we explicitly asked participants to focus on just one modality at a time (i.e. visual or auditory). Experiment 3 was conducted to ascertain whether any multimodal effects found

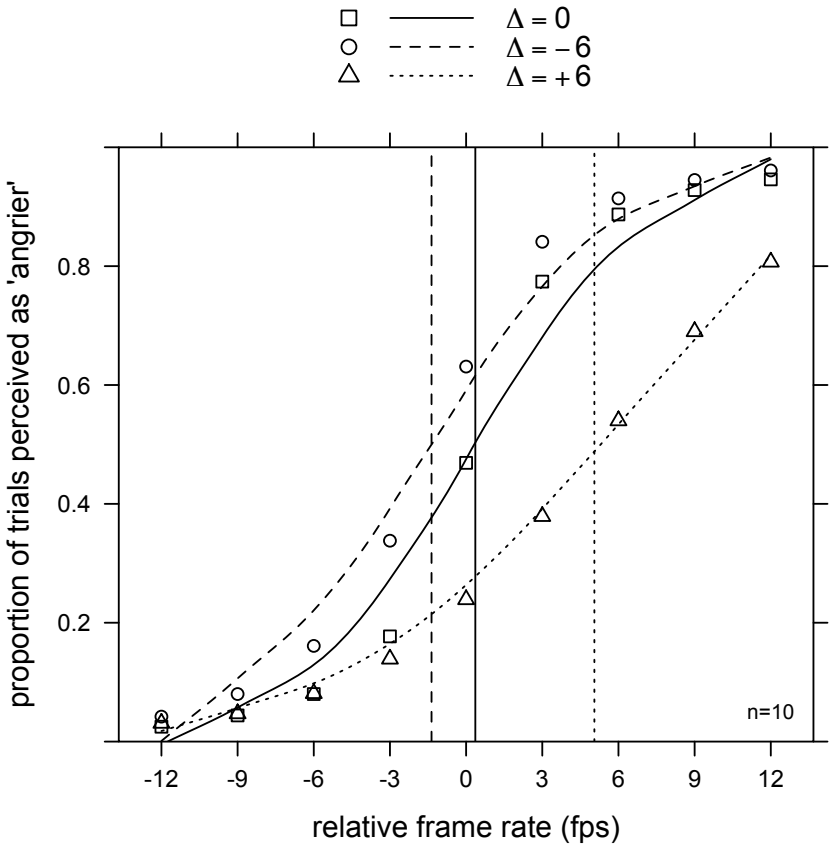


Figure 45: Average results for the group of participants performing one bimodal congruent and two bimodal incongruent conditions. The levels of cue conflict for the standard stimuli are represented here as -6, 0, and +6 fps for the visual and +6, 0, and -6 for the auditory. A shift of the dotted line with triangles toward +6 indicates that participants are relying more on the auditory information, whereas a shift toward -6 indicates that they are relying more on the visual. The opposite is the case for the dashed line and circles. The solid line refers to the congruent bimodal condition (zero conflict between the cues), as in the same condition in Figure 43.

in Experiments 1 and 2 were due to automatic processes and would not disappear when participants were asked to ignore one of the two modalities (Massaro & Egan, 1996; de Gelder & Vroomen, 2000). Experiment 4 concluded our attempts to apply the stimulus set to multisensory studies by using a cue combination paradigm. Specifically, we examined whether combining visual and auditory cues reduces sensory uncertainty for high-level factors such as perceived emotions in a social context. For this purpose, we used only a single, highly salient angry display where the frame rate was parametrically increased or decreased and presented to participants within a similar methodological framework to that of Ernst & Banks (2002).

In the first two experiments, we found the best performance in emotion judgements for both unmodified bimodal congruent and unmodified auditory displays. Participants' performance was worst when they viewed visual displays without the voice dialogue. Our findings are in line with the prediction of 'inverse effectiveness' which states that the outcome of multimodal integration is inversely proportional to the effectiveness of the relevant stimuli (Stein & Meredith, 1993; Collignon *et al.*, 2008; Petrini *et al.*, 2010). Namely, we failed to find a significant difference between the auditory and bimodal congruent condition because the auditory information was highly effective in delivering the intended emotions.

When participants viewed bimodal incongruent displays (different affect expressions in both modalities), they followed emotion in the auditory rather than the visual modality. This tendency to follow emotion in the auditory modality was weaker when the audio dialogues were filtered (in both Experiments 1 and 2). These results indicated that decreasing the reliability of auditory stimuli increases the salience of visual stimuli, although not sufficiently for observers to shift their judgements entirely to visual information.

Filtering the auditory signal with either brown noise (Exp. 1) or a low-pass filter (Exp. 2) had some effect on participants' performance, but the effect was limited to bimodal conditions. Specifically, when filtered audio dialogues were presented on their own they were easily identified, but in conditions where filtered audio dialogues were combined with a visual signal, filtering had some effect on performance. However, it showed that filtering had a more 'disruptive' effect on the general dominance of the auditory signal. In contrast, Collignon *et al.* (2008) and de Gelder & Vroomen (2000), who used faces and filtered voices, found that filtering the audio gradually shifted participants' judgements towards the emotion represented by the visual signal. This 'disruptive' rather than 'shifting' influence of visual in-

formation on the dominant auditory signal persisted in Experiment 3 when we asked participants to focus only on a single modality at a time. The addition of a visual incongruent signal to the filtered auditory signal made the participants' performance worse, and, with the filtered auditory, participants attended as much to vision as to sound because the sound reliability had decreased. However, because the reliability of the unmodified auditory signal was higher than for the filtered signal, participants attended more to the sound than the vision in the condition where the sound reliability was much higher. In short, participants kept using the sound much more in their emotion judgements even when the sound was filtered. However, this did not depend on the participants' ability to ignore the visual information since they were less efficient when the visual cue expressed a different emotion to the auditory cue and this was true for both filtered and unfiltered conditions.

In Experiment 4, we further tested whether participants were able to optimally integrate visual and auditory signals when we parametrically manipulated the displays' frame rate. This resulted in the video and audio components of the displays appearing either slower or faster. Participants had to make judgements as to which of the two presented displays was 'angrier' in a procedure described in Section 5.5.1.3. Psychometric functions were fitted to the proportion of 'angrier' responses given by each participant as a function of comparison stimulus relative frame rate. In short, those results further confirmed what we found from Experiments 1, 2 and 3, that participants mainly used the auditory information and ignored or underestimated the visual information. Even after excluding participants whose discriminations thresholds were beyond the estimated limits, we still found that both bimodal congruent and auditory-only, but not visual-only, discrimination thresholds were equally well predicted by the maximum likelihood (MLE) model. Correspondingly, we found that participants were relying more on auditory rather than visual cues when they were presented with bimodal displays with conflicting cues. Results from Experiment 4 further support the notion that voice is a highly salient and dominant cue comparing with body movement when observers make judgements of emotions watching other people interacting.

Studies on the perception of emotions from a single actor's face and voice show that observers make their judgements based mainly on faces rather than voices, although such dominance can shift depending on the visual and auditory reliability of the stimuli (de Gelder & Vroomen, 2000; Collignon *et al.*, 2008). In contrast, our results sug-

gest that voice plays a particularly important role in the perception of emotional social interactions. It is possible that people rely more on auditory cues when making judgements while observing dyadic interactions, because the voice is a more immediate and a clearer source of emotional information while body movement is more complex and ambiguous, and the presence of two actors might also have an impact on this. As we discussed in previous chapters, there are inherent differences between the perception of a single person versus dyadic interactions and this relates to such factors as the ability of action prediction when viewing dyads (Manera *et al.*, 2011, 2013), difference in judgements from the perspective of facing the agent and observing the agent (Schouten *et al.*, 2013), and detection of meaningful actions (Neri *et al.*, 2006). Faces and voices are frequently experienced during direct, face-to-face conversation facing the speaker. In contrast, when we observe interaction from a distance, we may not pay attention to facial cues or those cues might not be as prominent as body movement and especially voice. It is only a speculation to conclude that there is a specific hierarchy of multisensory cues depending on the mode of communication. One way of testing it would be to establish a three-cue paradigm using faces, voices and body movement, to better understand how these three cues contribute to emotion identification in a social context.

One drawback of our methodological approach in Experiment 4 was that we did not degrade the auditory signal using a different or more controlled method. We could use both brown noise and low-pass filtering combined together as a method of filtering. We could also adjust the auditory signal-to-noise ratio of the voice dialogues to lower the accurate discrimination of the stimuli presented only aurally in a more controlled manner, for example using the QUEST procedure (Watson & Pelli, 1983). However, it is important to stress that after the validation experiments described in Chapter 3, we picked a subset of displays that were judged with a very high accuracy of around 85%. The stimulus subset has also been selected to match the reliability between the visual and auditory components of each display. In Chapter 4, we also filtered the auditory signal to decrease its reliability and we concluded that such a manipulation was sufficient to match the visual and auditory signals.

To summarize, we systematically created and validated a data set of happy, angry and neutral audio-visual social interactions between two actors. We found that the auditory signal dominated the visual signal in the perception of emotions from social interactions. Although participants' judgements were better in the bimodal condition, the

performance was similar in both bimodal and auditory-only conditions. When participants watched emotionally mismatched bimodal displays, they predominantly oriented their responses towards the auditory signal rather than the visual signal. This auditory dominance persisted even when the reliability of an auditory signal was degraded with brown noise or low-pass filtering, although visual information had some effect on judgements of emotion when it was combined with a degraded auditory signal. Our results suggest that when judging emotions from observed social interactions, we rely primarily on vocal cues from conversation rather than visual cues from body movement. More studies are required to further examine the nature of this auditory salience but our studies described in this chapter are a first effort to use an audio-visual stimulus set for the study of emotional social interactions.

6

GENERAL DISCUSSION

6.1 OVERVIEW

The initial goal of the project described in this thesis was to create a stimulus set that incorporated body movement and voice dialogue between two actors with the interactions being happy, angry or neutral at different levels of intensity. After this set was created, the aim was to validate it and to examine the perception of happy and angry emotions from movement and voice when observers watched social interactions between others. In this social context, we examined how different methods of movement and voice distortion, such as inversion and scrambling of movement or filtering of voice, influenced judgements of happiness and anger. Finally, we focused on the application of the developed stimulus set to examine how voice and body movement are integrated when observers watch social emotional interaction, and how the emotional congruence of voice and body movement influences perception. As a part of multisensory studies, we also examined whether combining visual and auditory cues reduces sensory uncertainty for high-level factors such as perceived emotions in a social context.

6.2 MULTIMODAL STIMULUS SET FOR THE STUDY OF EMOTIONAL SOCIAL INTERACTIONS

To achieve the goals outlined in this thesis, the first step was to create a novel stimulus set and this process was described in Chapter 2. This novel stimulus set was constructed using methodology typically utilized in capturing and creating point-light displays (Dekeyser *et al.*, 2002; Troje, 2002; Ma *et al.*, 2006) and voice dialogue databases (Scherer & Ceschi, 1997; Douglas-Cowie *et al.*, 2003; Ververidis & Kotropoulos, 2006). This was achieved by using a passive optical motion capture system synchronized with an audio capture system. The final stimulus set consisted of 256 unique clips that present happy, angry and neutral interactions at low, medium and high levels of emotional intensity.

There are a number of advantages and features that made the stimulus set described in Chapter 2 unique and suitable for broader exper-

imental application, both within and outside the goals of this thesis. Primarily, we captured both body movement and voice in a synchronized manner and therefore provided the first data set to study audiovisual interactions. Existing stimulus sets that utilize point-light displays typically include only a single actor's actions (Dekeyser *et al.*, 2002; Vanrie & Verfaillie, 2004) and frequently use only a part of the body such as hands (Busso *et al.*, 2008). Other types of sets with emotional body movements focus on full-body displays (de Gelder & Van den Stock, 2011) which make it difficult to dissociate body movement from other cues such as clothing or body shape. In the case of voice, there is a large number of very naturalistic datasets (Scherer & Ceschi, 1997; Roach *et al.*, 2009; Douglas-Cowie *et al.*, 2003). However, most of the existing sets with visual and auditory stimuli are unimodal (i.e. only voice samples without movement or only movement without voice). These include a small number of existing stimulus sets with dyadic point-light interaction that do not include a voice component and do not employ different levels of emotional intensity (Manera *et al.*, 2010). Our set raises the bar by including both components of dyadic, emotional point-light interactions as well as natural emotional voice dialogues.

Because of the inherent nature of the data provided by the motion capture technique, we were able to design the stimulus set in a manner that allowed flexible manipulation of different parameters of movement, and the same applies (to some extent) for voice dialogues. For example, features of body movement and voice such as orientation of actors on the screen, size of point-lights, or intelligibility of dialogue conversation can be easily modified, making our set a useful tool for a wide range of studies focusing on perception of emotions from body movement and voice. Finally, we also attempted to make the set as realistic as possible by repeating interactions, using experienced and inexperienced actors, and creating role-plays that helped actors to better embed themselves in the role they played.

6.3 HIGH SALIENCE OF ANGER AND THE EFFECT OF EMOTIONAL INTENSITY

After the stimulus set was developed and processed, the next goal was to establish whether happy and angry interactions portrayed in the displays were identified accurately by the observers. We used different levels of emotional intensity and therefore we expected that for lower intensity displays observers would make more identification errors. Through a series of experiments described in Chapter 3, we

showed that happy and angry emotions were recognized accurately from the point-light displays of two actors engaged in conversation as well as from voice dialogues. These results confirmed earlier studies by Clarke *et al.* (2005) and Dittrich *et al.* (1996). Initially, the main difference in our result was that happy displays were identified more accurately than angry displays, although interestingly this effect was strongly related to the level of emotional intensity. Accuracy of angry judgements increased between low, medium and high intensity, but such an effect was not observed in the case of happy judgements. This effect of emotional intensity on angry rather than happy judgements may be related to an overall higher sensitivity to anger. For example, Chouchourelou *et al.* (2006) examined detection of emotions when point-light walkers were embedded in a mask of moving dots and they found that among the five different emotions (afraid, happy, angry, sad and neutral), the greatest visual sensitivity was found for angry walkers. Ikeda & Watanabe (2009) examined whether the relationship between gait detection and emotion detection from biological motion differed between angry and happy walkers. The authors found significant correlations with gait detection performance for anger detection but not necessarily for happiness detection, implying that the detection of anger may be more strongly linked to explicit gait detection. There is also a large body of evidence which emerges from studies on detection of facial expression, showing that angry faces are detected more accurately and quickly among other emotional faces (Hansen & Hansen, 1988; Fox *et al.*, 2000; Eastwood *et al.*, 2001; Ohman *et al.*, 2001). They found a much stronger correlation with gait detection performance for anger rather than happiness detection, concluding that anger may be more strongly linked to explicit gait detection.

It is important to note that when we selected a high intensity subset of stimuli for the further experiments described in Chapters 4 and 5, angry interactions were always identified more accurately than happy interactions. These arguments link to the popular threat advantage hypothesis stating that observers preferentially attend threats (Pichon *et al.*, 2011; Ohman *et al.*, 2001). A number of studies argue that humans have a high sensitivity to negative affect as an indicator of threat (Pichon *et al.*, 2008). Strong sensitivity to anger displays have been shown for emotional faces (Fox *et al.*, 2000), voices (Green *et al.*, 2010) and body movement (Ikeda & Watanabe, 2009). In the perception of emotions from voice, anger is generally best recognized, followed by sadness and fear (Scherer, 2003). As Scherer (2003) points out, there is a clear adaptive advantage in being able to threaten foes in anger

over large distances - something for which the voice is ideally suited. A number of brain imaging studies have also revealed clear activation in regions of the brain linked to autonomic reactions and motor responses related to defensive behaviours (e.g. Fox *et al.*, 2000; Ohman *et al.*, 2001; Pichon *et al.*, 2008). Our results support such accounts but the results described in Chapter 3 also suggest that emotional intensity played an additional role in judgements of anger. It is possible that the threat detection mechanism only activates when the emotional intensity of stimuli reaches a particular level. Because of the high sensitivity to angry stimuli, observers are also sensitive to small changes in the emotional intensity of this stimulus. Some parameters in low and medium intensity angry displays were more ambiguous to the observers. For example, such features as high velocity or acceleration of body movement, and high intensity or pace of voice were found to be related to angry expressions (Scherer, 2003; Ma *et al.*, 2006). It is possible that low and medium intensity angry displays lacked some of these cues and such displays were misidentified as happy because observers detected specific cues that were irrelevant to angry expressions. The effect of emotional intensity was further evident when participants rated their confidence in their emotion judgements in the experiments described in Chapter 3. Confidence increased with higher intensity displays. Such an effect was not observed for happiness further supporting the argument that the mechanism for processing happy interactions may be less specific than that processing angry interactions.

6.4 THE AMBIGUOUS ROLE OF ACTORS' EXPERIENCE

We also examined whether actors' experience had any effect on the perception of emotions. Here, we found that emotions portrayed by less experienced actors were perceived more accurately than those portrayed by experienced actors but only in cases when visual displays were presented rather than voice or a combination of visual and voice. In the debriefing, participants also reported that some of the expressions were exaggerated. This result supports earlier suggestions by Busso (2008), and Roether *et al.* (2009) that experienced actors may exaggerate their emotional expression because they are more aware of the components of specific emotional expressions. However, it is difficult to speculate on the exact nature of this effect as it is clear that some actors without acting experience also exaggerate their emotional expressions, especially in high intensity conditions. A more intuitive explanation of the lower accuracy for experienced actors'

movement may be related to the notion of surface and deep acting. Surface acting involves "pushing down" one's authentic expression of self in favour of an emotional mask, while deep acting involves "pumping up" by trying to bring the required and one's true feelings into alignment (Brotheridge & Lee, 2003). Drama studies show that affective delivery ratings are negatively related to surface acting but positively related to deep acting (Grandey, 2003). In our capture sessions, we used a realistic role-play scenario but it is possible that not every actor immersed themselves fully into the actual experience of emotions (deep acting). Anecdotally, we observed that some pairs of actors were more engaged in acting than others. Speculatively, it is possible that the difference we observed between interactions portrayed by experienced and inexperienced actors may have been due to a variance in the level of engagement of the actors rather than a variance in the actors' experience.

However, we found that both experienced and inexperienced actors were recognized with equal levels of accuracy when voice was present as a cue to emotions in auditory and audio-visual groups. Such a result indicates that voices might be easier to use as a tool for emotional expression by both experienced and inexperienced actors. One explanation could be that the voice is easier to control than body movement when faking emotion in the acting context. The body is a complex system to control, especially in a relatively constrained environment such as the motion capture system we used to record the interactions. It is difficult to validate this claim in the context of existing studies, although existing research in deception indicates that, overall, there are no differences between face, voice and body movement in deception efficiency (Ekman *et al.*, 1991).

It was also interesting to find that participants tended to grossly underestimate or overestimate the number of actors they think were used to create the displays, but only in the group where they viewed only body movement without sound. Their estimates were much more accurate when voice dialogue was available. This has some interesting potential to explore in the context of identity studies, suggesting that the voice is a much better cue for identity than body movement. It is possible that movement on its own does not provide clear identity information and it is easier to establish identity from the voice. Indeed, point-light display studies on identity recognition have shown that observers can recognize their friends only by their point-light walk display (Cutting & Kozlowski, 1977) and that observers can learn and subsequently recognize individuals from their arm movement (Hill & Pollick, 2000). However, it is clear that detection of different,

unfamiliar identities from body movement only is a much more difficult task. In contrast, the voice has a large number of acoustic properties that figure prominently in the literature on talker recognition (Bricker & Pruzansky, 1976; Laver & Trudgill, 1979). These properties include the fundamental frequency of phonation, the typical frequencies of vocal tract resonances, the structure of glottal harmonics, and the fine-grained power spectra of nasals and vowels (Sheffert *et al.*, 2003). For example, vocal pitch is an extremely salient component of vocal quality and accounts for most of the variance in studies on talker recognition (Matsumoto *et al.*, 1973; Gelfer, 1988). In this context, the result that participants were most accurate in establishing the number of actors in the auditory and audio-visual groups were not surprising.

6.5 INVERSION, SCRAMBLING, VIEWPOINT AND INTERACTIONS

After establishing the baseline for the perception of emotions using the newly developed stimulus set, we wanted to examine the effect of distortion of point-light displays and voice on emotion perception. Previous studies indicated that observers can recognize a range of emotions from just a few point-lights representing a single person, although this recognition is impaired if the point-lights are scrambled (e.g. Chouchourelou *et al.* 2006), inverted (e.g. Clarke *et al.* 2005) or presented from different viewpoints (e.g. Jokisch & Troje 2003). Similarly, listeners could recognize emotions from voice, although methods applied to speech such as white noise (You *et al.*, 2006; Hammerschmidt & Jürgens, 2007) or low-pass filtering (Rogers *et al.*, 1971; Frick, 1985; Knoll *et al.*, 2009) impair emotion recognition. We know that observation of two people makes a difference compared to observation of a single actor, but we do not know whether this social aspect would impair recognition of emotions if the stimuli become degraded. The motivation behind the series of experiments described in Chapter 4 was to replicate the studies that examined the effect of distortion of point-light displays and voice, but to replicate them in a dyadic rather than single actor setting. We also wanted to further examine whether the intelligibility of dialogue also influences the judgements by either distorting or removing intelligible content completely and preserving only the prosodic features.

As a method of stimulus distortion, we used inversion and scrambling of point lights, as well as viewpoint shift. We found that inversion and scrambling decreased the accuracy of emotion judgements by approximately 15-20%. Even with such a decrease in accuracy, par-

ticipants were still able to recognize emotions correctly around 60% of the time compared to 75-80% with unmodified stimuli. Clarke *et al.* (2005) also found a lower accuracy of identification for upside-down displays but still they got the identification accuracy above the level of chance. Similarly, using point-light displays of professional dancers conveying different emotions, Dittrich *et al.* (1996) found that inversion of the displays reduced biological-motion performance close to, but still significantly above, the level of chance. These results clearly show that the effect of inversion in our study was less evident than the findings of Dittrich *et al.* (1996) and Clarke *et al.* (2005) with recognition of inverted angry and happy point-light displays being well above the level of chance. One suggestion is that the presence of a second agent made it easier for observers to recognize emotion from the inverted orientation. Such a view would be consistent with suggestions by Neri *et al.* (2006) and Manera *et al.* (2011) that observation of communicative interactions improves the detection of agents and the emotions they express. Our results on inversion also suggest that the effect of local cues is more powerful than assumed. Inversion might distort perception of walking direction, but kinetic features such as velocity and direction are still highly recognisable, even from severely distorted movies.

Scrambling had a similar effect on overall accuracy compared to inversion, decreasing it by approximately 15-20%, although participants were still able to identify the emotions above the level of chance. In a series of psychophysical studies, Chouchourelou *et al.* (2006) reported that observers were able to identify emotions of happiness, fear, sadness and anger above the chance level from scrambled point-light walkers. Ikeda & Watanabe (2009) also showed that angry and happy point-light walkers can be detected behind a scrambled mask, and again detection of anger was stronger, a point which is also supported by our findings. Pollick *et al.* (2001) showed that even when the phase and position relationships were distorted, participants categorised the stimuli under correct emotions above the chance level despite the fact that the stimuli did not resemble humans. Thus, it seems that the detection of emotion from dynamic stimuli can be sustained by kinematics of the body such as the local velocity signals, although an importance of such factors as global extrinsic motion and global spatial structure (Thurman & Lu, 2013), and existence of basic perceptual mechanisms such as "life detectors" (Troje & Westhoff, 2006) should not be ignored.

Considering the effect in our study, participants found it easier to identify angry, happy and neutral displays from a side view rather

than an oblique view. Placing these results in the context of existing research is challenging mainly because of the different modes of viewpoint investigated by other researchers. The majority of studies on the effect of viewpoint argue for an overall advantage of frontal views compared to profile (i.e. side) and half-profile views (Mather & Murdoch, 1994; Troje, 2002; Troje *et al.*, 2005). However, all these studies focus on the perception of a single actor rather than the interaction between two actors. When looking at the interactions, it is intuitively plausible to assume that a side view would be the optimal viewpoint. When we observe people interacting around us, they do not face us, but rather they face each other. Our result supports this intuitive assumption that a side viewpoint is optimal in the observation of interactions between other people.

We also found that inversion and scrambling had a different effect depending on the viewpoint from which those displays were presented. In the case of both inverted and scrambled displays, there was no difference between the two viewpoints. It is possible that viewpoint had an effect in the case of unmodified displays, because the oblique view presented a more complex motion compared to a side view, which is more commonly encountered when we observe other people interacting around us. The viewpoint became less relevant when inversion impaired configural processing, and scrambled displays lacked coherent global structures. It is also possible that from oblique viewpoint, there was an overlap and occlusion between points while actors were dynamically interacting. One explanation may be related to a low-level perceptual crowding effect (Nishida, 2011; Ikeda *et al.*, 2013). Crowding is an observation that the identification of closely spaced objects is more difficult than for isolated objects (Bouma, 1970; Flom *et al.*, 1963). Therefore, closely overlapping points between two point-light actors created a scene that was harder to decode for the observers. Because such an effect does not occur with the side viewpoint where the actors were visually separated, it is possible that crowding contributed to lower emotion identification accuracy in oblique viewpoint. But higher-level perceptual and cognitive mechanisms cannot be ignored in the context of emotion perception from dyadic interaction in different viewpoints. In the context of single agent studies, Coulson (2004) also highlights the role of occluding effects of particular viewpoints on some classes of stimuli. For example, the more closed and downward looking postures for disgust, fear and sadness appear smaller from the front, and present less information to the viewer than from side and rear views. The overall preference for frontal views suggests that attributing emotion to a body posture

is a great deal easier when the person adopting the posture is facing the perceiver. Such an orientation, while not necessarily ideal for perceiving the three-dimensional relationships between body segments, may nonetheless enhance recognition due to its interpersonal significance. While optimal perception from a frontal viewpoint applies to interpersonal perception, it is different in the context of intrapersonal perception. Our finding clearly suggests that a side view may be the optimal viewpoint for intrapersonal identification of emotions.

6.6 THE DOMINANCE OF VOICE OVER MOVEMENT IN SOCIAL INTERACTIONS

One of the major goals of the research conducted as part of this thesis was to apply our stimulus set to the study of multisensory perception of emotional social interactions. This was the initial reason why we not only captured the movement of interacting actors but also the short dialogue exchange between them. After the series of validation experiments described in Chapter 3, we found that participants presented with dialogues or presented with a combination of dialogues and movement were much more accurate in their judgements of emotions than participants presented only with movement. These results suggested that voice dialogue during interactions was much more salient than movement, but initially we thought it was due to intelligibility of dialogues. Therefore, in Chapter 4, we degraded the dialogues using low-pass filtering or addition of brown noise to observe whether these methods would be sufficient to lower highly accurate judgements of emotional dialogues. Indeed, both filtering methods degraded the reliability of the voice dialogues to the level similar to that of the unmodified visual signal. With this audio-visual set, matched for reliability, we conducted the four experiments described in Chapter 5. The experiments were specifically aimed to test our stimulus set for the multisensory studies and to investigate how people integrate emotional signals from movement and voice when they observe social interaction. Such studies have been previously conducted using a single actor's face and voice (de Gelder & Vroomen, 2000; Kreifelts *et al.*, 2007; Collignon *et al.*, 2008), a single actor's full-light body movement and voice (Pichon *et al.*, 2008; Stienen *et al.*, 2011), and a single actor's full-light body movement and music (Petrini *et al.*, 2010), but not dyadic point-light interactions with full voice dialogues. Two aspects of our stimulus set allowed us to introduce novelty into the investigation of multisensory integration. First, the social context inherent in our stimulus set, and second, the realis-

tic content of our stimuli. Other authors typically tested multisensory integration with a stimulus that had relatively low ecological validity - it did not resemble real-life interactions. Our goal was to emulate real-life situations using simulated role-play scenarios, snapshots of intelligible dialogues, and a 3-second conversation that was sufficient for a single exchange of information. Following a research paradigm similar to Collignon *et al.* (2008) and Petrini *et al.* (2010), we introduced a condition where there was emotional incongruence between body movement and voice. Angry voices were combined with happy body movements, and happy voices with angry movements. We also introduced conditions where participants were explicitly asked to focus either on visual or auditory information, ignoring the other. This served to establish whether the irrelevant modality affected the judgements.

We repeatedly found that the auditory signal dominated the visual signal in the perception of emotions from social interactions. In summary, participants were less accurate, slower and worse in discriminating emotions when making judgements from body movement only, compared to conditions when body movement was combined with dialogue or dialogue was presented on its own. Additionally, there was no audio-visual facilitation when participants were presented with combined movement and voice - they performed equally well with only auditory information. However, some effects were noticed when participants were presented with emotionally incongruent combinations of displays (i.e. happy dialogue with angry movement or happy movement with angry dialogue). In such incongruent conditions, visual information seemed to have a disruptive influence on the interpretation of emotions from the auditory signal. Specifically, adding incongruent visual information to dialogue made the participants' performance worse compared to a congruent combination between movement and voice. This confirmed earlier results by de Gelder & Vroomen (2000) and Collignon *et al.* (2008) that an incongruent combination of two signals would cause some disruption in the emotion interpretation of those signals. However, such conditions were still not sufficient to break the strong influence of the auditory signal on participants' judgements. This was further confirmed in the cue combination experiment where we asked participants to judge which display was more angry and we indirectly manipulated the speed of actions (by parametrically changing the displays' frame rate). Participants were making optimal decisions equally likely when they were presented with voice only, as well as a combination of voice and movement. The auditory information was much more reliable than

the visual information to judge the level of 'angriness' and therefore participants ignored the visual signal. Both bimodal congruent and auditory-only discrimination thresholds were well predicted by the optimal (ideal observer) estimate, while the threshold for the visual signal was much higher. In short, we found that when judging emotions from observed social interactions, we rely primarily on vocal cues from conversation, rather than visual cues from body movement.

It is not easy to explain those results in the context of existing results due to the lack of studies using social and emotional multimodal stimuli. In the context of face and voice literature, it's clear that facial expression is typically a more dominant signal then voice (Massaro & Egan, 1996; de Gelder & Vroomen, 2000; Collignon *et al.*, 2008; Jessen *et al.*, 2012). However, there are some indirect indicators from the literature on deception and nonverbal communication suggesting that the voice plays a leading role comparing to body movement when it comes to expression of emotions. Ekman *et al.* (1976) found that measures of hand movements and voice were interrelated but changed incongruently when a person shifted from honest to deceptive expressions. Specifically, the amount of symbolic hand movements decreased in deception, while pitch variance into high tones increased with deception, making the voice more accessible as cue as well as creating a discrepancy between voice and body movement. Moreover, studies on body movement and speech rhythm in social conversation clearly show that speakers tend to use their body movement to highlight specific aspects of their spoken messages (Dittmann & Llewellyn, 1969). Movement output and speech output were found to be quite closely correlated (Boomer, 1963). Renneker (1963, p. 155) described what he called speech-accompanying gestures, which "seek to complement, modify, and dramatize the meanings of words". Freedman & Hoffman (1967) separated what they called punctuating movements from other speech-related movements. In such contexts, it is possible that, in a conversational context, body movements play an accenting function to the voice - a claim also supported and suggested by Ekman (1965) regarding nonverbal behaviour in general. Referring back to our studies, such a finding from the study on nonverbal communication and speech would explain the high salience and dominance of dialogue over body movement when observers judged emotional interactions. It is plausible to say that the voice was a primary focus in decoding emotional signals when we observed interactions, while body movement played a secondary accenting and supportive role to the voice. More research is required to better understand this effect of voice saliency compared to body movement, but it suggest

the interesting possibility that there is a specific multisensory integration hierarchy between voice and body movement when we watch social interactions between people.

6.7 LIMITATIONS AND POTENTIAL FOR FUTURE RESEARCH

A few minor limitations were identified in the process of conducting the research described in this thesis. One can argue that it is a limitation that we used only male participants when capturing the stimulus set. Introducing female participants would increase the ecological validity of the stimulus set by creating an additional layer of interaction - between two females or a male and a female. The reason we decided to avoid more gender combinations was related to the fact that such combinations would introduce an entirely new level of complexity and intergender dynamics (see Feingold (1994) for metanalysis of research on personality differences between males and females). While this is an important and interesting topic, it was not within the scope of this thesis. We were more interested in establishing how observers decode basic emotional signals when watching interactions rather than focusing on higher-level factors of intergender interactions. There was also a practical constraint related to capturing more than ten actor pairs and we preferred to increase the variance by introducing different actors rather than more gender combinations. Finally, this topic has also been explored in earlier studies investigating gender identification from point-light displays (Pollick *et al.*, 2002, 2005; van der Zwan *et al.*, 2009; Poom, 2012). Still, it provides a clear potential for future research in the domain of cognitive social psychology to expand the stimulus set by additional gender combinations and to investigate the role of gender in the perception of emotional social interactions.

Another area for future research would be to investigate the role of cultural factors in the perception of emotion from body movement. In the research described throughout this thesis, we used actors and participants from only a single cultural context (i.e. British). However, one question is the extent to which expression and perception of emotion from movement are universal across cultures. Matsumoto (1992, 2010) argues that cultural variants affect the way we understand meaning in situations and alter the frequency of occurrence of emotion-inducing situations. Classic research on identification of emotions from facial expression shows that observers across cultures can accurately identify a range of basic emotions (Ekman & Friesen, 1971) but recent studies argue for cross-cultural differences (Jack *et al.*,

2009). Intuitively, there are vast differences between cultures in body expression; however, there has been limited research conducted into cross-cultural differences in the perception of emotions from body movement. There is a clear potential for future application of the stimulus set developed in this thesis in the cross-cultural context in perception of emotions. For example, the set could be used to test the differences in perception of emotion from movement and voice between Western Caucasian and East Asian observers. However, such a study introduces an entirely new level of challenges such as creating a stimulus set using participants representative of different cultures, and again such research was beyond the scope and realistic time frame of this thesis.

In relation to the dominance of the auditory signal in social interaction, one limitation of the cue combination experiment was that we did not parametrically manipulate the quality of voice dialogues with noise or filtering. Instead, we applied noise and filtering on a single level, matching auditory reliability with the visual signal in the experiment described in Chapter 4. The reason for this was that we wanted to keep a broad variety between the different types of emotional interactions we used. Parametric manipulation would enable us to better test the effect of different levels of noise on perception of emotions from voice dialogues, but it would prevent us from keeping the broad variety of stimuli.

In the analysis and validation conducted with the stimulus set, we decided to omit the kinematic and kinetic analysis of the movement, as well as the speech analysis of the voice. This omission was due to the constraint related to the huge variance of factors and variables related to body movement and voice. Besides the large number of interactions, an additional level of complexity related to the fact that we used dyadic interaction instead of a single actor. Analysis methods available in the research literature have only been performed on the kinematic models of a single actor (e.g. Montepare & Zebrowitz-McArthur 1988; Wallbott 1998; Atkinson *et al.* 2007). The joint angle trajectories define complex spatiotemporal patterns and countless possible features could be analysed in order to investigate how these trajectories change with emotion (Roether *et al.*, 2009). Adding interaction effects related to dyadic interaction poses a challenge that was simply beyond the technical scope of this thesis. Nevertheless, it is clear that the analysis method could be developed and applied to our stimuli to deconstruct the movement and voice patterns. One of the 'holy grails' in the study of emotional body expression is to pinpoint critical features that people use to make emotional judgments. Such

critical features have been established for facial expression by Ekman & Friesen 1971 who created *Facial Action Coding System* (FACS). FACS defines exact sets of muscle contraction that produce basic emotional facial expressions. Although there is no such detailed coding system for body movement, a current direction in social neuroscience focuses increasingly on decoding emotions from the patterns of movement (see reviews by de Gelder & Hortensius 2014 and Kleinsmith & Bianchi-Berthouze 2013). For example, a number of studies focused on finding local configurations of movements that produce particular emotional impressions in observers (e.g. Roether *et al.* 2009). While the stimulus set developed in this thesis is not sufficient to achieve such an ambitious goal, it is a good starting point in the attempt and potential approach to study patterns of detection of emotions from human movement. Instead of the approach frequently used by researchers who attempt to map the broad spectrum of emotional interactions, one could instead focus on simple affective differences between happy and angry interactions at different levels of intensity. Decoding visual and auditory patterns of such simple negative and positive interactions could be the first step to creating a methodology for studying the broader spectrum of emotions. Such a simplified approach could be easily applied in the security field where detection of threat is crucially dependent on differentiating between angry, happy and neutral interactions.

Studies on degrading point-light displays and voice described in Chapter 4 are also relevant to the security field. For example, an oblique viewpoint is a typical perspective from which a CCTV operator watches the street from a camera. As we clearly demonstrated in Chapter 4, interactions are harder to identify from an oblique viewpoint compared to a side viewpoint. In a practical sense, such results suggest that CCTV operators would have more trouble in detecting hostile actions when watching typical CCTV images, compared to situations when they watch a side view image. However, we also found that observers could accurately identify emotions to a high standard even if the interaction was presented from an oblique viewpoint in a scrambled or inverted format. This suggests that CCTV operators may detect hostile actions even when CCTV images are distorted with ambient light or poor weather conditions. The next question could ask what exact information CCTV operators use to make their judgements. Answering such questions would help to understand how people detect a threat when observing others interact. Research in this thesis may be a good starting point in terms of the methodological approach used to look at the complex social scene

and attempt to understand how people detect threat, anger, friendly conversations and other social signals.

Overall, our stimulus set provides a simple solution to study the perception of body movement and voice in a social context. Previous studies have shown that information in point-light displays is sufficient for a clear recognition of communicative actions (Dittrich, 1993; Clarke *et al.*, 2005; Lorey *et al.*, 2012; Manera *et al.*, 2011) as well as identification of emotions (Dittrich *et al.*, 1996; Pollick *et al.*, 2001; Atkinson *et al.*, 2004). By testing our stimulus set for multisensory perception, we identified some of its limitations related to the high salience of voices, but this could be solved by applying more parametric methods of auditory filtering to make sure that the visual and auditory information is better matched for reliability. Nevertheless, we clearly highlight the powerful role of the voice as a social cue to emotions. Further studies are required to better understand how the voice contributes to our judgements of emotions in relation to cues from facial expression and body movement. Because our stimulus set emulates natural social scenes, it allows us to expand the investigation on movement and voice perception to a more realistic context. We already know that the presence of a second agent is an important factor in the perception of meaningful actions (Neri *et al.*, 2006) but we can gain a better insight into how emotional judgements are driven by the social interaction.

Another potential use of the stimulus set involves the computational and behavioural analyses of how emotion and identity are encoded and decoded from human movement and voice. Thanks to the flexibility related to the point-light display method of coding movement, the stimulus set can be easily adapted to neuroimaging studies focusing on social and emotional aspects of multisensory perception (Saygin *et al.*, 2008; Mendonça *et al.*, 2011). Because of its flexibility, such a social stimulus set would be ideal in examining various hypotheses related to autistic spectrum disorder (Moore *et al.*, 1997; Hubert *et al.*, 2007; Nackaerts *et al.*, 2012) and the mirror neuron system (Chaminade *et al.*, 2007; Centelles *et al.*, 2011).

6.8 CONCLUSION

At the beginning of this thesis, we used a simple pub example as a metaphor for the complexity of the social environment in which humans continuously observe interaction between each other and the richness of social stimulation. This thesis is an attempt to methodically isolate a simplified, short and emotional social interaction be-

tween two people, and examine how observers judge emotions when they watch others from movement and voice in different perceptual conditions. We can conclude from findings in studies with single actors that observers can very easily recognise happy and angry interactions even when they watch simplified, point-light interactions or hear short dialogues. Observers are better at identifying anger than happiness and this is also expressed by the fact that they are much more responsive to changes in emotional intensity with angry rather than happy displays. Observers' judgements of emotions of dyadic interactions are viewpoint dependent and can be distorted by worsening the reliability of the visual and auditory conditions. However, even if severe distortion methods such as inversion or scrambling of point-light location are applied, observers can still identify emotions above the level of chance. Additionally, observers judging emotions rely primarily on the voice rather than the body movement of observed interactions. This auditory dominance opens the possibility that body movement plays a secondary accenting and supportive role to voice when people watch emotional interactions between other people. It is hoped that the stimulus set, as well as the findings from this thesis, will give researchers a new toolbox and framework to examine various aspects related to perception of emotions in complex, multisensory, emotional social scenes.

A

APPENDIX

Name	Description and measurement unit
Body Mass	Patient mass (kg).
Height	Patient height (mm).
Ankle Width	The medio-lateral distance across the malleoli. Measure with patient standing, if possible (mm).
Knee Width	The medio-lateral width of the knee across the line of the knee axis. Measure with patient standing, if possible (mm).
LegLength	Full leg length, measured between the ASIS marker and the medial malleolus, via the knee joint. Measure with patient standing, if possible. If the patient is standing in the crouch position, this measurement is NOT the shortest distance between the ASIS and medial malleoli, but rather the measure of the skeletal leg length (mm).
Elbow Width	Width of elbow along flexion axis (roughly between the medial and lateral epicondyles of the humerus) (mm).
Hand Thickness	Anterior/Posterior thickness between the dorsum and palmar surfaces of the hand (mm).
Shoulder Offset	Vertical offset from the base of the acromion marker to shoulder joint centre (mm).
Wrist Width	Anterior/Posterior thickness of wrist at position where wrist marker bar is attached (mm).

Table A.1: Description of measurements taken from participants for Plug-in Gait model.

Label	Definition	Position on Patient
LFHD	Left front head	Left temple
RFHD	Right front head	Right temple
LBHD	Left back head	Left back of head
RBHD	Right back head	Right back of head
C7	7th cervical vertebra	On the spinous process of the 7th cervical vertebra
T10	10th thoracic vertebra	On the spinous process of the 10th thoracic vertebra
CLAV	Clavicle	On the jugular notch where the clavicles meet the sternum
STRN	Sternum	On the xiphoid process of the sternum
RBAK	Right back	Anywhere over the right scapula
LASI	Left ASIS	Left anterior superior iliac spine
RASI	Right ASIS	Right anterior superior iliac spine
LPSI	Left PSI	Left posterior superior iliac spine
RPSI	Right PSI	Right posterior superior iliac spine
LSHO	Left shoulder	On the acromio-clavicular joint
LUPA	Left upper arm	On the upper lateral 1/3 surface of the left arm
LELB	Left elbow	On the lateral epicondyle
LFRM	Left forearm	On the lower lateral 1/3 surface of the left forearm
LWRA	Left wrist marker A	At the thumb side on the posterior of the left wrist, close to the wrist joint
LWRB	Left wrist marker B	At the little finger side on the posterior of the left wrist, close to the wrist joint
LFIN	Left finger	Just proximal to the middle knuckle on the left hand
RSHO	Right shoulder	On the acromio-clavicular joint
RUPA	Right upper arm	On the lower lateral 1/3 surface of the right arm
RELB	Right elbow	On the lateral epicondyle approximating the elbow joint axis
RFRM	Right forearm	On the lower lateral 1/3 surface of the right forearm
RWRA	Right wrist marker A	At the thumb side of a bar on the posterior of the right wrist
RWRB	Right wrist marker B	At the little finger side of a bar on the posterior of the right wrist
RFIN	Right finger	Just proximal to the middle knuckle on the right hand.
LTHI	Left thigh	Over the lower lateral 1/3 surface of the left thigh in line with the hip and knee joint centres
LKNE	Left knee	On the flexion-extension axis of the left knee
LTIB	Left tibia	Over the lower 1/3 surface of the left shank
LANK	Left ankle	On the lateral malleolus along an imaginary line that passes through the transmalleolar axis
LHEE	Left heel	On the calcaneous at the same height above the plantar surface of the foot as the toe marker
LTOE	Left toe	Over the second metatarsal head, on the mid-foot side of the equinus break between fore-foot and mid-foot
RTHI	Right thigh	Over the lower lateral 1/3 surface of the right thigh
RKNE	Right knee	On the flexion-extension axis of the right knee
RTIB	Right tibia	Over the lower 1/3 surface of the right shank
RANK	Right ankle	On the lateral malleolus along an imaginary line that passes through the transmalleolar axis
RHEE	Right heel	On the calcaneous at the same height above the plantar surface of the foot as the toe marker
RTOE	Right toe	Over the second metatarsal head, on the mid-foot side of the equinus break between fore-foot and mid-foot

Table A.2: Anatomical location of markers for Plug-in Gait model.

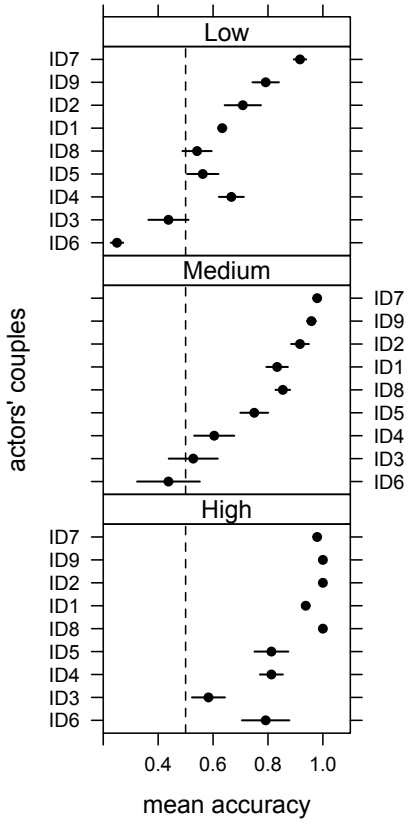


Figure A.1: Mean accuracy of emotion judgments for angry displays on low, medium and high intensity level with specific actors' couples (ID1-ID9) in auditory experiment in Chapter 3. The error bars represent one standard error of the mean and the dashed line represents the level of chance (0.5).

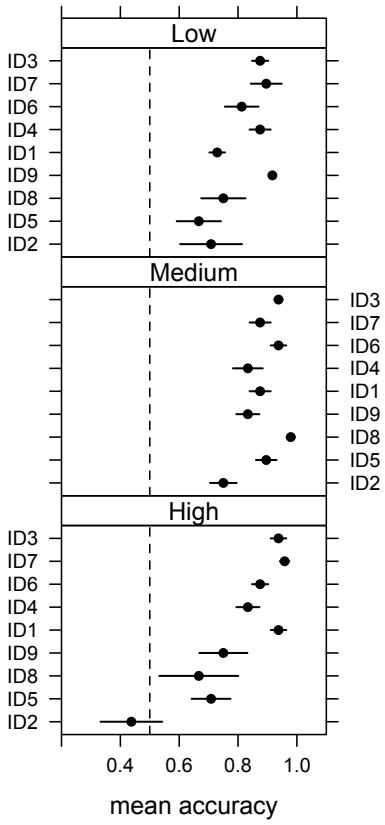


Figure A.2: Mean accuracy of emotion judgments for happy displays on low, medium and high intensity level with specific actors' couples (ID1-ID9) in auditory experiment in Chapter 3. The error bars represent one standard error of the mean and the dashed line represents the level of chance (0.5).

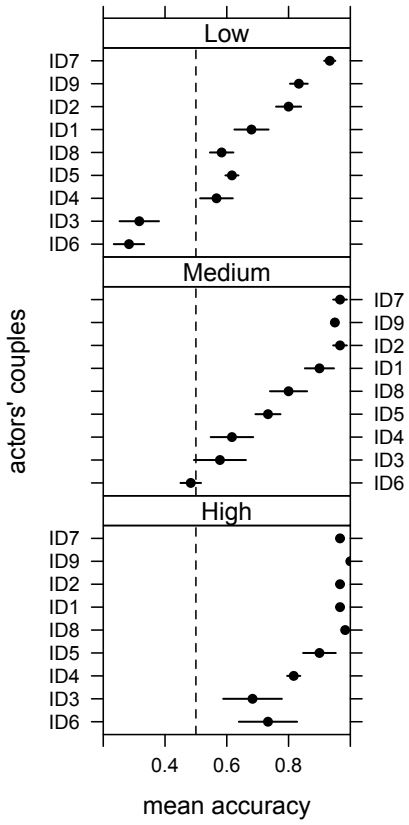


Figure A.3: Mean accuracy of emotion judgments for angry displays on low, medium and high intensity level with specific actors' couples (ID1-ID9) in audio-visual experiment in Chapter 3. The error bars represent one standard error of the mean and the dashed line represents the level of chance (0.5).

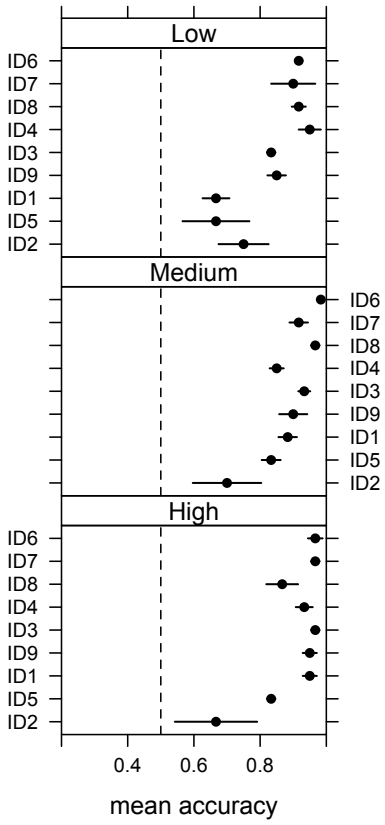


Figure A.4: Mean accuracy of emotion judgments for happy displays on low, medium and high intensity level with specific actors' couples (ID1-ID9) in audio-visual experiment in Chapter 3. The error bars represent one standard error of the mean and the dashed line represents the level of chance (0.5).

emotion	accuracy (%)			confidence score			display type	
	av	a	v	av	a	v	intensity	actors' ID
angry	100	100	93	8.73	8.50	8.60	High	ID1
	100	100	93	8.47	7.67	8.73	High	ID2
	100	100	87	9.00	8.08	8.80	High	ID3
	100	100	87	8.87	9.00	6.47	High	ID4
	100	100	87	7.13	7.75	7.87	High	ID5
	93	100	87	8.33	8.67	6.67	High	ID6
	93	92	100	7.13	8.33	7.13	High	ID7
	87	92	93	6.07	5.67	7.20	High	ID8
happy	100	100	93	6.73	6.92	6.47	Medium	ID1
	100	100	87	8.47	7.83	6.07	Medium	ID2
	100	92	93	8.47	6.67	7.67	High	ID3
	100	92	87	8.47	6.42	7.33	High	ID4
	100	92	87	6.93	6.42	5.73	Low	ID5
	93	100	93	6.87	6.08	6.80	Medium	ID6
	93	100	87	7.33	7.75	6.47	Medium	ID7
	93	92	93	7.53	7.67	5.67	Medium	ID8
neutral	53	58	39	5.20	3.67	4.80	-	ID1
	40	42	43	4.93	5.50	4.60	-	ID2
	47	50	33	5.33	4.83	4.27	-	ID3
	40	42	40	5.80	4.50	5.33	-	ID4
	40	58	57	4.20	4.58	4.93	-	ID5
	53	42	67	4.40	4.58	4.07	-	ID6
	40	45	35	4.93	4.67	4.93	-	ID7
	47	67	40	5.27	5.50	5.00	-	ID8

Table A.3: Summary of unique displays composing the stimuli subset used in Chapters 4 and 5. From the original stimulus set described in Chapter 2 and 3 we selected eight angry and eight happy displays that were identified with an accuracy of 85% or higher, and an average confidence rating of five or higher. Mean accuracy ratings are summarized based on the judgements obtained from three groups in validation experiments (av - audio-visual group, a - auditory group, v - visual group) described in Chapter 3. We also chose eight neutral displays that received an approximately equal number of happy and angry judgements in each experimental group described in Chapter 3 (between 40-60%).

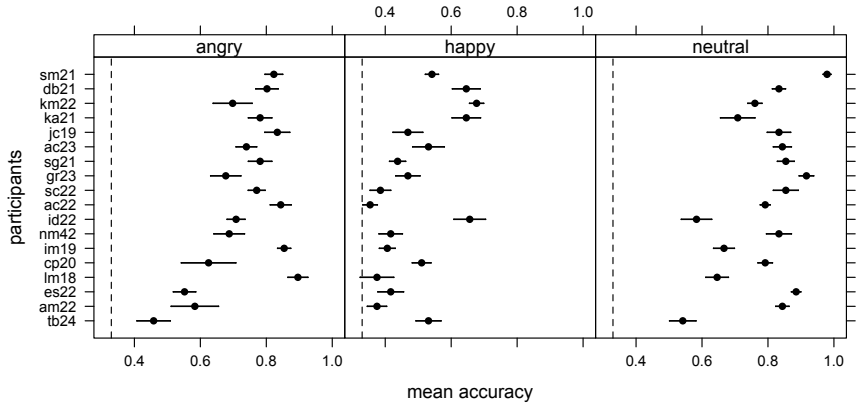


Figure A.5: Mean accuracy of emotion judgments for angry, happy and neutral displays for different participants (initials and age for specific participants given) for visual condition in Chapter 4. The error bars represent one standard error of the mean and the dashed line represents the level of chance (0.33).

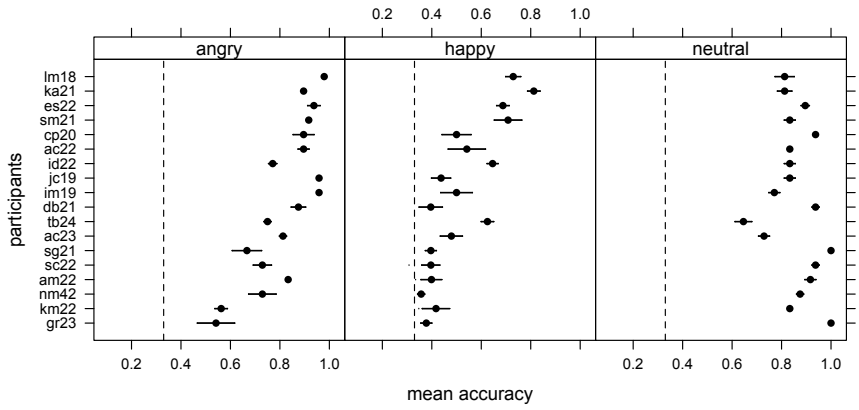


Figure A.6: Mean accuracy of emotion judgments for angry, happy and neutral displays for different participants (initials and age for specific participants given) for auditory condition in Chapter 4. The error bars represent one standard error of the mean and the dashed line represents the level of chance (0.33).

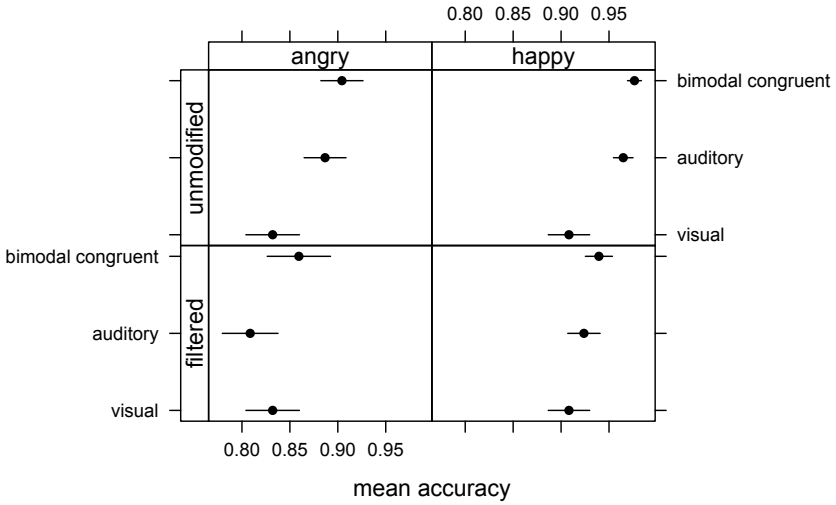


Figure A.7: Mean accuracy of emotion judgements and standard errors obtained in Experiment 1 (Chapter 5) for unimodal stimuli (auditory and visual) and congruent bimodal stimuli for both emotional expressions. The figure displays the results obtained with unmodified auditory stimuli (top row labeled *unmodified*) and brown noise filtered auditory stimuli (bottom row labeled *filtered*).

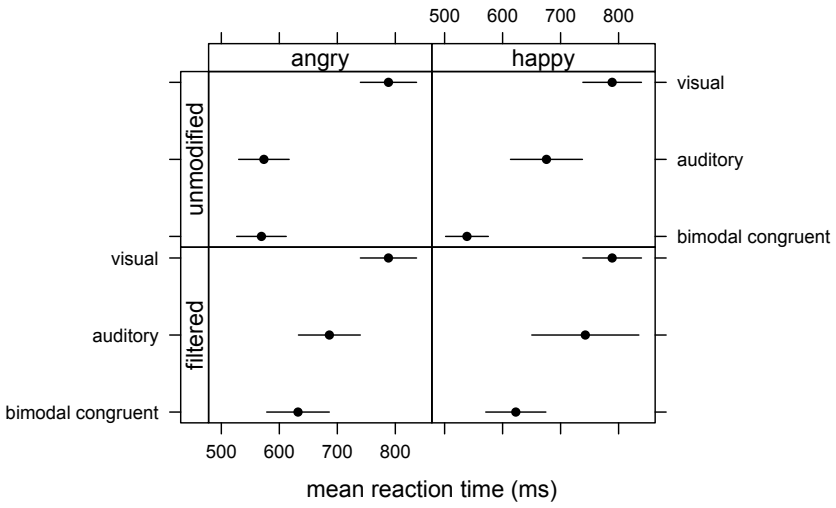


Figure A.8: Mean reaction times (in milliseconds) of emotion judgements and standard errors obtained in Experiment 1 (Chapter 5) for unimodal stimuli (auditory and visual) and congruent bimodal stimuli for both emotional expressions. The figure displays the results obtained with unmodified auditory stimuli (top row labeled *unmodified*) and brown noise filtered auditory stimuli (bottom row labeled *filtered*).

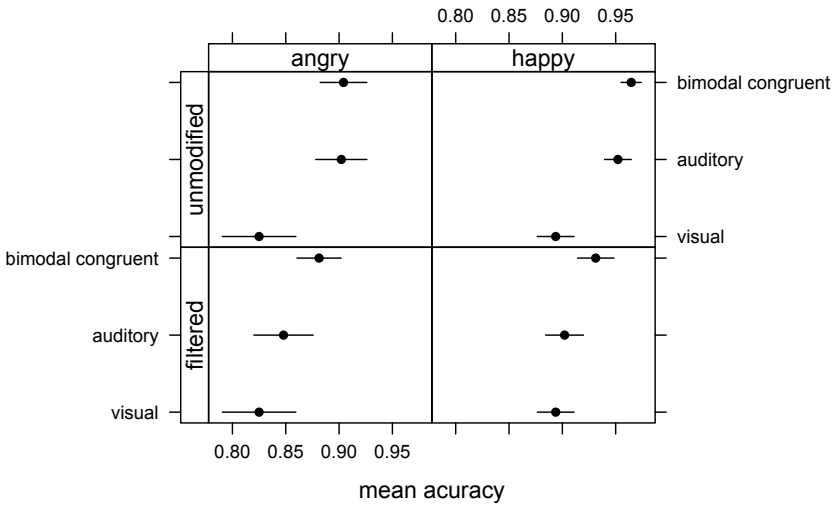


Figure A.9: Mean accuracy of emotion judgements and standard errors obtained in Experiment 2 (Chapter 5) for unimodal stimuli (auditory and visual) and congruent bimodal stimuli for both emotional expressions. The figure displays the results obtained with unmodified auditory stimuli (top row labeled *unmodified*) and low-pass filtered auditory stimuli (bottom row labeled *filtered*).

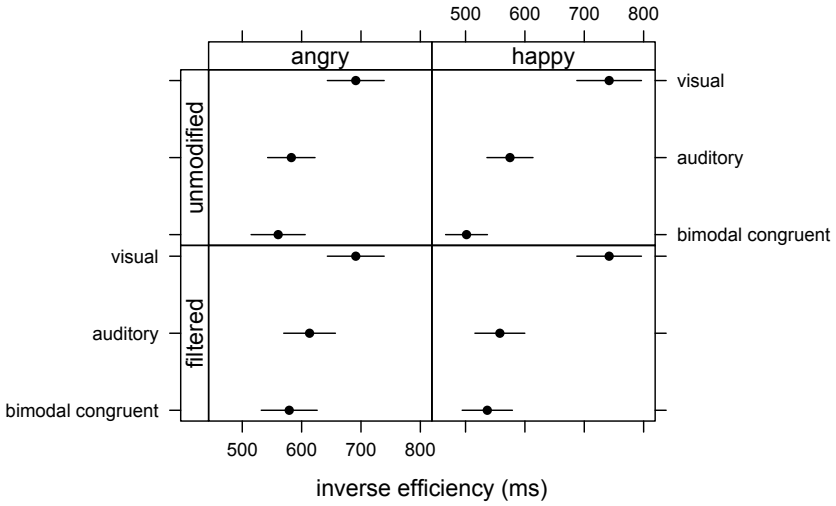


Figure A.10: Mean reaction times (in milliseconds) of emotion judgements and standard errors obtained in Experiment 2 (Chapter 5) for unimodal stimuli (auditory and visual) and congruent bimodal stimuli for both emotional expressions. The figure displays the results obtained with unmodified auditory stimuli (top row labeled *unmodified*) and low-pass filtered auditory stimuli (bottom row labeled *filtered*).

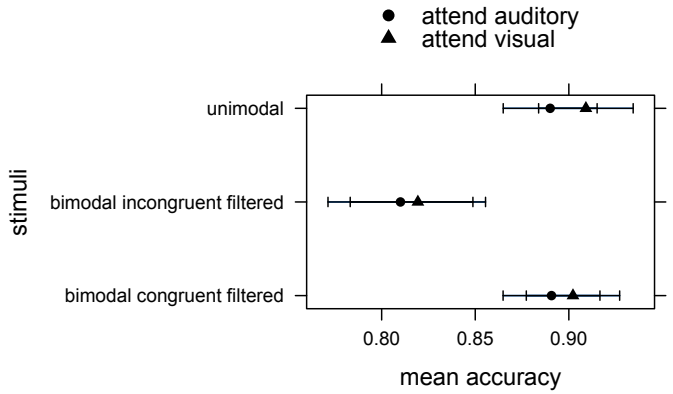


Figure A.11: Mean accuracy of emotion identification and standard errors obtained in Experiment 3 (Chapter 5) for unimodal, and congruent and incongruent bimodal filtered stimuli. Circles display performance when participants were instructed to attend the emotion expressed aurally, while triangles display performance when participants were instructed to attend the emotion expressed visually.

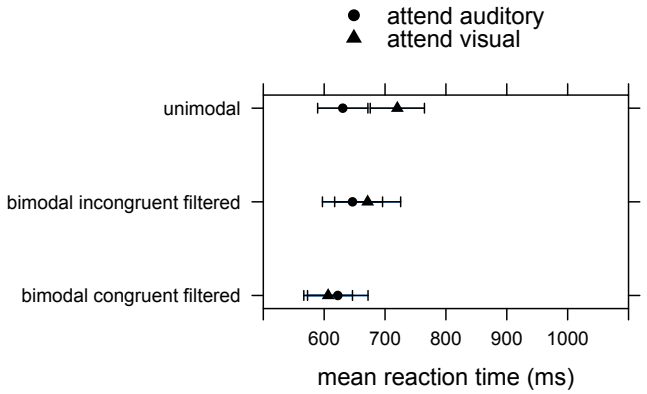


Figure A.12: Mean reaction times (milliseconds) and standard errors obtained in Experiment 3 (Chapter 5) for unimodal, and congruent and incongruent bimodal filtered stimuli. Circles display performance when participants were instructed to attend the emotion expressed aurally, while triangles display performance when participants were instructed to attend the emotion expressed visually.

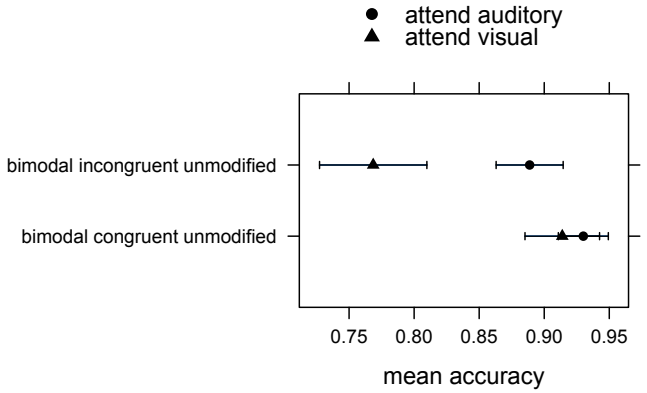


Figure A.13: Mean accuracy of emotion identification and standard errors obtained in Experiment 3 (Chapter 5) for congruent and incongruent bimodal unmodified stimuli. Circles display performance when participants were instructed to attend the emotion expressed aurally, while triangles display performance when participants were instructed to attend the emotion expressed visually.

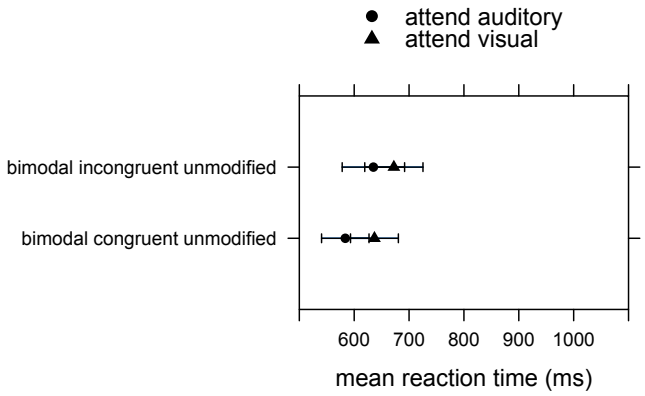


Figure A.14: Mean reaction times (milliseconds) and standard errors obtained in Experiment 3 (Chapter 5) for congruent and incongruent bimodal unmodified stimuli. Circles display performance when participants were instructed to attend the emotion expressed aurally, while triangles display performance when participants were instructed to attend the emotion expressed visually.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications and presentations:

Piwek, L., Petrini, K. & Pollick, F. (2012). Auditory signal dominates visual in the perception of emotional social interactions. *Seeing and Perceiving*, **25**, 112;

Petrini, K., Piwek, L. & Crabbe, F. (2011). The Precuneus role in third-person perspective of dyadic social interaction. *Perception*, **40**, 87;

Pollick, F., Steel, W., Tan, H., Piwek, L. & Ahlström, U. (2011). A New Action Library For Localising Brain Activity Specific To Biological Motion. *Journal of Vision*, **11**, 683;

Piwek, L., Petrini, K. & Pollick, F. (2010). Multimodal integration of the auditory and visual signals in dyadic point-light interactions. *Journal of Vision*, **10**, 788.

BIBLIOGRAPHY

Ada, M., Suda, K. & Ishi, M. (2003). Expression of emotions in dance: relation between arm movement characteristics and emotion. *Perceptual and motor skills*, **97**, 697–708.

Ahlström, V., Blake, R. & Ahlström, U. (1997). Perception of biological motion. *Perception*, **26**, 1539–48.

Alais, D. & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology : CB*, **14**, 257–62.

Armony, J.L. (2012). Current Emotion Research in Behavioral Neuroscience: The Role(s) of the Amygdala. *Emotion Review*, **5**, 104–115.

Armstrong, A., Stokoe, W. & Wilcox, S. (1995). *Gesture and the Nature of Language.*. Cam- bridge, UK: Cambridge Univ. Press.

Atkinson, A.P., Dittrich, W.H., Gemmell, A.J. & Young, A.W. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, **33**, 717–746.

Atkinson, A.P., Tunstall, M.L. & Dittrich, W.H. (2007). Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures. *Cognition*, **104**, 59–72.

Averill, J.R. & More, T.A. (2000). Happiness (2nd ed.) (pp.). . In M. Lewis & J.M. Haviland, eds., *Handbook of Emotions.*, 666–676, New York: Guilford Publications, 2nd edn.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior research methods*, **37**, 379–84.

Banse, R. & Scherer, K.R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, **70**, 614–636.

Barclay, C.D., Cutting, J.E. & Kozlowski, L.T. (1978). Temporal and spatial factors in gait perception that influence gender recognition. *Perception & psychophysics*, **23**, 145–52.

Barnes, J. & Allan, D. (1966). A statistical model of flicker noise. *Proceedings of the IEEE*, **54**, 176–178.

Beauchamp, M.S., Lee, K.E., Argall, B.D. & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, **41**, 809–23.

Bechara, A. (2000). Emotion, Decision Making and the Orbitofrontal Cortex. *Cerebral Cortex*, **10**, 295–307.

Belin, P., Fecteau, S. & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences*, **8**, 129–35.

Belin, P., Fillion-Bilodeau, S. & Gosselin, F. (2008). The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, **40**, 531–539.

Bertenthal, B.I. & Pinto, J. (1994). Global Processing of Biological Motions. *Psychological Science*, **5**, 221–224.

Bezzoijen, R. & Boves, L. (1986). The effects of low-pass filtering and random splicing on the perception of speech. *Journal of Psycholinguistic Research*, **15**, 403–417.

Boomer, D.S. (1963). Speech disturbance and body movement in interviews. *The Journal of nervous and mental disease*, **136**, 263–6.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, **226**, 177–8.

Bresciani, J.P. & Ernst, M.O. (2007). Signal reliability modulates auditory-tactile integration for event counting. *Neuroreport*, **18**, 1157–61.

Bricker, P. & Pruzansky, S. (1976). Speaker recognition. In N. Lass, ed., *Contemporary issues in experimental phonetics*, 295–326, New York: Academic Press.

Brotheridge, C.M. & Lee, R.T. (2003). Development and validation of the Emotional Labour Scale. *Journal of Occupational and Organizational Psychology*, **76**, 365–379.

Bryant, G.a. & Barrett, H.C. (2007). Recognizing intentions in infant-directed speech: evidence for universals. *Psychological science : a journal of the American Psychological Society / APS*, **18**, 746–51.

Bryant, G.a. & Barrett, H.C. (2008). Vocal Emotion Recognition Across Disparate Cultures. *Journal of Cognition and Culture*, **8**, 135–148.

Bryant, G.a. & Fox Tree, J.E. (2005). Is there an Ironic Tone of Voice? *Language and Speech*, **48**, 257–277.

Bull, P. (1990). What does gesture add to the spoken word. In H. Barlow & C. Blakemore, eds., *Images and Understanding*, 108–121, Cambridge: Cambridge University Press.

Burke, D., Taubert, J. & Higman, T. (2007). Are face representations viewpoint dependent? A stereo advantage for generalizing across different views of faces. *Vision research*, **47**, 2164–9.

Busso, C. (2008). Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database. *Ninth Annual Conference of the International Speech*.

Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. & Narayanan, S.S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, **42**, 335–359.

Campanella, S. & Belin, P. (2007). Integrating face and voice in person perception. *Trends in cognitive sciences*, **11**, 535–43.

Campos, J.J., Campos, R.G. & Barrett, K.C. (1989). Emergent themes in the study of emotional development and emotion regulation. *Developmental Psychology*, **25**, 394–402.

Centelles, L., Assaiante, C., Nazarian, B., Anton, J.L. & Schmitz, C. (2011). Recruitment of Both the Mirror and the Mentalizing Networks When Observing Social Interactions Depicted by Point-Lights: A Neuroimaging Study. *PLoS ONE*, **6**, e15749.

Chaminade, T., Hodgins, J. & Kawato, M. (2007). Anthropomorphism influences perception of computer-animated characters' actions. *Social cognitive and affective neuroscience*, **2**, 206–16.

Chang, D.H.F. & Troje, N.F. (2008). Perception of animacy and direction from local biological motion signals. *Journal of vision*, **8**, 3.1–10.

Chouchourelou, A., Matsuka, T., Harber, K. & Shiffrar, M. (2006). The visual analysis of emotional actions. *Social neuroscience*, **1**, 63–74.

Clarke, T.J., Bradshaw, M.F., Field, D.T., Hampson, S.E. & Rose, D. (2005). The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception*, **34**, 1171–1180.

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M. & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain research*, **1242**, 126–35.

CORBALLIS, M. (2002). *From hand to mouth: The origins of language*. Princeton: Princeton Univ. Press.

COSMIDES, L. (1983). Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 864.

COULSON, M. (2004). Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence. *Journal of Nonverbal Behavior*, **28**, 117–139.

CRAGGS, R. & WOOD, M. (2003). Annotating emotion in dialogue. In *4th SIGdial Workshop on Discourse and Dialogue*.

CUTTING, J.E. (1978). A program to generate synthetic walkers as dynamic point-light displays. *Behavior Research Methods & Instrumentation*, **10**, 91–94.

CUTTING, J.E. (1981). Coding theory adapted to gait perception. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 71–87.

CUTTING, J.E. & KOZLOWSKI, L.T. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, **9**, 353–356.

CUTTING, J.E., MOORE, C. & MORRISON, R. (1988). Masking the motions of human gait. *Perception & psychophysics*, **44**, 339–47.

DAEL, N., MORTILLARO, M. & SCHERER, K.R. (2012). Emotion expression in body action and posture. *Emotion*, **12**, 1085–101.

DAEMS, A. & VERFAILLIE, K. (1999). Viewpoint-dependent Priming Effects in the Perception of Human Actions and Body Postures. *Visual Cognition*, **6**, 665–693.

DARWIN, C. (1872). The Expression of the Emotions in Man and Animals. *The American Journal of the Medical Sciences*.

DAVIS, R.B., ÕUNPUU, S., TYBURSKI, D. & GAGE, J.R. (1991). A gait analysis data collection and reduction technique. *Human Movement Science*, **10**, 575–587.

DE GELDER, B. (2006). Towards the neurobiology of emotional body language. *Nature reviews. Neuroscience*, **7**, 242–9.

DE GELDER, B. & HORTENSIUS, R. (2014). The Many Faces of the Emotional Body. In J. Decety & Y. Christen, eds., *New Frontiers in Social Neuroscience*, 153–164, Springer International Publishing.

DE GELDER, B. & VAN DEN STOCK, J. (2011). The Bodily Expressive Action Stimulus Test (BEAST). Construction and Validation of a Stimulus Basis for Measuring Perception of Whole Body Expression of Emotions. *Frontiers in psychology*, **2**, 181.

DE GELDER, B. & VROOMEN, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion*, **14**, 289–311.

DE GELDER, B., BÖCKER, K.B., TUOMAINEN, J., HENSEN, M. & VROOMEN, J. (1999). The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses. *Neuroscience letters*, **260**, 133–6.

DEKEYSER, M., VERFAILLIE, K. & VANRIE, J. (2002). Creating stimuli for the study of biological-motion perception. *Behavior research methods*, **34**, 375–382.

DEPAULO, B.M., LINDSAY, J.J., MALONE, B.E., MUHLENBRUCK, L., CHARLTON, K. & COOPER, H. (2003). Cues to deception. *Psychological Bulletin*, **129**, 74–118.

DERING, B., MARTIN, C.D., MORO, S., PEGNA, A.J. & THIERRY, G. (2011). Face-sensitive processes one hundred milliseconds after picture onset. *Frontiers in human neuroscience*, **5**, 93.

DIEBOLD, F. (2006). *Elements of forecasting*. South-Western College Publishing.

DIRKS, D.D. (1970). Effect of Forward and Backward Masking on Speech Intelligibility. *The Journal of the Acoustical Society of America*, **47**, 1003.

DITTMANN, A.T. & LLEWELLYN, L.G. (1969). Body movement and speech rhythm in social conversation. *Journal of personality and social psychology*, **11**, 98–106.

DITTRICH, W.H. (1993). Action categories and the perception of biological motion. *Perception*, **22**, 15–22.

DITTRICH, W.H., TROSCIANKO, T., LEA, S.E. & MORGAN, D. (1996). Perception of emotion from dynamic point-light displays represented in dance. *Perception*, **25**, 727–38.

DOLAN, R.J., MORRIS, J.S. & DE GELDER, B. (2001). Crossmodal binding of fear in voice and face. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 10006–10.

Douglas-Cowie, E., Campbell, N., Cowie, R. & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, **40**, 33–60.

Dunbar, R.I. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, **22**, 469–493.

Eastwood, J.D., Smilek, D. & Merikle, P.M. (2001). Differential attentional guidance by unattended faces expressing positive and negative emotion. *Perception & psychophysics*, **63**, 1004–13.

Ekman, P. (1965). Communication through nonverbal behavior: A source of information about an interpersonal relationship. In S. Tomkins & C. Izard, eds., *Affect, cognition and personality.*, New York: Springer.

Ekman, P. (1994). Antecedent events and emotion metaphors. In P. Ekman & R.J. Davidson, eds., *The Nature of Emotion: Fundamental Questions.*, 146–149, New York: Oxford University Press.

Ekman, P. (1999). Emotional And Conversational Nonverbal Signals. In L. Messing & R. Campbell, eds., *Gesture, Speech and Sign.*, vol. 99, 45–55, London: Oxford University Press.

Ekman, P. & Friesen, W. (1978). *Facial action coding system: A technique for the measurement of facial movement. Palo Alto*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P. & Friesen, W. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.

Ekman, P. & Friesen, W.V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, **17**, 124–129.

Ekman, P. & O'Sullivan, M. (1991). Who can catch a liar? *The American psychologist*, **46**, 913–20.

Ekman, P., Friesen, W.V. & Scherer, K. (1976). Body movement and voice pitch in deceptive interaction. *Semiotica*, **16**, 23–27.

Ekman, P., O'Sullivan, M., Friesen, W.V. & Scherer, K.R. (1991). Invited article: Face, voice, and body in detecting deceit. *Journal of Nonverbal Behavior*, **15**, 125–135.

Ekman, P., O'Sullivan, M. & Frank, M.G. (1999). A Few Can Catch a Liar. *Psychological Science*, **10**, 263–266.

Ernst, M.O. & Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, **415**, 429–33.

Ethofer, T., Anders, S., Erb, M., Droll, C., Royen, L., Saur, R., Reiterer, S., Grodd, W. & Wildgruber, D. (2006). Impact of voice on emotional judgment of faces: an event-related fMRI study. *Human brain mapping*, **27**, 707–14.

Farah, M.J., Tanaka, J.W. & Drain, H.M. (1995). What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 628–634.

Feingold, A. (1994). Gender differences in personality: a meta-analysis. *Psychological bulletin*, **116**, 429–56.

Fernandez, E. & Cairns, H.S. (2010). *Fundamentals of psycholinguistics*. John Wiley & Sons.

Ferri, F., Ebisch, S.J.H., Costantini, M., Salone, A., Arciero, G., Mazzola, V., Ferro, F.M., Romani, G.L. & Gallese, V. (2013). Binding action and emotion in social understanding. *PloS one*, **8**, e54091.

Fischer, A.H. & Manstead, A.S. (2008). Social functions of emotion. In M. Lewis, J. Haviland-Jones & L. Barrett, eds., *Handbook of Emotions.*, 456–468, New York, NY: THE GUILFORD PRESS.

Flom, M.C., Weymouth, F.W. & Kahneman, D. (1963). Visual resolution and contour interaction. *Journal of the Optical Society of America*, **53**, 1026–32.

Fox, E. (2008). *Emotion science: An integration of cognitive and neuroscientific approaches*. New York, NY: Palgrave Macmillan.

Fox, E., Lester, V., Russo, R., Bowles, R.J., Pichler, A. & Dutton, K. (2000). Facial Expressions of Emotion: Are Angry Faces Detected More Efficiently? *Cognition & emotion*, **14**, 61–92.

Freedman, N. & Hoffman, S.P. (1967). Kinetic behavior in altered clinical states: approach to objective analysis of motor behavior during clinical interviews. *Perceptual and motor skills*, **24**, 527–39.

French, N.R. & Steinberg, J.C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, **19**, 90–119.

Frick, R.W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, **97**, 412–429.

FRIJDA, N.H. (1986). *The emotions.*. New York: Cambridge University Press.

FRITH, C.D. (2007). The social brain? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **362**, 671–8.

FRÜND, I., HAENEL, N.V. & WICHMANN, F.A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of vision*, **11**, 1–19.

GALLACE, A. & SPENCE, C. (2010). The science of interpersonal touch: an overview. *Neuroscience and biobehavioral reviews*, **34**, 246–59.

GALLAGHER, H.L. & FRITH, C.D. (2003). Functional imaging of 'theory of mind'. *Trends in cognitive sciences*, **7**, 77–83.

GALLESE, V., KEYSERS, C. & RIZZOLATTI, G. (2004). A unifying view of the basis of social cognition. *Trends in cognitive sciences*, **8**, 396–403.

GARDINER, C. (2009). *Stochastic methods: a handbook for the natural and social sciences*. Springer, New York, 4th edn.

GELFER, M. (1988). Perceptual attributes of voice: Development and use of rating scales. *Journal of Voice*, **2**, 320–326.

GRANDEY, A.A. (2003). When "The Show Must Go On": Surface Acting And Deep Acting As Determinants Of Emotional Exhaustion And Peer-Rated Service Delivery. *Academy of Management Journal*, **46**, 86–96.

GRAY, H. (2007). *Anatomy of the human body*. Lea & Febiger, 37th edn.

GRAY, J.A. (1982). *The neuropsychology of anxiety.*. Oxford: Oxford University Press.

GREEN, J., WHITNEY, P. & GUSTAFSON, G. (2010). Vocal expressions of anger. *International Handbook of Anger*, 139–156.

GRILL-SPECTOR, K., KUSHNIR, T., EDELMAN, S., AVIDAN, G., ITZCHAK, Y. & MALACH, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, **24**, 187–203.

GROSS, M.M., CRANE, E.A. & FREDRICKSON, B.L. (2010). Methodology for Assessing Bodily Expression of Emotion. *Journal of Nonverbal Behavior*, **34**, 223–248.

GROSSMAN, E.D. & BLAKE, R. (2002). Brain Areas Active during Visual Perception of Biological Motion. *Neuron*, **35**, 1167–75.

HALL, E. (1966). *The hidden dimension*. Anchor Books.

HALL, J., CARTER, J. & HORGAN, T. (2000). Gender differences in nonverbal communication of emotion. In A.H. Fischer, ed., *Gender and emotion: Social psychological perspectives*, 97–118, Cambridge University Press, NY.

HAMMERSCHMIDT, K. & JÜRGENS, U. (2007). Acoustical correlates of affective prosody. *Journal of voice : official journal of the Voice Foundation*, **21**, 531–40.

HANSEN, C.H. & HANSEN, R.D. (1988). Finding the face in the crowd: an anger superiority effect. *Journal of personality and social psychology*, **54**, 917–24.

HARLOW, H. (1958). The nature of love. *American Psychologist*, **13**, 673–685.

HARRIGAN, J. (2005). Proxemics, kinesics, and gaze. In J.A. Harrigan, R. Rosenthal & K.R. Scherer, eds., *The new handbook of methods in nonverbal behavior research*, 137–198, New York, NY: Oxford University Press.

HAYWARD, W.G. (2003). After the viewpoint debate: where next in object recognition? *Trends in Cognitive Sciences*, **7**, 425–427.

HEIMAN, G.W. (2002). *Research Methods in Psychology.*. Boston & New York. Houghton Mifflin Company., 3rd edn.

HILL, H. & POLLICK, F.E. (2000). Exaggerating temporal differences enhances recognition of individuals from point light displays. *Psychological science : a journal of the American Psychological Society / APS*, **11**, 223–8.

HILL, H., SCHYNS, P.G. & AKAMATSU, S. (1997). Information and viewpoint dependence in face recognition. *Cognition*, **62**, 201–22.

HUBERT, B., WICKER, B., MOORE, D.G., MONFARDINI, E., DUVERGER, H., DA FONSÉCA, D. & DERUELLE, C. (2007). Brief report: recognition of emotional and non-emotional biological motion in individuals with autistic spectrum disorders. *Journal of autism and developmental disorders*, **37**, 1386–92.

IKEDA, H. & WATANABE, K. (2009). Anger and happiness are linked differently to the explicit detection of biological motion. *Perception*, **38**, 1002–1011.

Ikeda, H., Blake, R. & Watanabe, K. (2005). Eccentric perception of biological motion is unscalably poor. *Vision research*, **45**, 1935–43.

Ikeda, H., Watanabe, K. & Cavanagh, P. (2013). Crowding of biological motion stimuli. *Journal of vision*, **13**, 20.

Jack, R.E., Blais, C., Scheepers, C., Schyns, P.G. & Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Current biology*, **19**, 1543–8.

Jessen, S., Obleser, J. & Kotz, S.a. (2012). How bodies and voices interact in early emotion perception. *PloS one*, **7**, e36070.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perceiving events and objects*, **14**, 201–211.

Johnstone, T., van Reekum, C.M., Oakes, T.R. & Davidson, R.J. (2006). The voice of emotion: an FMRI study of neural responses to angry and happy vocal expressions. *Social cognitive and affective neuroscience*, **1**, 242–9.

Jokisch, D. & Troje, N.F. (2003). Biological motion as a cue for the perception of size. *Journal of vision*, **3**, 252–64.

Jokisch, D., Daum, I. & Troje, N.F. (2006). Self recognition versus recognition of others by biological motion: Viewpoint-dependent effects. *Perception*, **35**, 911–920.

Kadaba, M.P., Ramakrishnan, H.K. & Wootten, M.E. (1990). Measurement of lower extremity kinematics during level walking. *Journal of orthopaedic research : official publication of the Orthopaedic Research Society*, **8**, 383–92.

Kilner, J.M., Marchant, J.L. & Frith, C.D. (2006). Modulation of the mirror system by social relevance. *Social cognitive and affective neuroscience*, **1**, 143–8.

Kleinsmith, A. & Bianchi-Berthouze, N. (2013). Affective Body Expression Perception and Recognition: A Survey. *IEEE Transactions on Affective Computing*, **4**, 15–33.

Kleinsmith, A., Bianchi-Berthouze, N. & Steed, A. (2011). Automatic Recognition of Non-Acted Affective Postures. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, **41**, 1027–1038.

Klette, R. & Tee, G. (2008). Understanding human motion: A historic review. *Human Motion*.

KLINNERT, M.D., EMDE, R.N., BUTTERFIELD, P. & CAMPOS, J.J. (1986). Social referencing: The infant's use of emotional signals from a friendly adult with mother present. *Developmental Psychology*, **22**, 427–432.

KNAPP, M. & HALL, J. (2009). *Nonverbal communication in human interaction*. Wadsworth Publishing Company.

KNOLL, M., UTHER, M. & COSTALL, A. (2009). Effects of low-pass filtering on the judgment of vocal affect in speech directed to infants, adults and foreigners. *Speech Communication*, **51**, 210–216.

KNUTSON, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*, **20**, 165–182.

KNYAZEV, G.G., BOCHAROV, A.V. & SLOBODSKOJ-PLUSNIN, J.Y. (2009). Hostility- and gender-related differences in oscillatory responses to emotional facial expressions. *Aggressive behavior*, **35**, 502–13.

KONIJN, E. (2000). *Acting emotions: Shaping emotions on stage*. Amsterdam University Press.

KRABBE, E. & WALTON, D. (1995). *Commitment in dialogue*. Albany: State University of New York Press.

KRAMER, E. (1963). Judgment of personal characteristics and emotions from nonverbal properties of speech. *Psychological Bulletin*, **60**, 408–420.

KRAMER, E. (1964). Elimination of verbal cues in judgments of emotion from voice. *Journal of abnormal psychology*, **68**, 390–6.

KREIFELTS, B., ETHOFER, T., GRODD, W., ERB, M. & WILDGRUBER, D. (2007). Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. *NeuroImage*, **37**, 1445–56.

KUHLMANN, S., DE LUSSANET, M.H.E. & LAPPE, M. (2009). Perception of limited-lifetime biological motion from different viewpoints. *Journal of vision*, **9**, 11.1–14.

LATTNER, S., MEYER, M.E. & FRIEDERICI, A.D. (2005). Voice perception: Sex, pitch, and the right hemisphere. *Human brain mapping*, **24**, 11–20.

LAVER, J. & TRUDGILL, P. (1979). Phonetic and linguistic markers in speech. In K. Scherer & H. Giles, eds., *Social markers in speech*, 1–31, Cambridge, England: Cambridge University Press.

LEE, Y., MATSUMIYA, K. & WILSON, H.R. (2006). Size-invariant but viewpoint-dependent representation of faces. *Vision research*, **46**, 1901–10.

LOREY, B., KALETSCH, M., PILGRAMM, S., BISCHOFF, M., KINDERMANN, S., SAUERBIER, I., STARK, R., ZENTGRAF, K. & MUNZERT, J. (2012). Confidence in emotion perception in point-light displays varies with the ability to perceive own emotions. *PloS one*, **7**, e42169.

MA, Y., PATERSON, H.M., DOLIA, A., CHO, S.B., UDE, A. & POLLICK, F.E. (2004). Towards a biologically-inspired representation of human affect. *Brain Inspired Cognitive Systems*.

MA, Y., PATERSON, H.M. & POLLICK, F.E. (2006). A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior research methods*, **38**, 134–41.

MACCALLUM, J.K., OLSZEWSKI, A.E., ZHANG, Y. & JIANG, J.J. (2011). Effects of low-pass filtering on acoustic analysis of voice. *Journal of voice : official journal of the Voice Foundation*, **25**, 15–20.

MACDONNELL, K. (1972). *Eadweard Muybridge, The Man Who Invented Moving Pictures*. Weidenfeld and Nicolson (London).

MANERA, V., SCHOUTEN, B., BECCHIO, C., BARA, B.G. & VERFAILLIE, K. (2010). Inferring intentions from biological motion: a stimulus set of point-light communicative interactions. *Behavior research methods*, **42**, 168–78.

MANERA, V., BECCHIO, C., SCHOUTEN, B., BARA, B.G. & VERFAILLIE, K. (2011). Communicative interactions improve visual detection of biological motion. *PloS one*, **6**, e14594.

MANERA, V., SCHOUTEN, B., VERFAILLIE, K. & BECCHIO, C. (2013). Time Will Show: Real Time Predictions during Interpersonal Action Perception. *PloS one*, **8**, e54949.

MAREY, E. (1898). Analyse des mouvements du cheval par la chronophotographie. *La nature*, **1306**.

MASSARO, D.W. & EGAN, P.B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, **3**, 215–221.

MATHER, G. & MURDOCH, L. (1994). Gender Discrimination in Biological Motion Displays Based on Dynamic Cues. *Proceedings of the Royal Society B: Biological Sciences*, **258**, 273–279.

MATHER, G., RADFORD, K. & WEST, S. (1992). Low-level visual processing of biological motion. *Proceedings. Biological sciences / The Royal Society*, **249**, 149–55.

MATSUMOTO, D. (1992). American-Japanese Cultural Differences in the Recognition of Universal Facial Expressions. *Journal of Cross-Cultural Psychology*, **23**, 72–84.

MATSUMOTO, D. (1993). Ethnic differences in affect intensity, emotion judgments, display rule attitudes, and self-reported emotional expression in an American sample. *Motivation and Emotion*, **17**, 107–123.

MATSUMOTO, D. (2010). The Expression of Anger Across Cultures. *International handbook of anger*, 125–137.

MATSUMOTO, H., HIKI, S., SONE, T. & NIMURA, T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics*, **21**, 428–436.

MCALEER, P. & POLLICK, F.E. (2008). Understanding intention from minimal displays of human activity. *Behavior Research Methods*, **40**, 830–839.

MCCLAVE, E. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, **32**, 855–878.

MEEREN, H.K.M., VAN HEIJNSBERGEN, C.C.R.J. & DE GELDER, B. (2005). Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 16518–23.

MEHRABIAN, A. & BLUM, J.S. (1997). Physical appearance, attractiveness, and the mediating role of emotions. *Current Psychology*, **16**, 20–42.

MEIJER, M. (1989). The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, **13**, 247–268.

MENDONÇA, C., SANTOS, J.A. & LÓPEZ-MOLINER, J. (2011). The benefit of multisensory integration with biological motion signals. *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale*, **213**, 185–192.

MOESLUND, T.B. & GRANUM, E. (2001). A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, **81**, 231–268.

MONTEPARE, J.M. & ZEBROWITZ-MCARTHUR, L. (1988). Impressions of people created by age-related qualities of their gaits. *Journal of personality and social psychology*, **55**, 547–56.

MOORE, D.G., HOBSON, R.P. & LEE, A. (1997). Components of person perception: An investigation with autistic, non-autistic retarded and typically developing children and adolescents. *British Journal of Developmental Psychology*, **15**, 401–423.

MURRAY, I.R. & ARNOTT, J.L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, **93**, 1097.

NACKAERTS, E., WAGEMANS, J., HELSEN, W., SWINNEN, S.P., WENDEROTH, N. & ALAERTS, K. (2012). Recognizing biological motion and emotions from point-light displays in autism spectrum disorders. *PloS one*, **7**, e44473.

NERI, P., LUU, J.Y. & LEVI, D.M. (2006). Meaningful interactions can enhance visual discrimination of human agents. *Nature neuroscience*, **9**, 1186–92.

NESSE, R.M. (1990). Evolutionary explanations of emotions. *Human Nature*, **1**, 261–289.

NEWELL, F.N., CHIRORO, P. & VALENTINE, T. (1999). Recognizing unfamiliar faces: the effects of distinctiveness and view. *The Quarterly journal of experimental psychology. A, Human experimental psychology*, **52**, 509–34.

NISHIDA, S. (2011). Advancement of motion psychophysics: review 2001-2010. *Journal of vision*, **11**, 11.

OHMAN, A., LUNDQVIST, D. & ESTEVES, F. (2001). The face in the crowd revisited: a threat advantage with schematic stimuli. *Journal of personality and social psychology*, **80**, 381–96.

OLEJNIK, S. & ALGINA, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods*, **8**, 434–47.

OLOFSSON, U., NYBERG, L. & NILSSON, L.G. (1997). Priming and Recognition of Human Motion Patterns. *Visual Cognition*, **4**, 373–382.

Ortony, A. & Turner, T.J. (1990). What's basic about basic emotions? *Psychological review*, **97**, 315–31.

Palmer, S., Rosch, E. & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley, eds., *Attention and performance IX*, Lawrence Erlbaum, Hillsdale, NJ.

Pavlova, M. & Sokolov, A. (2000). Orientation specificity in biological motion perception. *Perception & psychophysics*, **62**, 889–99.

Peirce, J.W. (2007). PsychoPy–Psychophysics software in Python. *Journal of neuroscience methods*, **162**, 8–13.

Perrett, D.I., Harries, M.H., Bevan, R., Thomas, S., Benson, P.J., Mistlin, A.J., Chitty, A.J., Hietanen, J.K. & Ortega, J.E. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *The Journal of experimental biology*, **146**, 87–113.

Perrett, D.I., Hietanen, J.K., Oram, M.W. & Benson, P.J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **335**, 23–30.

Petrini, K., McAleer, P. & Pollick, F.E. (2010). Audiovisual integration of emotional signals from music improvisation does not depend on temporal correspondence. *Brain research*, **1323**, 139–48.

Petrini, K., Piwek, L. & Crabbe, F. (2011). The Precuneus role in third-person perspective of dyadic social interaction. *Perception*, **40**, 87.

Petrini, K., Remark, A., Smith, L. & Nardini, M. (2012). When vision is not an option: Development of haptic-auditory integration. *Seeing and Perceiving*, **25**, 205–205.

Pichon, S., de Gelder, B. & Grezes, J. (2008). Emotional modulation of visual and motor areas by dynamic body expressions of anger. *Social neuroscience*, **3**, 199–212.

Pichon, S., de Gelder, B. & Grèzes, J. (2011). Threat Prompts Defensive Brain Responses Independently of Attentional Control. *Cerebral cortex (New York, N.Y. : 1991)*.

Piwek, L., Petrini, K. & Pollick, F. (2010). Multimodal integration of the auditory and visual signals in dyadic point-light interactions. *Journal of Vision*, **10**, 788.

Piwek, L., Petrini, K. & Pollick, F. (2012). Auditory signal dominates visual in the perception of emotional social interactions. *Seeing and Perceiving*, **25**, 112.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. New York: Harper and Row.

Pollack, I. (1948). Effects of High Pass and Low Pass Filtering on the Intelligibility of Speech in Noise. *The Journal of the Acoustical Society of America*, **20**, 259–266.

Pollick, F., Steel, W., Tan, H., Piwek, L. & Ahlström, U. (2011). A New Action Library For Localising Brain Activity Specific To Biological Motion. *Journal of Vision*, **11**, 683.

Pollick, F.E., Paterson, H.M., Bruderlin, A. & Sanford, A.J. (2001). Perceiving affect from arm movement. *Cognition*, **82**, B51–61.

Pollick, F.E., Lestou, V., Ryu, J. & Cho, S.B. (2002). Estimating the efficiency of recognizing gender and affect from biological motion. *Vision research*, **42**, 2345–55.

Pollick, F.E., Kay, J.W., Heim, K. & Stringer, R. (2005). Gender recognition from point-light walkers. *Journal of experimental psychology. Human perception and performance*, **31**, 1247–65.

Poom, L. (2012). Memory of gender and gait direction from biological motion: Gender fades away but directions stay. *Journal of experimental psychology. Human perception and performance*, **38**, 1091–7.

Pourtois, G., Gelder, B.D., Bol, A. & Crommelinck, M. (2005). Perception of facial expressions and voices and of their combination in the human brain. *Cortex*, **41**, 49–59.

Remmington, N.A., Fabrigar, L.R. & Visser, P.S. (2000). Reexamining the circumplex model of affect. *Journal of personality and social psychology*, **79**, 286–300.

Renneker, R. (1963). Kinesic research and therapeutic processes: Further discussion. In P.H. Knapp, ed., *Expression oj the emotions in man.*, New York: International Universities Press.

Risko, E.F., Laidlaw, K., Freeth, M., Foulsham, T. & Kingstone, A. (2012). Social attention with real versus reel stimuli: toward an empirical approach to concerns about ecological validity. *Frontiers in human neuroscience*, **6**, 143.

Rizzolatti, G. & Arbib, M.A. (1998). Language within our grasp. *Trends in neurosciences*, **21**, 188–94.

Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annual review of neuroscience*, **27**, 169–92.

Rizzolatti, G., Fogassi, L. & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature reviews. Neuroscience*, **2**, 661–70.

Roach, P., Stibbard, R., Osborne, J., Arnfield, S. & Setter, J. (2009). Transcription of Prosodic and Paralinguistic Features of Emotional Speech. *Journal of the International Phonetic Association*, **28**, 83.

Röder, B., Kusmierek, A., Spence, C. & Schicke, T. (2007). Developmental vision determines the reference frame for the multisensory control of action. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 4753–8.

Roether, C.L., Omlor, L., Christensen, A. & Giese, M.A. (2009). Critical features for the perception of emotion from gait. *Journal of vision*, **9**, 15.1–32.

Rogers, P.L., Scherer, K.R. & Rosenthal, R. (1971). Content- filtering human speech: A simple electronic system. *Behavioral Research Methods and Instrumentation*, **3**, 16–18.

Rose, D. & Clarke, T.J. (2009). Look who's talking: Visual detection of speech from whole-body biological motion cues during emotive interpersonal conversation. *Perception*, **38**, 153–156.

Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, **39**, 1161–1178.

Sander, D., Grafman, J. & Zalla, T. (2003). The human amygdala: an evolved system for relevance detection. *Reviews in the neurosciences*, **14**, 303–16.

Sauter, D.a., Eisner, F., Calder, A.J. & Scott, S.K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly journal of experimental psychology (2006)*, **63**, 2251–72.

Saxe, R. (2006). Uniquely human social cognition. *Current opinion in neurobiology*, **16**, 235–9.

Saygin, A.P., Driver, J. & de Sa, V.R. (2008). In the footsteps of biological motion and multisensory perception: judgments of audiovisual temporal relations are enhanced for upright walkers. *Psychological*

science : a journal of the American Psychological Society / APS, **19**, 469–75.

Scherer, K.R. (1971). Randomized splicing: A note on a simple technique for masking speech content. *Journal of Experimental Research in Personality*, **5**, 155–159.

Scherer, K.R. (1986). Vocal affect expression: a review and a model for future research. *Psychological bulletin*, **99**, 143–65.

Scherer, K.R. (1989). Vocal correlates of emotion. In H. Wagner & A. Manstead, eds., *Handbook of Psychophysiology: Emotion and Social Behavior.*, 165–197, Wiley, London.

Scherer, K.R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, **40**, 227–256.

Scherer, K.R. (2005). What are emotions? And how can they be measured? *Social Science Information*, **44**, 695–729.

Scherer, K.R. & Ceschi, G. (1997). Lost Luggage: A Field Study of Emotion-Antecedent Appraisal. *Motivation and emotion*, **21**, 211–235.

Scherer, K.R. & Oshinsky, J.S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, **1**, 331–346.

Scherer, K.R. & Tannenbaum, P.H. (1986). Emotional experiences in everyday life: A survey approach. *Motivation and Emotion*, **10**, 295–314.

Schouten, B., Davila, A. & Verfaillie, K. (2013). Further explorations of the facing bias in biological motion perception: perspective cues, observer sex, and response times. *PloS one*, **8**, e56978.

Sevdalis, V. & Keller, P.E. (2011). Perceiving performer identity and intended expression intensity in point-light displays of dance. *Psychological research*, **75**, 423–34.

Sheffert, S.M., Pisoni, D.B., Fellowes, J.M. & Remez, R.E. (2003). Learning to recognize talkers from natural, sinewave and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance*, **28**, 1447–1469.

Shiffrar, M. (2011). People watching: visual, motor, and social processes in the perception of human movement. *Wiley Interdisciplinary Reviews: Cognitive Science*, **2**, 68–78.

Soskin, W.F. & Kauffman, P.E. (1961). Judgement of emotion in word-free voice samples. *Journal of Communication*, **11**, 73–80.

SPENCE, C., SHORE, D.I., GAZZANIGA, M.S., SOTO-FARACO, S. & KINGSTONE, A. (2001). Failure to remap visuotactile space across the midline in the split-brain. *Canadian journal of experimental psychology = Revue canadienne de psychologie expérimentale*, **55**, 133–40.

STARKWEATHER, J.A. (1956). Content-free speech as a source of information about the speaker. *The Journal of Abnormal and Social Psychology*, **52**, 394–402.

STEARNS, P. (1995). History of Emotions: Issues of Change and Impact. In M. Lewis, J. Haviland-Jones & L. Barrett, eds., *Handbook of Emotions.*, vol. 24, 17–32, New York, NY: THE GUILFORD PRESS.

STEIN, B. & MEREDITH, M. (1993). *The merging of the senses.*. Cambridge, MA, US: The MIT Press.

STIENEN, B.M.C., TANAKA, A. & DE GELDER, B. (2011). Emotional Voice and Emotional Body Postures Influence Each Other Independently of Visual Awareness. *PLoS ONE*, **6**, e25517.

TARR, M.J. & PINKER, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, **21**, 233–282.

THURMAN, S.M. & LU, H. (2013). Physical and biological constraints govern perceived animacy of scrambled human forms. *Psychological science*, **24**, 1133–41.

TOWNSEND, J. & ASHBY, F. (1978). Methods of modeling capacity in simple processing systems. In N.J. Castellan & F. Restle, eds., *Cognitive Theory, vol. 3*, 199–239, Hillsdale, NJ.

TOWNSEND, J. & ASHBY, F. (1983). *Stochastic Modelling of Elementary Psychological Processes*. Cambridge University Press, NY.

TROJE, N.F. (2002). Decomposing biological motion: a framework for analysis and synthesis of human gait patterns. *Journal of vision*, **2**, 371–87.

TROJE, N.F. & BÜLTHOFF, H.H. (1996). Face recognition under varying poses: the role of texture and shape. *Vision research*, **36**, 1761–71.

TROJE, N.F. & WESTHOFF, C. (2006). The inversion effect in biological motion perception: evidence for a "life detector"? *Current biology*, **16**, 821–4.

TROJE, N.F., WESTHOFF, C. & LAVROV, M. (2005). Person identification from biological motion: effects of structural and kinematic cues. *Perception & psychophysics*, **67**, 667–75.

VALENTINE, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, **79**, 471–491.

VAN BOXTEL, J.J.A. & LU, H. (2012). Signature movements lead to efficient search for threatening actions. *PloS one*, **7**, e37085.

VAN DEN STOCK, J., RIGHART, R. & DE GELDER, B. (2007). Body expressions influence recognition of emotions in the face and voice. *Emotion*, **7**, 487–94.

VAN DEN STOCK, J., GRÈZES, J. & DE GELDER, B. (2008). Human and animal sounds influence recognition of body language. *Brain research*, **1242**, 185–90.

VAN DER ZWAN, R., MACHATCH, C., KOZLOWSKI, D., TROJE, N.F., BLANKE, O. & BROOKS, A. (2009). Gender bending: auditory cues affect visual judgements of gender in biological motion displays. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, **198**, 373–82.

VANRIE, J. & VERFAILLIE, K. (2004). Perception of biological motion: a stimulus set of human point-light actions. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, **36**, 625–9.

VANRIE, J., DEKEYSER, M. & VERFAILLIE, K. (2004). Bistability and biasing effects in the perception of ambiguous point-light walkers. *Perception*, **33**, 547–560.

VATAKIS, A. & SPENCE, C. (2008). Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli. *Acta psychologica*, **127**, 12–23.

VERFAILLIE, K. (1993). Orientation-dependent priming effects in the perception of biological motion. *Journal of Experimental Psychology: Human Perception and Performance*, **19**, 992.

VERVERIDIS, D. & KOTROPOULOS, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, **48**, 1162–1181.

VINES, B.W., KRUMHANSL, C.L., WANDERLEY, M.M. & LEVITIN, D.J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition*, **101**, 80–113.

VLASIC, D., ADELSBERGER, R., VANNUCCI, G., BARNWELL, J., GROSS, M., MATUSIK, W. & POPOVIĆ, J. (2007). Practical motion capture in everyday surroundings. *ACM Transactions on Graphics*, **26**, 35.

VROOMEN, J. & DE GELDER, B. (2000). Sound enhances visual perception: cross-modal effects of auditory organization on vision. *Journal of experimental psychology. Human perception and performance*, **26**, 1583–90.

WALLBOTT, H.G. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, **28**, 879–896.

WATSON, A. & PELLI, D. (1983). QUEST: A Bayesian adaptive psychometric method. *Attention, Perception, & Psychophysics*.

WOLPERT, D.M., DOYA, K. & KAWATO, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **358**, 593–602.

WOLPERT, D.M., DIEDRICHSEN, J. & FLANAGAN, J.R. (2011). Principles of sensorimotor learning. *Nature reviews. Neuroscience*, **12**, 739–51.

WUERGER, S.M., CROCKER-BUQUE, A. & MEYER, G.F. (2012). Evidence for Auditory-Visual Processing Specific to Biological Motion. *Seeing and Perceiving*, **25**, 15–28.

YOU, M., CHEN, C., BU, J., LIU, J. & TAO, J. (2006). Emotion Recognition from Noisy Speech. In *2006 IEEE International Conference on Multimedia and Expo*, 1653–1656, IEEE.